Relationship of sediment toxicity to benthic community impacts and water body management decisions

Steven M. Bay, Joseph R. Gully[1], Darrin J. Greenstein, Philip J. Markle[1] and J. Ananda Ranasinghe

Abstract

Assessments of the impacts of sediment contaminants on the benthic community often rely on multiple lines of evidence such as the sediment quality triad (sediment chemistry, sediment toxicity, and benthic community condition), which was the approach recently approved for California's sediment quality objectives (SQO). However, other California water quality programs allow the use of sediment toxicity data alone as the basis for determining sediment quality, with the implicit assumption that toxicity is predictive of impacts to the benthic community. The accuracy of this assumption is uncertain, as thresholds have not been developed to relate laboratory measures of sediment toxicity directly to impacts on benthic communities. The objectives of this study were to determine the predictive ability of toxicity thresholds with respect to benthic community condition, and to investigate the environmental management implications of using toxicity as the sole indicator of biological impacts. A data set of 441 matched sediment toxicity (10-day amphipod survival tests using *Eohaustorius estuarius*) and benthic condition measurements (four benthic indices) for southern California marine embayments was used. Toxicity results were significantly correlated with benthic condition, but the strength of this correlation was weak and there was relatively low agreement between assessment categories used in the SQO program. Evaluation of various threshold values indicated that toxicity had low efficiency in predicting benthic community condition, as thresholds that resulted in the highest accuracy for predicting disturbed benthos also had low sensitivity. Reliance on just toxicity information for 303(d) impaired water body listing decisions could result in the listing of many water bodies with healthy benthic conditions or the delisting of water bodies with degraded conditions. Use of a multiple line of evidence approach is the most effective way to balance the strengths and limitations of each of the lines of evidence used to assess sediment quality.

Introduction

The health of the benthic community is the indicator of choice for monitoring and assessing the impacts of sediment contamination on marine ecosystems. The diverse community comprising the benthos has limited mobility and is exposed to sediment contaminants through direct contact and feeding activities, resulting in high sensitivity and response. In addition, the benthos are important components of aquatic food webs, serving as forage for bottom-feeding fishes and a pathway of contaminant exposure to fish, wildlife, and humans. Although measuring the health of benthos remains challenging in some habitats, such as sediments beneath low salinity waters, where reliable indices of benthic community condition are not available, assessing the condition of the benthos is a fundamental part of coastal monitoring programs nationwide (United States Environmental Protection

[1] *Sanitation Districts of Los Angeles County, Whittier, CA.*

Agency (USEPA) 1998, 2004), and is a primary focus of California's Water Quality Control Plan for Enclosed Bays and Estuaries (SWRCB 2009).

Linking the health of the benthos to pollution impacts is often challenging because impacts of natural factors such as salinity variation and physical disturbance confound the ability to associate impacts on the benthos with contaminant exposure. Benthic responses to natural and anthropogenic disturbances are sufficiently similar that no examples successfully differentiating these types of impacts exist. Additional indicators (lines of evidence) of sediment quality, such as sediment chemistry and toxicity, are commonly used to provide greater confidence in the determination of adverse impacts related to contamination (Adams *et al* 2005). This sediment quality triad approach has been adopted by the California State Water Resources Control Board as the means for determining compliance with the Sediment Quality Objective (SQO) for benthic community health (SWRCB 2009).

While the SQO assessment framework requires the use of three lines of evidence, other water quality programs allow the use of toxicity data alone, or in combination with chemistry data, as the basis for determining sediment quality in relation to benthic community protection. An example is California's listing policy for impaired water bodies under Section 303(d) of the Federal Clean Water Act (SWRCB 2004); sediment toxicity alone may be used for both listing and delisting purposes. Such flexibility was intended to facilitate management decisions for water bodies having limited monitoring resources or data, and is based on the assumption that changes in sediment toxicity correspond to ecological impact.

Interpretation of toxicity data as the sole indicator of biological effects is potentially problematic, however, as standardized thresholds have not been developed to relate laboratory measures of sediment toxicity to impacts on benthic communities. While whole sediment toxicity tests, such as the widely used 10-day amphipod survival test, have been shown to correlate with benthic community impacts, the relationship is often weak and appears to vary depending on geographic location, study design, test species, and spatial scale (Long *et al* 2001). Agencies using toxicity as the sole line of evidence for sediment quality assessment often use data interpretation thresholds developed for other purposes (e.g., regional sediment quality triad surveys or the SQO program), incorrectly assuming that the thresholds have been calibrated to accurately predict biologically or ecologically significant effects. The accuracy of using uncertain toxicity thresholds as a surrogate for benthic community condition assessments is unknown and could result in inaccurate water body assessments and flawed environmental management decisions.

The objectives of this study were to determine the predictive ability of various toxicity thresholds with respect to benthic community condition, and to investigate the environmental management implications of using toxicity as the sole indicator of benthic community condition. Using data from samples collected in southern California, both objectives were evaluated using the toxicity thresholds specified in the California SQO assessment framework as well as other thresholds selected to control specific types of errors in the assessment.

Methods

Data

Matched toxicity, chemistry, and benthic macrofaunal species abundance data from 0.1m^2 Van Veen sediment grab samples collected at 441 stations were used in the analysis. These data were obtained from a database containing sediment quality information from California embayments (http://www.sccwrp.org/Data/SearchAndMapData/DataCatalog/CaliforniaSedimentQualityObjectivesDatabase.aspx). The samples used in this study were collected in southern California embayments during six regional monitoring or research surveys conducted between 1998 and 2003. All stations were in enclosed bays or harbors at subtidal depths.

The toxicity of whole sediment to amphipods was determined using a 10-day survival test with *Eohaustorius estuarius* (USEPA 1994, ASTM 2002). Toxicity response ranges based on thresholds for *E. estuarius* specified in the CA SQO program were used to classify the results into four categories: 1) Nontoxic – response equivalent to that in controls (0-10% mortality); 2) Low Toxicity – a response of variable reproducibility where control adjusted mortality was greater than the controls, but less than the threshold for moderate toxicity (11-18% mortality); 3) Moderate Toxicity – a consistent and strong response above a threshold of 18% mortality, representing the 90th percentile minimum significant difference (MSD) for the test (19-41% mortality); and 4) High Toxicity – where mortality was greater than 41%, a threshold representing the average of two effect values: the 99th percentile MSD and the 75th percentile of responses in samples that were statistically significant from the control. Under the SQO policy, categories 3 and 4 represent sediments where confidence is high that a toxic effect is present whereas categories 1 and 2 indicate sediments where the response is within normal test variability for nontoxic sediments. For statistical analyses of predictive ability requiring a binary classification of toxicity, samples having a toxicity category of 1 or 2 were considered nontoxic and samples with toxicity categories of 3 or 4 were considered toxic.

A multi-step process was used to determine benthic community condition at each site. Samples were collected in the field and processed in the laboratory to produce species abundance data. The species abundance data were used to calculate four benthic indices: 1) the Index of Biotic Integrity (IBI; Thompson and Lowe 2004); 2) the Relative Benthic Index (RBI; Hunt *et al.* 2001); 3) the River Invertebrate Prediction and Classification System (RIVPACS; Wright *et al.* 1993, Van Sickle *et al.* 2006); and 4) the Benthic Response Index (BRI; Smith *et al.* 2001, Smith *et al.* 2003, Ranasinghe *et al.* 2009). Each index value was assigned to one of four condition categories using threshold values that were established during index development and validated with southern California marine bay data (Bay *et al.* 2009, Ranasinghe *et al.* 2009).

The four benthic condition categories were: 1) Undisturbed - a community that would occur at a reference site for that habitat; 2) Low Disturbance - a marginal deviation from reference with a benthic community that exhibits some indication of stress, but within the measurement variability of reference condition; 3) Moderate Disturbance - a community that exhibits evidence of physical, chemical, natural or anthropogenic stress; and 4) High Disturbance - a community exhibiting a high magnitude of stress. Categories 2, 3, and 4 are characterized by loss of 5, 25, and 50%, respectively, of the potential reference species in Category 1 (Smith *et al*. 2003). They correspond to southern California coastal assessment categories characterized by changes in relative abundance of species, loss of biodiversity due to the loss of sensitive species from the assemblage, and loss of community function, where taxonomic groups, particularly arthropods and ophiuroids, are for the most part excluded. The overall benthic condition category (benthic line of evidence or benthic LOE value) for each site was evaluated as the median of the numeric condition categories (Undisturbed=1, Severe disturbance=4) of the four indices (Ranasinghe *et al.* 2009). If the median for the index combination fell between categories, it was rounded up to the higher effects category. Samples where the median score was in categories 3 and 4 were considered to be in poor condition (disturbed), while those within categories 1 and 2 were considered to be in good condition (undisturbed).

Data Analysis

Three approaches were used to describe relationships between toxicity and benthic community condition. First, associations between amphipod mortality and benthic measures, independent of toxicity and benthic index thresholds, were examined by calculating nonparametric Spearman correlation coefficients. Second, the correspondence between the SQO assessment categories for the toxicity and benthic lines of evidence was measured by calculating percent agreement between toxicity and benthic condition categories.

Finally, the effectiveness of various toxicity thresholds for predicting benthic community condition was quantified using five metrics: sensitivity, specificity, positive predictive value, negative predictive value, and overall efficiency (Shine *et al.* 2003). Sensitivity is the probability of a sample with degraded benthos (disturbed) exceeding the toxicity threshold (e.g., correctly classifies a sample with disturbed benthos). Specificity is the probability of a site with undisturbed benthos having a response below the toxic threshold, correctly classifying a healthy benthic condition. The positive predictive value is the probability that a sample exceeding the toxicity threshold will have a degraded benthic condition (e.g., correct prediction of benthic disturbance). The negative predictive value is the probability that a sample below the toxicity threshold will have a healthy benthic condition. The overall efficiency represents the likelihood of making a correct prediction of healthy or degraded benthos based on the toxicity threshold.

The Shine *et al*. (2003) calculations are based on the binary disturbed or undisturbed results for benthos, and toxic or nontoxic toxicity tests.

Equations 1 through 5 were used to calculate the performance metrics (after Shine *et al.* 2003).

Sensitivity = B/(A + B) Eq. 1

Specificity = C/(C + D) Eq. 2

Positive predictive value = B/(B + D) Eq. 3

Negative predictive value = C/A + C) Eq. 4

Overall efficiency = B + C/(A + B + C + D) Eq. 5

where: A = number of disturbed samples below the toxicity threshold; B = number of disturbed samples above the toxicity threshold; C = number of non disturbed samples below the toxicity threshold; and D = number of non disturbed samples above the toxicity threshold

The effectiveness of toxicity test data alone to assess benthic community condition for impaired waterbody listing and delisting decisions was also evaluated. For this analysis, the probability of correctly listing or delisting a water body based on various toxicity thresholds was determined using data sets that were a random subsample from a larger data set where the benthic community condition was known to be either healthy (undisturbed) or disturbed. Each random evaluation data set was evaluated for listing and delisting using criteria specified in the State policy where ≥3 toxicity exceedances results in a listing and ≤2 exceedances is required for delisting (SWRCB 2004).

The integrated benthic condition score was used to segregate the data into undisturbed and disturbed benthos subsets. Of the 441 sites, 281 comprised the undisturbed benthos subset and 160 were placed in the disturbed benthos subset. Thirty random sites, with replacement, were subsampled from each subset 10,000 times to derive listing and delisting probability curves for integer toxicity thresholds ranging from 0% to 100% amphipod mortality. In each iteration, the percent amphipod mortality associated with each of the 30 samples was compared against potential toxicity thresholds ranging from 0% to 100% mortality. At each integer threshold, samples where the observed mortality in the toxicity test was greater than the threshold were scored as toxic while those equal to or below the threshold were scored as nontoxic. The total number of toxic samples at each threshold was used to determine the listing or delisting outcome at each integer threshold value.

Three different approaches were used to select discrete toxicity thresholds for calculation of listing and delisting outcomes associated with undisturbed and disturbed benthos conditions. The first approach was based on using the 18% mortality threshold used in the SQO assessment process to distinguish sediments with moderate or high toxicity from those with less toxicity. The second approach selected the toxicity thresholds that resulted in high sensitivity in listing disturbed water bodies (e.g., >90% of disturbed areas listed as impaired and <10% disturbed areas delisted). The final approach selected toxicity thresholds that resulted in few listing errors (e.g., <10% of undisturbed areas listed and >90% of undisturbed areas delisted). Effectiveness of these toxicity thresholds for impaired waterbody listing purposes was evaluated by comparing the percentage of correct decisions (i.e., listing disturbed benthos and delisting undisturbed benthos) to incorrect decisions (i.e., listing undisturbed benthos and delisting disturbed benthos).

Results

Scores for each of the four indices comprising the benthic LOE showed highly variable associations with toxicity (percent mortality). For example, nontoxic samples (<10% mortality) were classified as having benthic conditions ranging from Reference to High Disturbance (Figure 1). Similarly, some samples with very high toxicity (>80% mortality) were classified by each index as having a reference benthic condition.

Significant correlations between benthic condition and mortality were observed for the BRI, RBI and IBI (Figure 1). While these correlations were highly significant, the correlation coefficients were all less than 0.35, indicating relatively weak associations. The RIVPACS index was not correlated with mortality ($p=0.227$). Differences in categorization were also apparent among the indices. For example, the RBI classified far more samples in the High Disturbance category than the other indices. All subsequent analyses used the SQO approach to integrate benthic condition as the median of the four index condition categories, in order to minimize differences among index results. Index integration resulted in relatively few (22 of 441) samples being classified as High Disturbance (category 4) although another 138 samples were classified as Moderate Disturbance (category 3).

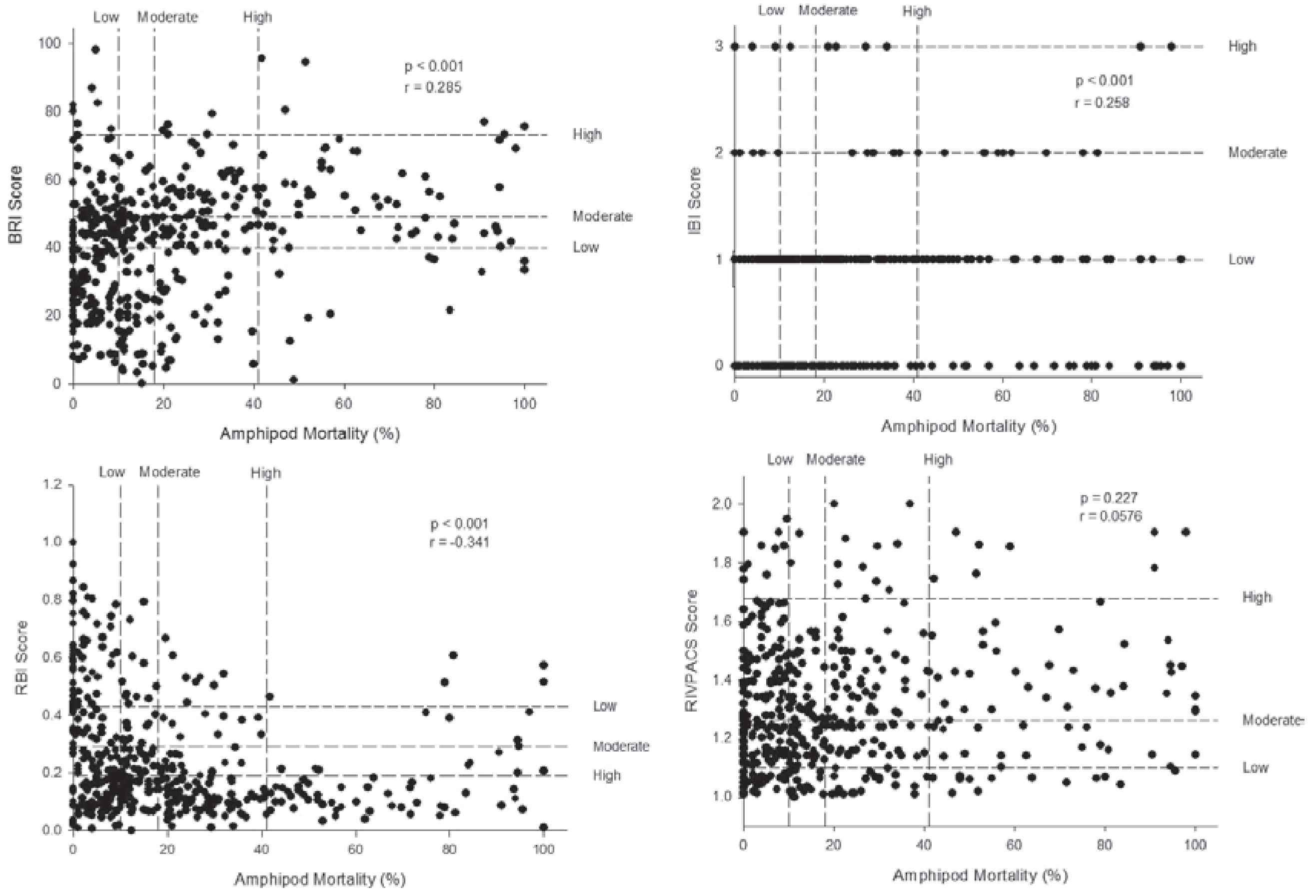

Figure 1. Benthic index scores versus amphipod mortality for the four indices used to assess benthic condition. BRI (top left); RBI (bottom left); IBI (top right); RIVPACS (bottom right). P value and r are from Spearman's correlations. Dashed lines indicate the respective SQO thresholds for benthic condition and toxicity.

The toxicity and integrated benthic condition categories were in agreement (i.e., same category score) for only 35% of the samples (156 of 441 samples; Table 1). The lowest rates of agreement (i.e., along diagonal of Table 1) were found at the extremes of toxicity response (Nontoxic and High Toxicity categories), a result that corresponds with the high variation in results seen in the individual index scatter plots (Figure 1). These results indicate that predictions of benthic condition based on toxicity may be in error more than 50% of the time, even at extremes of the toxicity response distribution.

Predictive ability and classification error rates were calculated using a binary classification system (i.e., not toxic/toxic and non disturbed/disturbed) instead of the four categories in order to represent typical assessment scenarios. An example of the binary classification results is shown for the Moderate Toxicity threshold (18% mortality) in Figure 2. Use of this toxicity threshold results in 92

Table 1. Contingency table of toxicity and integrated benthic categories. Shaded cells indicate agreement between categories.

		Toxicity Category				
		1	2	3	4	Total
Benthic Category	1	73 36.3%	10 15.9%	10 9.4%	5 7.0%	98
	2	77 38.3	35 55.6%	45 42.5%	26 36.6%	183
	3	43 21.4%	17 27.0%	43 40.6%	35 49.3%	138
	4	8 4.0%	1 1.6%	8 7.5%	5 7.0%	22
Total		201	63	106	71	441

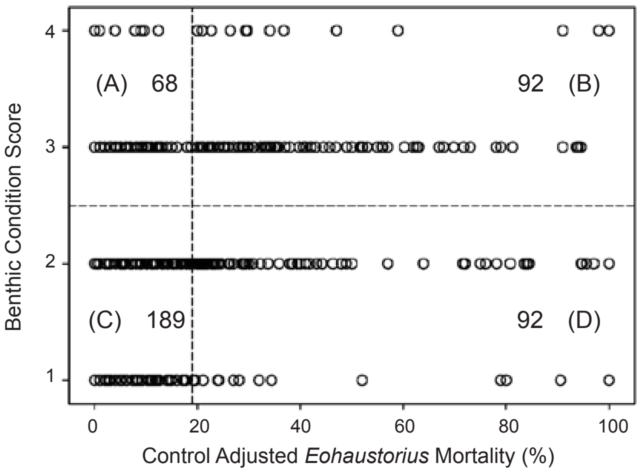

Figure 2. Relationship of *E. estuarius* toxicity response to integrated benthic biological condition. Vertical and horizontal dashed lines represent condensed SQO toxic/non-toxic and disturbed/undisturbed biological condition thresholds, respectively. Letters for the four quadrants correspond to Equations 1 through 5. Numbers are the count of samples in each quadrant.

false positives (quadrant D) and 68 false negatives (quadrant A), compared to 281 correct predictions of benthic condition (sum of quadrants B and C).

The classification metrics for the three SQO toxicity thresholds show that performance tradeoffs occur with the use of each threshold (Table 2). For example, use of the High toxicity threshold results in low sensitivity (0.26), meaning that this threshold would fail to correctly classify 74% of the samples having benthic disturbance. Use of the High threshold has a higher positive predictive value (0.57) relative to the other thresholds, but the lowest negative predictive value. The Low threshold had the highest accuracy in predicting non disturbed benthos (negative predictive value), but low specificity (only 39% of non disturbed samples below Low threshold). Overall prediction efficiency for these three thresholds was moderate (53 to 66%).

Examination of the predictive values for all possible mortality thresholds values failed to identify a threshold with high values for both positive and negative predictive ability (Figure 3). Positive predictive values declined and were unstable at thresholds of >58% mortality, contrary to expectations. The maximum negative predictive value of 0.87 occurred at 0% mortality and after an initial decline through 10% mortality, remained similar at higher mortality thresholds.

Plots of sensitivity and specificity varied in opposite directions, indicating that there was no toxicity threshold that yielded low rates of both false negative and positive errors (Figure 4). The two curves intersected at 15% mortality with a value of 0.60 for both sensitivity and specificity, representing a balance of false positive and negative error rates (40%). These plots indicate that use of thresholds representing extremes of response (e.g., nontoxic or > 60% mortality) will have one-directional error rates >70%.

The maximum overall prediction efficiency of 0.67 was obtained at 24% mortality (Figure 5). Efficiency varied little between 30 and 100% mortality, suggesting little benefit in performance at these higher thresholds. Any increase in accuracy in prediction (predictive value) was offset by reduced sensitivity.

The limited ability of toxicity tests to accurately predict benthic conditions also resulted in inaccurate decisions regarding the placement or removal of areas from the impaired waterbody list. The

Table 2. Quantitative measures of the effectiveness of the toxicity Line of Evidence to predict benthic community impacts at the three toxicity threshold levels.

Metric	Toxicity Threshold		
	Low (10% Raw Mort)	Moderate (18% Con. Adj)	High (42 % Con. Adj)
Sensitivity	0.78	0.57	0.26
Specificity	0.39	0.68	0.89
Positive Predictive Value	0.42	0.50	0.57
Negative Predictive Value	0.76	0.73	0.68
Overall Efficiency	0.53	0.64	0.66

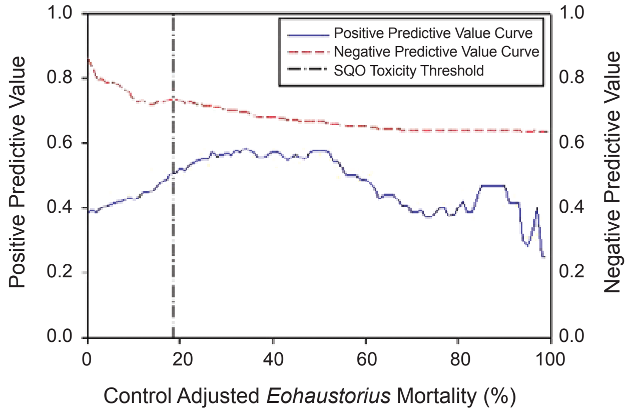

Figure 3. Positive (solid line) and negative (dashed line) predictive values for benthic condition as a function of acute toxicity threshold. Vertical reference line indicates SQO-based toxicity threshold of 18% mortality.

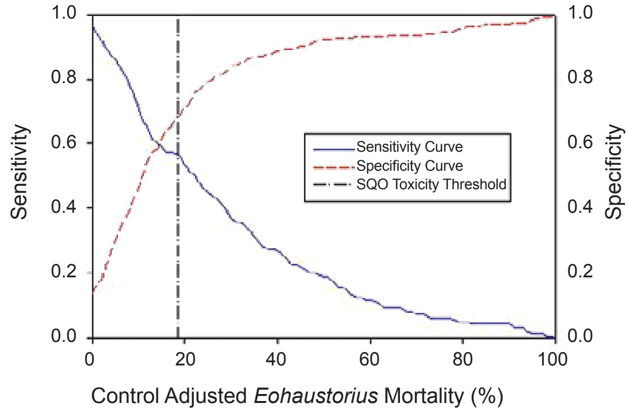

Figure 4. Sensitivity (solid line) and specificity (dashed line) for benthic condition as a function of acute toxicity threshold. Vertical reference line indicates SQO-based toxicity threshold of 18% mortality.

listing probability curves for areas of disturbed and undisturbed benthos (Figure 6) indicate that at toxicity thresholds of 16% amphipod mortality or below, there is certainty (100% probability) that an area will be listed as impaired, regardless of whether the benthos is disturbed or undisturbed. The greatest overall accuracy in listing was obtained using a threshold of 52% amphipod mortality, where there was a 92% probability of listing disturbed areas and a 39% probability of listing undisturbed areas.

Similar results were obtained in terms of delisting decisions (Figure 7). The delisting probability curves indicated that there is no possibility of any area being delisted, regardless of benthic condition, using toxicity thresholds at or below 16% amphipod mortality. As with the listing curves, maximum

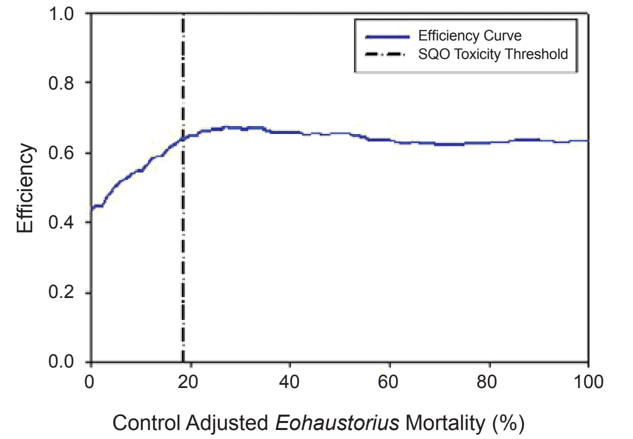

Figure 5. Overall efficiency of toxicity thresholds for predicting benthic condition. Vertical reference line indicates SQO-based toxicity threshold of 18% mortality.

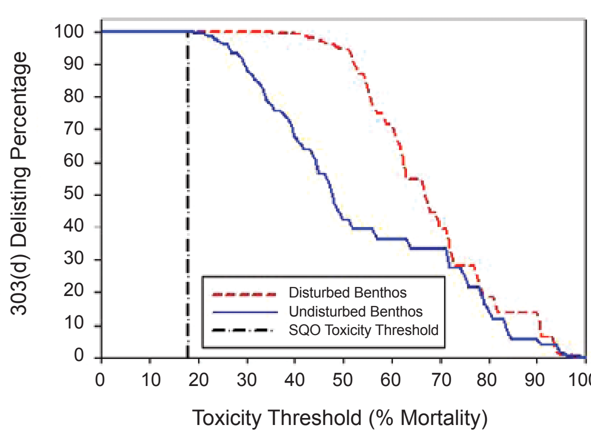

Figure 6. Simulated relationships between toxicity threshold and the probabilities of listing disturbed (solid line) and undisturbed (dashed line) benthic communities. Vertical reference line indicates SQO-based toxicity threshold of 18% mortality.

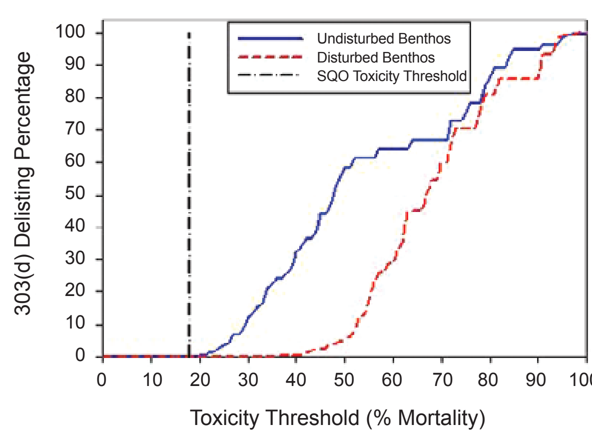

Figure 7. Simulated relationships between toxicity threshold and the probabilities of delisting undisturbed (solid line) and disturbed (dashed line) benthic communities. Vertical reference line indicates SQO-based toxicity threshold of 18% mortality.

accuracy for delisting was obtained using a threshold of 52% amphipod mortality. This threshold resulted in delisting 61% of undisturbed sites and 8% of disturbed sites.

Use of the SQO toxicity threshold of 18% mortality also resulted in highly ineffective listing/delisting decisions. When this threshold was used for listing purposes, 99.9% of areas with undisturbed benthos were listed as not meeting water quality standards (Table 3). Use of the 18% mortality threshold also made it virtually impossible to delist a water body having healthy benthos, only 0.01% of data sets with undisturbed benthos met the criteria for delisting using this threshold (Table 4).

Selection of toxicity thresholds to reduce listing errors had limited success. Use of a 52% mortality threshold provided high sensitivity for listing disturbed sites (>90%, Table 3), but also resulted in a nearly 40% probability of incorrectly listing undisturbed sites. When a threshold resulting in few (<10%) listing errors was used (84% mortality), this threshold was ineffective at identifying those sites known to be disturbed (13.9%).

Identical toxicity threshold values for these two environmental management approaches were obtained using the delisting probability curves. A toxicity threshold of 52% amphipod mortality ensures that few (<10%) of the disturbed sites are mistakenly delisted as impaired (8.0%) but would only identify 61.2% of the undisturbed sites that should be delisted (Table 4). If the threshold was selected to ensure that most (>90%) undisturbed areas are delisted (84% mortality), this threshold would also incorrectly delist 86% of disturbed sites (Table 4).

Discussion

This study expands upon previous investigations of the relationship between acute sediment toxicity and benthic community condition in southern California. Most previous studies were limited by a small sample size, few data for *E. estuarius*, and a lack of robust indices of benthic community condition (Long *et al*. 2001). In spite of differences in their data and approach, our results are generally similar and show a weak, but significant correspondence between toxicity and benthic disturbance (Bay *et al*. 2003, Anderson *et al*. 2001, Long *et al*. 2001). These previous studies also found a high level of uncertainty in the relationship between toxicity to amphipods and benthic community composition, with a wide range of variation in benthic community metrics for nontoxic samples (e.g., <10% mortality), as well as for samples with substantial toxicity (Anderson *et al*. 2001).

Sediment toxicity and benthic community condition represent two independent measures of sediment quality, and both lines of evidence provide important and complementary information regarding sediment quality. However, we were unable to identify an approach or toxicity threshold that resulted in accurate and effective predictions of benthic condition. Our analyses explored several methods of relating toxicity to benthic community condition (e.g., individual indices, integrated condition index, multiple response categories, binary response categories) and none of these approaches provided much better than a 50% level of accuracy. The best ability to predict a disturbed benthic condition was generally obtained when a toxicity

Table 3. Probability of toxicity based 303(d) listing of disturbed (unhealthy) and undisturbed (healthy) benthic communities under three environmental management scenarios.

Management Scenario	Listing Probability		
	Toxicity Threshold	Disturbed Benthos	Undisturbed Benthos
List Using SQO Policy Toxicity Threshold	18%	100%	100%
List Most (~90%) Disturbed Areas	52%	92%	39%
List Few (~10%) Undisturbed Areas	84%	14%	12%

Table 4. Probability of toxicity based 303(d) delisting of undisturbed (healthy) and disturbed (unhealthy) benthic communities under three environmental management scenarios.

Management Scenario	Delisting Probability		
	Toxicity Threshold	Disturbed Benthos	Undisturbed Benthos
Delist Using SQO Policy Toxicity Threshold	18%	0%	0%
Delist Most (~90%) Undisturbed Areas	84%	86%	93%
Delist Few (~10%) Disturbed Areas	52%	8%	61%

threshold of 20-40% was applied, but even this threshold range would incorrectly classify 20-30% of samples with healthy benthos as disturbed and fail to identify at least 50% of the samples with disturbed benthos.

The weak relationship between toxicity and benthic community condition can lead to poor environmental management decisions if toxicity is used as the sole measure of sediment quality. The implications of this approach were illustrated using California's 303(d) listing methodology, where it was shown that listing or delisting decisions based on toxicity rarely corresponded with the actual condition of the benthos. Use of 20% mortality as a threshold for determining an ecologically significant level of toxicity (a common practice), would result in a >95% chance of listing a water body with an undisturbed benthic community as not meeting water quality standards.

The variable correspondence between toxicity and benthic community condition is likely due to fundamental differences between these measures, rather than errors in measurement. Benthic communities respond to a broad range of environmental factors (Diaz *et al.* 2004, Pinto *et al.* 2009, Neto *et al. In press*), while the organisms in toxicity tests respond specifically to the effects of chemical contaminants present in the sediments. In addition to sediment contaminants, benthic composition may be affected by many natural and anthropogenic factors (Neto *et al. In press*). Natural disturbances such as strong currents and wave turbulence, salinity changes due to rainfall or tides, predation and burrowing by rays and bottom feeding fishes are reflected in the benthic community and, therefore, in benthic community assessment results. Even factors such as land use patterns and nutrient loads have been linked to benthic community condition (Dauer *et al.* 2000, Borja and Dauer 2008).

Discrepancies between the temporal and spatial scales on which benthic communities and toxicity tests respond may also account for assessment differences. Benthic communities integrate the effects of multiple events over time and an anoxic or hypoxic event may depress benthic abundances, diversity, and assessment categories until the next surge of recruitment, which may be months later (Dauer *et al.* 1992, Diaz and Rosenberg 1995, Diaz 2001, Seitz *et al.* 2009). Benthic condition is typically assessed based on animals collected in one grab, while several other grabs are often required to collect sufficient sediments for toxicity tests. Thus, small scale spatial variability in contaminant concentrations may also account for differences between benthic and toxicity assessments.

Differences between laboratory and field exposure conditions may also account for a lack of concordance in response. Laboratory toxicity tests use standardized methods for handling the sediment and exposure of animals that are intended to reduce variability and confounding factors. However, the laboratory test procedures such as sediment homogenization and acclimation time can influence biological and geochemical properties of the sediments with resulting effects on toxicity (Word *et al*. 2005). Laboratory exposure methods may increase or decrease contaminant bioavailability relative to the environment, resulting in corresponding changes in toxicity that will not be reflected in benthic community condition.

Measurements of sediment toxicity obtained using laboratory versus *in situ* tests with the same species often differ (e.g., Anderson *et al*. 2004) and illustrate that laboratory tests may not accurately represent the responses of benthic organisms in the field.

Each of the lines of evidence commonly used to assess sediment quality (chemistry, toxicity, benthic community condition, and bioaccumulation) has substantial strengths and limitations. Reliance on just one line of evidence (e.g., toxicity or benthic community condition) for sediment quality assessment can result in significant uncertainties, and the use of multiple lines of evidence is considered the best approach to improve the accuracy of sediment quality assessments and management decisions (Adams *et al*. 2005). The sediment quality objectives recently adopted by California utilizes a multiple line of evidence approach and provides guidance for integrating this information for assessing sediment quality in bays and estuaries (Bay and Weisberg 2009).

Literature Cited

Adams, W.J., A.S. Green, W. Ahlf, S.S. Brown, G.A. Burton, B. Chadwick, M. Crane, R. Gouguet, K.T, Ho, C. Hogstrand, T.B. Reynoldson, A.H. Ringwood, J.S. Savits, R.K. Sibley. 2005. Using sediment assessment tools and a weight-of-evidence approach. pp. 163-226 *in:* R.J. Wenning, G.E. Bately, C.G. Ingersoll, and D.W. Moore (eds.), Use of Sediment Quality Guidelines and Related Tools for the Assessment of Contaminated Sediments. Society of Environmental Toxicology and Chemistry. Pensacola, FL.

American Society for Testing and Materials (ASTM). 2002. Standard guide for conducting 10-day static sediment toxicity tests with marine and estuarine amphipods (E1367). pp. 693-719, 2002 Annual Book of ASTM Standards, Vol. 11.05. ASTM. West Conshohocken, PA.

Anderson, B.S., J.W. Hunt, B.M. Phillips, R. Fairey, C.A. Roberts, J.M. Oakden, H.M. Puckett, M. Stephenson, R.S. Tjeerdema, E.R. Long, C.J. Wilson and J.M. Lyons. 2001. Sediment quality in Los Angeles Harbor, USA: A triad assessment. *Environmental Toxicology and Chemistry* 20:359-370.

Anderson, B.S., J.W. Hunt, B.M. Phillips, P.A. Nicely, R.S. Tjeerdema and M. Martin. 2004. A comparison of *in situ* and laboratory toxicity tests with the estuarine amphipod *Eohaustorius estuarius*. *Archives of Environmental Contamination and Toxicology* 46:52-60.

Bay, S.M. and S.B. Weisberg. 2009. Framework for Interpreting Sediment Quality Triad Data. Integrated Environmental Assessment and Management. DOI: 10.1897/ieam.118.

Bay, S.M., D.J. Greenstein, J.A. Ranasinghe, D.W. Diehl and A.E. Fetscher. 2009. Sediment quality assessment draft technical support manual. Technical Report 582. Southern California Coastal Water Research Project. Costa Mesa, CA.

Bay, S.M., M.J. Allen, J. Anderson, D.E. Montagne, J.A. Noblet, J.A. Ranasinghe and S.B. Weisberg. 2003. Integration of the coastal ecology indicators. pp. 91-99 and Appendices *in*: J.A. Ranasinghe, D.E. Montagne, R.W. Smith, T.K. Mikel, S.B. Weisberg, D. Cadien, R. Velarde and A. Dalkey (eds.), Southern California Bight 1998 regional monitoring program: VII. Benthic macrofauna. Southern California Coastal Water Research Project. Westminster, CA.

Borja, Á. and D.M. Dauer. 2008. Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. *Ecological Indicators* 8:331-337.

Dauer, D.M., A.J. Rodi and J.A. Ranasinghe. 1992. Effects of low dissolved-oxygen events on the macrobenthos of the lower Chesapeake Bay. *Estuaries* 15:384-391.

Dauer, D.M., J.A. Ranasinghe and S.B. Weisberg. 2000. Relationships between benthic community condition, water quality, sediment quality, nutrient loads, and land use patterns in Chesapeake Bay. *Estuaries* 23:80-96.

Diaz, R.J. and R. Rosenberg. 1995. Marine benthic hypoxia: A review of its ecological effects and the behavioural responses of benthic macrofauna. *Oceanography and Marine Biology: An Annual Review* 33:245-303.

Diaz, R.J. 2001. Overview of hypoxia around the world. *Journal of Environmental Quality* 30:275-281.

Diaz, R.J., M. Solan and R.M. Valente. 2004. A review of approaches for classifying benthic habitats and evaluating habitat quality. *Journal of Environmental Management* 73:165-181.

Hunt, J.W., B.S. Anderson, B.M. Phillips, R.S. Tjeerdema, K.M. Taberski, C.J. Wilson, H.M. Puckett, M. Stephenson, R. Fairey and J.M. Oakden. 2001. A large-scale categorization of sites in San Francisco Bay, USA, based on the sediment quality triad, toxicity identification evaluations, and gradient studies. *Environmental Toxicology and Chemistry* 20:1252-1265.

Long, E.R., C.B. Hong and C.G. Severn. 2001. Relationships between acute sediment toxicity in laboratory tests and abundance and diversity of benthic infauna in marine sediments: A review. *Environmental Toxicology and Chemistry* 20:46-60.

Neto, J.M., H. Teixeira, J. Patricio, A. Baeta, H. Verissimo, R. Pinto and J.C. Marques. In press. The response of estuarine macrobenthic communities to natural- and human-Induced changes: Dynamics and ecological quality. *Estuaries and Coasts*.

Pinto, R.I.C., J. Patricio, A. Baeta, B.D. Fath, J.M. Neto and J.C. Marques. 2009. Review and evaluation of estuarine biotic indices to assess benthic condition. *Ecological Indicators* 9:1-25.

Ranasinghe, J.A., S.B. Weisberg, R.W. Smith, D.E. Montagne, B. Thompson, J.M. Oakden, D.D. Huff, D.B. Cadien, R.G. Velarde and K.J. Ritter. 2009. Calibration and evaluation of five indicators of benthic community condition in two California bay and estuary habitats. *Marine Pollution Bulletin* 59:5-13.

Seitz, R.D., D.M. Dauer, R.J. Llanso and W.C. Long. 2009. Broad-scale effects of hypoxia on benthic community structure in Chesapeake Bay, USA. *Journal of Experimental Marine Biology and Ecology* 381:S4-S12.

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

Smith, R.W., M. Bergen, S.B. Weisberg, D.B. Cadien, A. Dalkey, D.E. Montagne, J.K. Stull and R.G. Velarde. 2001. Benthic response index for assessing infaunal communities on the southern California mainland shelf. *Ecological Applications* 11:1073-1087.

Smith, R.W., J.A. Ranasinghe, S.B. Weisberg, D.E. Montagne, D.B. Cadien, T.K. Mikel, R.G. Velarde and A. Dalkey. 2003. Extending the southern California Benthic Response Index to assess benthic condition in bays. Technical Report 410. Southern California Coastal Water Research Project. Westminster, CA.

State Water Resources Control Board (SWRCB). 2004. Water quality control policy for developing California's Clean Water Act Section 303(d) list: Appendix A. SWCRB. Sacramento, CA.

SWRCB. 2009. Water quality control plan for enclosed bays and estuaries; Part I: Sediment quality. State Water Resources Control Board. Sacramento, CA.

Thompson, B. and S. Lowe. 2004. Assessment of macrobenthos response to sediment contamination in the San Francisco Estuary, California, USA. *Environmental Toxicology and Chemistry* 23:2178-2187.

United States Environmental Protection Agency (USEPA). 1994. Methods for assessing the toxicity of sediment-associated contaminants with estuarine and marine amphipods. EPA/600/R-94/025. USEPA Office of Research and Development. Narragansett, RI.

USEPA. 1998. Condition of the Mid-Atlantic Estuaries. EPA/600/R-98/147. USEPA Office of Research and Development. Washington, DC.

USEPA. 2004. National Coastal Condition Report II. EPA-620/R-03/002. USEPA Office of Research and Development. Washington, DC.

Van Sickle, J., D.D. Huff and C.P. Hawkins. 2006. Selecting discriminant function models for predicting the expected richness of aquatic macroinvertebrates. *Freshwater Biology* 51:359-372.

Word, J.Q., W.W. Gardiner and D.W. Moore. 2005. Influence of confounding factors on SQGs and their application to estuarine and marine sediment evaluations. pp. 633-686 *in:* R.J. Wenning, G.E. Bately, C.G. Ingersoll, and D.W. Moore (eds.), Use of Sediment Quality Guidelines and Related Tools for the Assessment of Contaminated Sediments. Society of Environmental Toxicology and Chemistry. Pensacola, FL.

Wright, J.F., M.T. Furse and P.D. Armitage. 1993. RIVPACS: a technique for evaluating the biological water quality of rivers in the UK. *European Water Pollution Control* 3:15-25.

Acknowledgments

The authors thank Chi-Li Tang (Sanitation Districts of Los Angeles County) for development of the spreadsheet macros used to develop 303(d) listing and delisting probability curves. We thank Kerry Ritter for assistance with data preparation and analysis. We also thank the many water quality regulation, management, and research organizations in California that have participated in the Southern California Bight Regional Monitoring Program and helped generate the data used in this study.