DEVELOPMENT AND EVALUATION OF SEDIMENT QUALITY GUIDELINES BASED ON BENTHIC MACROFAUNA RESPONSES

Draft Final Report

Kerry J. Ritter[1], Steven M. Bay[1], Robert W. Smith[2], Doris E. Vidal-Dorsch[1], and L. Jay Field[3]

[1]Southern California Coastal Water Research Project
3535 Harbor Blvd., Suite 110, Costa Mesa, CA 92626
www.sccwrp.org

[2]Deceased

[3]National Oceanic and Atmospheric Administration, Seattle, WA

November 14, 2007

Technical Report

EXECUTIVE SUMMARY

Sediment quality guidelines (SQGs) are often used to assess the potential of sediment contamination to adversely affect benthic macrofauna, yet the predictive ability of these guidelines for benthic community condition is poorly documented. Most studies of SQG predictive ability have examined the relationship with toxicity, which may not accurately reflect changes in benthic community condition. Here we compare the performance of six SQG approaches applied to a common data set of matched chemistry, toxicity, and benthic macrofauna measurements from southern California embayments. The predictive ability of each approach for both toxicity and benthic community effects was determined. Four of the SQG approaches were derived in previous national studies (ERM, LRM, SQGQ1, Consensus) and one was a regional variation of the LRM calibrated to California data (CA LRM). The final SQG approach was the Chemical Score Index (CSI), a new method that was derived using southern California benthic community condition data. An index of overall contamination (i.e., mean quotient, mean score, or maximum probability of toxicity) was calculated for each approach, and a set of three thresholds were selected for each index that classified the results into four categories of biological effect. The performance of each SQG approach was evaluated in terms of the correlation and classification accuracy with respect to amphipod mortality or benthic community disturbance.

All of the SQG indices were significantly correlated with changes in sediment toxicity and benthic community condition. Furthermore, the SQG index category thresholds for toxicity and benthic community condition were similar, indicating that the sediment toxicity tests and benthic indices had similar sensitivities to contamination. The benthos-based CSI had the highest classification accuracy for both toxicity and benthic community condition, but the differences were small relative to the other SQG approaches. The high degree of similarity in SQG performance between effects measures differed from other studies and may have been related to several factors: 1) the use of a consistent method for selecting classification thresholds, 2) optimization of the thresholds for California data, and 3) use of a relatively sensitive sediment toxicity test. Several options for increasing the relevance of SQGs for assessing the potential effects of sediment contamination on benthic macrofauna are described, including the use of multiple SQG approaches so that maximum relevance for each type of biological effect is achieved.

ACKNOWLEDGMENTS

We thank Chris Beegan from the California Water Resources Control Board, and Mike Connor and Bruce Thompson of the San Francisco Estuary Institute for their suggestions on the design of this study. Peggy Myre of Exa Data and Mapping and Ananda Ranasinghe compiled and standardized the data sets. Darrin Greenstein, Jeff Brown, and Diana Young assisted with data analysis. We also thank Peter Landrum, Ed Long, Todd Bridges, Tom Gries, Rob Burgess and Bob Van Dolah for their thoughtful review of the ideas contained within the document.

Work on this project was funded by the California State Water Resources Control Board under agreement 01-274-250-0.

INTRODUCTION

Resource managers often use sediment quality guidelines (SQGs) to interpret contamination data with respect to impacts on sediment dwelling organisms. The most widely used SQGs for this purpose are empirical guidelines that are based on the statistical analysis of matched sediment chemistry and biological effects data. Although the benthic community is often the receptor of concern for sediment quality assessments, laboratory sediment toxicity tests such as the 10-day amphipod survival test are frequently used as the sole measure of biological effects during SQG development (Wenning *et al*. 2005).

Various statistical approaches have been used to derive empirical SQGs. The commonly used approaches include the effects range-median (ERM; Long *et al*. 1995), logistic regression modeling (LRM; Field *et al*., 1999; 2002), sediment quality guideline quotient 1 (SQGQ1; Fairey *et al*. 2001), and consensus (MacDonald *et al*. 2000, Swartz 1999). ERMs are chemical values corresponding to the 50th (median) percentile of biological effects concentrations compiled from a variety of study types and sources. Logistic regression modeling (LRM) indicators model the probability that a sample is toxic at a given chemical concentration. Individual logistic regression models are developed for each chemical. The SQGQ1 is calculated using SQGs compiled from multiple sources. Consensus guidelines aggregate several types of SQGs in order to reduce the uncertainty of using a single SQG approach.

The predictive ability of empirical SQGs with respect to toxicity has been documented in many studies (summarized in Wenning *et al*. 2005). Predictive ability is often demonstrated by a high incidence of toxicity when multiple guidelines are exceeded (Long *et al*. 1998). Greater predictive ability is obtained when an integrative SQG index, such as the mean SQG quotient, is used to represent the magnitude of contamination (Long *et al*. 2000, 2006).

The relevance of toxicity-based SQGs for predicting benthic community effects is uncertain. There are many differences between laboratory toxicity tests and field biological effects studies that could affect the relationship between biological effects and sediment contaminants (Leung *et al*. 2005). First, toxicity tests are conducted in the laboratory under controlled conditions and may not reproduce the contaminant exposure characteristics and interactions between chemicals and environmental stressors found in nature. Second, the typical exposure duration of a sediment toxicity test is several days, while benthos are continually exposed to sediment contaminants during their lifespan. Third, benthic assemblages are composed of many species whose sensitivity to various contaminants may not be comparable to the few species used for toxicity testing. Finally, adaptation or acclimation of benthic species to contamination may reduce their response to specific chemicals.

In spite of these differences, the ability of empirical SQGs to predict benthic community impacts has been documented in several studies. Variations in mean ERM or PEL quotients have been shown to correspond with changes in a variety of benthic community metrics and indices (Hyland *et al*. 2003, Wenning *et al*. 2005). However, this correspondence has been documented for just a few SQG approaches and it is difficult to compare results between studies due to differences in data analysis methods. In addition, there is evidence that benthic community degradation may be occurring at much lower contaminant levels than those associated with the increased incidence of toxicity (Hyland *et al*. 2003, Leung *et al*. 2005), suggesting that currently applied thresholds for SQG interpretation may not be protective of benthic communities

One strategy to deal with these uncertainties is to use SQGs that are based on the responses of benthic macrofauna to sediment contamination. Several approaches have been used to derive benthic community-based SQGs: the apparent effects threshold (AET; Barrick *et al*. 1988), screening level concentration (SLC; Neff *et al*. 1986), and field-based species sensitivity distribution (f-SSD; Leung *et al*. 2005). While it is assumed that SQGs based on benthic community characteristics are better predictors of the adverse impacts of contaminants on benthic community condition, no systematic study has been conducted to compare their predictive ability relative to toxicity-based SQGs.

Here we compare the performance of six SQG approaches applied to a common data set of matched chemistry, toxicity, and benthic macrofauna measurements. Two issues were examined: 1) whether the predictive ability of each SQG approach is different for benthic community effects and toxicity and 2) whether the predictive ability for benthic community effects is improved by using SQGs based on benthic macrofauna responses to environmental stress.

METHODS

We assessed the performance of six SQG approaches by applying them to matched chemical concentration, amphipod mortality, and benthic macrofauna abundance data for southern California. An index of overall contamination (i.e., mean quotient, mean score, or maximum probability of toxicity) was calculated for each approach, and the correlation and classification accuracy determined with respect to amphipod mortality or benthic community indices. Four of the SQG approaches were derived in previous national studies (ERM, LRM, SQGQ1, Consensus) and one was a regional variation of the LRM calibrated to California data (CA LRM). The final SQG approach was the Chemical Score Index (CSI), a new method developed in this study that was derived using southern California benthic community condition data.

Data

Data consisting of 441 samples of matched toxicity, chemistry, and benthic macrofauna abundance data were compiled across 6 regional monitoring and research surveys in southern California embayments. The data were screened to select information that was of high quality and comparable. All stations were located in enclosed bays or harbors at subtidal depths and only data from surficial sediment (top 30 cm or less) were selected. The database included stations from Santa Barbara Harbor in the north to San Diego Bay in the south. More information on the studies used to populate this database can be found at http://www.sccwrp.org/data/2006_sqo.html.

Toxicity data consisted of solid-phase 10-d amphipod survival tests using *Rhepoxynius abronius* or *Eohaustorius estuarius* and conducted using standardized methods (USEPA 1994). The toxicity data were screened to ensure that conventional data quality objectives were met, including mean control survival >85% and overlying water ammonia concentrations below species-specific criteria (USEPA 1994). Amphipod survival was normalized to the control response for each test batch by expressing the results as a percentage of the control survival. Each toxicity result was classified into one of four categories of toxicological response that were based on classification systems used in other studies (Long *et al.* 2006). The toxicity categories were specific to each test species and were based on analyses of the minimum significant difference and magnitude of response (percent of control survival) to California samples (Bay *et al*. 2007a). The categories for *E. estuarius* were: Nontoxic (≥90% survival), Low Toxicity (≥82%), Moderate Toxicity (≥59%) and High Toxicity (<59%). The categories for *R. abronius* were: Nontoxic (≥90% survival), Low Toxicity (≥83%), Moderate Toxicity (≥70%) and High Toxicity (<70%). A wide range of toxicity was present in the dataset; however, most samples were classified in the Nontoxic or Low Toxicity categories (Figure 1).

Benthic community condition was determined based on species abundances collected from 1-mm sieves. Taxonomic inconsistencies among programs were eliminated by cross-correlating the species lists, identifying differences in nomenclature, and resolving discrepancies by consulting the taxonomists from each program. The Benthic Response Index (BRI) was used to describe the level of benthic disturbance for response comparison and SQG development. The BRI is the abundance-weighted average pollution tolerance score of organisms occurring in a sample. This index was originally developed for the southern California mainland shelf and extended into California's bays and estuaries (Smith *et al.* 2001, Smith *et al.* 2003). The BRI scores were used to classify each sample into one of four categories of disturbance (Ranasinghe *et al*. 2007): Reference (<39.96), Low Disturbance (≥39.96), Moderate Disturbance (≥49.15), or High Disturbance (>73.26). All categories of benthic condition were represented in the database, but most samples were classified as Reference or Low Disturbance; only 4% had High Disturbance (Figure 1).

The chemistry data selected for analysis had to meet several screening criteria that included a review of the data quality assessment by the study authors, use of comparable extraction/digestion methods, and measurement of a minimum suite of contaminants that included multiple metals and PAHs. Standardized sums of PAHs, DDTs, PCBs, and chlordanes were calculated using a consistent methodology for all samples. Low molecular weight PAHs were calculated as the sum of acenaphthene, anthracene, biphenyl, naphthalene, 2,6-dimethylnaphthalene, fluorene, 1-methylnaphthalene, 2-methylnaphthalene, 1-methylphenanthrene, and phenanthrene. High molecular weight PAHs was the sum of benzo(a)anthracene, benzo(a)pyrene, benzo(e)pyrene, chrysene, dibenz(a,h)anthracene, fluoranthene, perylene, and pyrene. Total PAHs was the sum of Low PAH and the High PAH values. Total PCBs were calculated from the sum of congeners 8, 18, 028, 44, 52, 66, 101, 105, 110, 118, 128, 138, 153, 180, 187, and 195. This sum was multiplied by 1.72 to estimate the total concentration of all congeners. Total DDTs represented the sum of p,p'-DDT, o,p'-DDT, p,p'-DDE, o,p'-DDE, p,p'-DDD, and o,p'-DDD. Total chlordane was the sum of alpha-chlordane (cis-chlordane), oxychlordane, trans-chlordane, trans-nonachlor, and gamma-chlordane.

Data were estimated for values reported as below reporting limits based on multiple regression imputation, taking advantage of covariation among the many chemical and sediment variables. Imputation produces lesser bias than conventional approaches for interpreting nondetect data, such as substituting zero or one-half of the reporting limit (Helsel 2005). SAS PROC MI (SAS Institute Inc, North Carolina, USA) was used to impute values in a sequential stepwise fashion by contaminant type. Metal data were estimated first, followed in order by pesticides, PAHs, and PCBs. The stepwise manner in which the groups of data variables were imputed was used because SAS PROC MI could not compute all imputations in a single step. The stepwise procedure also allowed for better control of the data variables used in the imputations for each chemical group.

The resulting data set contained a 1-2 order of magnitude range of concentration for each constituent (Table 1). Approximately 2/3 of the data (293 samples) were used for SQG development and threshold calibration. An independent validation data set, containing 1/3 of the data (148 samples) randomly selected from throughout the range of benthic community response, was used for performance evaluation of the SQG approaches.

Sediment Quality Indices

National SQGs

The ERM values used in the analyses were obtained from Long *et al.* (1995). The mean ERM quotient (mERMQ) for each sample in the data set was calculated by dividing each chemical concentration by its respective ERM and averaging the individual quotients (Long *et al*. 2000). The subset of ERM values used to calculate the mERMQ (Table 2) was the same as that used in previous mERMQ performance studies (Long *et al*. 2000).

The mean sediment quality guideline quotient 1 (SQGQ1) was calculated as described by Fairey *et al*. (2001). The SQG values used in the analysis are listed in Table 2.

The Consensus SQG values for PAHs and PCBs were midpoint effect concentrations obtained from Swartz (1999) and MacDonald *et al*. (2000), respectively. Values for DDTs, dieldrin, arsenic, cadmium, chromium, copper, lead, mercury, nickel, silver, and zinc were obtained from Vidal and Bay (2005). The mean Consensus quotient was calculated by dividing each chemical concentration by its respective SQG (Table 2) and averaging the individual quotients.

The Logistic Regression Model (LRM) approach was based on the statistical analysis of paired chemistry and amphipod toxicity data from studies throughout the U.S. (Field *et al.* 1999, 2002). The logistic model is described by the following equation:

$$p= e^{B0+B1\,(x)} / (1 + e^{B0+B1\,(x)})$$

where: p= probability of observing a toxic effect;
B0= intercept parameter;
B1= slope parameter; and,
x= concentration or log concentration of the chemical.

The chemical-specific models used in this study were based on an analysis of the accuracy for predicting toxicity for 37 candidate models. Models for 18 chemicals having low rates of false positives were selected for use (Table 3). The maximum probability of toxicity obtained from the individual models (P_{max}) for each sample was used as the index of overall contamination.

Regional SQGs

A version of the LRM calibrated using California data (CA LRM) was also evaluated, as recent analyses showed that the CA LRM had better performance for predicting the toxicity of southern California sediments relative to other SQG approaches (Bay *et al.* 2007b). The individual chemical models included in the CA LRM (Table 3) were selected from a library of candidate models that included national models, as well as models derived using a California data set (Bay *et al*. 2007b).

Chemical Score Index

The Chemical Score Index (CSI) describes the association between chemical contaminant concentrations and the severity of disturbance to Southern California benthic macrofauna. The index is the weighted mean of benthic community effect scores based on guidelines developed for 13 contaminants:

$$CSI = \Sigma(w \times cat)/\Sigma w$$

where: cat = benthic community category score for the contaminant and
w = weighting factor for that contaminant.

The benthic community disturbance category for each contaminant was determined by comparing the concentration to a set of three guideline values (Low, Moderate, High) developed using BRI results for southern California. These guidelines corresponded to the four BRI categories of benthic community condition: Reference (<Low guideline), Low Disturbance (≥Low guideline), Moderate Disturbance (≥Moderate guideline), and High Disturbance (>High guideline). A score of 1, 2, 3 or 4 was assigned to a category of Reference, Low, Moderate or High, respectively; the scores were then used to calculate the CSI.

The chemical-specific guidelines were developed using a statistical optimization procedure written in Splus that was based on maximizing the weighted kappa statistic (Cohen 1960, 1968). The weighted kappa statistic measures agreement, beyond that expected by chance, between two sets of sample classifications; in this case one based on CSI guidelines and one based on BRI categories. Weights for the kappa statistic were used to give partial credit for incorrect CSI classifications depending on how close they were to the BRI classifications. A linear weighting scheme was used that gave greater weight (i.e., more partial credit) when the disagreement was relatively small (Cicchetti-Allison 1971, Stokes *et al*. 2002). A weight of 1 was used for perfect agreement, and a weight of 1/3, 1/6, and 0 was assigned when the misclassification spanned two, three or four categories, respectively.

The guideline optimization procedure for each chemical was bootstrapped by conducting the analysis on subsamples containing an even distribution of BRI response categories. This step was incorporated in order to reduce the tendency of skewed data distributions to provide misleading kappa results (Kraemer and Bloch 1988, Fienstein and Cicchetti, 1990). Kappa is least susceptible to providing misleading measures of correspondence when distributions are even (Canran, *et al*. 2005, Lantz and Nebenzahl 1996). The data were subsampled by randomly selecting (with replacement) 40 affected (Moderate or High BRI classifications) and 40 unaffected (Reference or Low) samples from the development data set using PROC SURVEYSELECT (Figure 2A). We repeated the subsampling process, (i.e., bootstrapping) 50 times in order to ensure that nearly all samples were represented in at least one analysis and to create robustness of the thresholds to different data subsets.

The optimal set of three guideline values for each bootstrap sample was selected by computing weighted kappa statistics for all possible sets of guidelines occurring within a relatively dense set of possibilities. Mesh sizes for optimization reflected a distance between possible thresholds values of 5% of the range of data values for each chemical. In addition, distances between individual guidelines within each set were constrained to be no less than 10% of the range of data values for each chemical. These constraints ensured that optimization converged and resulting guidelines within a set were not too close to one another. For each data subset, the set of three guidelines that yielded the highest kappa statistic was selected (Figure 2B). The median of the guidelines across all subsets was chosen to be the final guideline value for the Low, Moderate, and High categories (Figure 2C).

The CSI weighting factor for each chemical was based on the median kappa value resulting from the guideline optimization analyses. The median kappa represented the relative strength of association between the variations in chemical concentration and benthic community condition. The kappa values were normalized to the largest median kappa for use as weighting factors in calculation of the CSI.

Threshold Development

Evaluating the indices with respect to categorical classification accuracy requires identification of category thresholds for each SQG index. Such thresholds are generally unavailable for these SQG approaches or vary in the method of development. The thresholds used in this study were developed for each SQG approach using a consistent methodology so that differences in performance would reflect inherent differences among approaches, rather than variations in how thresholds were assigned.

Two sets of thresholds, each defining four ranges of SQG index values that corresponded to biological effect categories, were established for each SQG approach. One set of three thresholds was based on sediment toxicity and the other set of thresholds was based on benthic community condition. The toxicity-based thresholds corresponded to the Nontoxic, Low Toxicity, Moderate Toxicity, and High Toxicity categories of amphipod survival defined previously. The thresholds based on benthic community condition also corresponded to four categories of biological response: Reference, Low Disturbance, Moderate Disturbance, and High Disturbance. The benthic community condition classification for each sample was based on the median of four benthic indices calibrated to southern California embayments (Ranasinghe *et al*. 2007): BRI, Index of Benthic Biotic Integrity (Thompson and Lowe 2004), Relative Benthic Index (Hunt *et al.* 2001, and River Invertebrate Prediction and Classification System (Wright *et al.* 1993, Van Sickle *et al.* 2006). The median benthic condition was used instead for threshold development instead of the BRI score used for CSI development because analyses conducted subsequent to CSI development indicated that the median approach provided a more accurate measure of benthic community condition (Ranasinghe *et al*. 2007).

Threshold selection for each SQG index relied on the same optimization method that was used to select the chemical-specific guideline values for the CSI (see above), except that optimization was based on percent agreement between the SQG index and biological response (benthos or toxicity) categories. Agreement was chosen over kappa for this step because thresholds based on agreement were less sensitive to data variations, particularly with respect to the High threshold.

Selection of the benthic community-based thresholds used the development data set described previously. The toxicity-based thresholds were selected using a larger Southern California data set that contained the benthic community condition data set described above plus toxicity data from additional studies were macrofauna abundance data was not available (Bay *et al*. 2007b). This larger data set was used in order to provide a more robust set of threshold optimization analyses.

Evaluation of SQG Performance

SQG performance was evaluated by quantifying the strength of association between chemistry and biological effect (i.e., toxicity or benthic community condition) in terms of both correlation and categorical classification accuracy. Correlation was measured as the nonparametric Spearman's correlation coefficient between the SQG index value (i.e., mean quotient, mean score, or P_{max}) and amphipod mortality or BRI score. Categorical classification accuracy was quantified as the percent agreement between the SQG index category (determined by applying the thresholds developed using the development data set) and the biological effect category (toxicity or median benthic community condition). The analyses were conducted across 50 bootstrap samplings of 40 unaffected and 40 affected (based on benthic community condition) samples. The 90^{th} percentile confidence limits of the highest correlation or percent agreement value were determined from the bootstrap samplings and used to identify the best performing SQG indices. All analyses were conducted using an independent validation data set that was not used for SQG or threshold development.

Percent agreement was based on the number of samples that are correctly classified and was calculated as:

$$A=(Nc/Nt)*100$$

where: A = percent agreement
Nc = number of samples correctly classified
Nt = total number of samples

RESULTS

CSI Development

There was substantial variability in the relationship between chemical concentration and benthic community condition. Plots of representative results and CSI guideline values are shown in Figure 3 for copper, which had one of the strongest relationships, and high molecular weight PAHs (weak relationship). Analyses for each contaminant usually showed a trend of greater frequency and severity of community disturbance at higher concentrations. A poor relationship was present at low chemical concentrations, where benthic community condition was highly variable and ranged from Reference to High Disturbance (Figure 3). High BRI scores at low concentrations may indicate the presence of samples impacted by stress due to noncontaminant factors such physical disturbance.

A set of guideline values was developed for 13 contaminants using the statistical optimization procedure (Table 4). The guidelines for copper, lead, and zinc produced the highest kappa values, indicating that classifications based on these metals had better predictive accuracy for benthic community disturbance than guidelines for other chemicals. The lowest kappa values were obtained for the high and low molecular weight PAHs. The normalized weighting factor for each contaminant, which was based on kappa, varied from 100 (copper) to 5 (LPAH).

SQG Index Performance

All of the SQG indices were significantly correlated with changes in sediment toxicity and benthic community condition (Table 5). The correlations were similar, both among indices and between biological effects measures. The toxicity correlations ranged from 0.42 (SQGQ1) to 0.54 (CSI), while the correlations with benthic community condition ranged from 0.46 (Consensus) to 0.53 (National LRM). The correlation between benthic condition and the new CSI approach was intermediate (0.49).

The SQG index category thresholds developed based on benthic condition were similar to the toxicity-based thresholds (Table 6). The toxicity-based thresholds were almost always lower than those based on benthos, but the differences were small (usually less than 10%). This similarity indicates that the toxicity and benthic indices had a similar sensitivity to the contamination patterns in the samples. The two High SQGQ1 thresholds showed the greatest variation: 0.81 versus 1.26.

Each of the SQG approaches had similar classification accuracy (percent agreement) for each type of biological effect (Table 5). The mean CSI score had the greatest percent agreement for both toxicity (50%) and benthic community condition (53%). The SQGQ1 had the lowest classification accuracy for each effects measure. Results for the other SQG indices varied within a relatively narrow range. There was no consistent trend in classification accuracy between the effects measures. Percent agreement was higher for benthos than for toxicity for CSI, CA LRM, and National LRM, but the trend was reversed for the other SQG approaches.

DISCUSSION

Each of the SQG approaches investigated in this study had predictive ability with respect to impacts on the benthic community. This ability was demonstrated by the presence of both significant correlations and moderate classification accuracy when the SQG indices were compared to measures of benthic community condition. These results are consistent with studies conducted in other locations in the United States (Long *et al*. 2006). Mean SQG quotients based on the ERMs have been shown to be predictive of marine benthic community disturbance in Florida (MacDonald *et al*. 2004), along the Atlantic and Gulf Coasts (Hyland *et al*. 2003), and in Washington (Long *et al*. 2005). This study documents for the first time that SQG indices derived from logistic regression models, consensus guidelines, and the CSI also have predictive ability with respect to benthic community impacts.

Each of the SQG indices had similar performance with respect to predicting sediment toxicity or benthic community impacts. The method of SQG derivation (e.g., ERM, LRM) and even the type of biological effect underlying the guidelines (i.e., toxicity or benthic community disturbance) did not have much influence on performance. This result seems counterintuitive, considering that the benthic community condition represents the combined response of many species, reflects a wider variety of endpoints (e.g., mortality, growth, reproduction), and incorporates a much longer exposure duration relative to the 10-day single species amphipod toxicity test used for comparison. This similarity underscores the robust nature of empirical SQGs and indicates that these guidelines are useful for interpreting chemistry data with respect to biological effects in both the laboratory and field.

Two factors likely contributed to the high similarity in classification accuracy among the SQG approaches. First, we used a consistent method for selecting SQG index classification thresholds that were optimized to southern California data. Second, thresholds were developed separately for toxicity and benthic community condition. Both of these factors minimized the potential for classification errors due to regional differences in toxicity test species, contamination pattern, or bioavailability. Had we used thresholds obtained from the literature for the national SQGs, it is likely that the classification accuracy would have been lower and not representative of the optimum performance of each approach. For example, a mean ERM quotient of >1.5 has been identified as indicative of a high risk of biological effects (Long *et al*. 2000) yet the high effect threshold providing the best classification accuracy for southern California sediment toxicity was much lower (0.4).

The classification thresholds for toxicity and benthic community condition were nearly identical, indicating that these laboratory and field biological effects measures had similar sensitivity in their response to sediment contamination. This finding differs from other studies, where benthic community responses were reported to be more sensitive. For example, Hyland *et al*. (2003) determined a very high level of risk to the benthos occurred at a mean ERM quotient of >0.361, whereas Long and MacDonald (1998) determined that a mean ERM quotient of >1.5 corresponded to a high risk of toxicity.

Leung *et al*. (2005) determined effect concentrations for cadmium and PAHs using species sensitivity distributions and noted that the cadmium values were much lower than corresponding values based on sediment toxicity; PAH effect concentrations were similar for benthos and toxicity, however.

The apparent disparity in relative sensitivity of toxicity and benthos among studies may be related to several factors. One factor is a difference in the source of the toxicity response thresholds used for comparison. Both Hyland *et al*. and Leung *et al*. compared benthic community effect thresholds that they derived from regional data sets to toxicity effect thresholds derived by others from different data sets and regions. The methods used to derive the toxicity and benthos effect thresholds also varied. The toxicity and benthic community thresholds compared in our study were developed using identical data sets and the same statistical approach. Regional differences in toxicity associations with sediment contamination, possibly due to differences in contaminant mixtures or bioavailability, have been identified in previous studies (Fairey *et al*. 2001, Long *et al*. 2006) and may have confounded threshold comparisons between studies.

Differences in relative sensitivity of toxicity or benthic community SQG thresholds may also be related to variations in test species or benthic community indices. The toxicity data used by Long *et al*. (1998) to identify SQG thresholds is dominated by survival tests using the marine amphipod *Ampelisca abdita*, whereas the southern California analyses were based on the survival of two other species: *Eohaustorius estuarius* and *Rhepoxynius abronius*. Tests of split samples indicate that *A. abdita* is less responsive to sediment contamination than *E. estuarius* or *R. abronius* (Bay *et al.* 2007a), suggesting that the similarity of toxicity and benthic based thresholds in our study may be due to the use of a more sensitive test species. Different benthic indices were also used in each study and their relative sensitivity to contamination may be different, potentially influencing threshold selection.

The CSI approach that we developed in this study represents a new method for interpreting sediment contamination with respect to benthic community impacts. The mean weighted CSI score had the highest classification accuracy of the SQGs evaluated, although the improvement was relatively small compared to the logistic regression models. The better performance of the CSI may have been related to several unique attributes: the approach was developed using southern California data and therefore incorporated regional differences in contaminant mixtures or bioavailability, the approach was calibrated directly to the benthic community response instead of toxicity, and finally, the contribution of individual contaminants to the mean score was weighted in proportion to the strength of association with biological effects.

The results of this suggest several options to increase the relevance of SQGs for predicting benthic community effects. The performance of established approaches such as the ERM can be enhanced through the use of thresholds based on local benthic community response patterns. Alternatively, it is feasible to develop and use SQGs specific to benthic macrofauna data. However, it may not be desirable to abandon existing SQG approaches, as these SQGs provide a link to toxicity, an independent and widely used measure of contaminant effects. Another option is to use multiple SQG approaches to interpret sediment chemistry data. Many sediment quality assessment frameworks integrate multiple indicators of contamination (Chapman *et al*. 2002, McDonald *et al*. 2007) and it may be desirable to use SQGs calibrated to toxicity in combination with SQGs based on benthic community response.

Using multiple SQG approaches, where each is optimized to a different biological effect endpoint, offers several potential advantages that may offset the added complexity. The classification accuracy of each SQG approach should be greater due to better threshold optimization. In addition, the use of multiple SQGs specific to different biological effects would provide additional information that could be helpful in discriminating or prioritizing among sites. Finally, the inclusion of an SQG approach directly linked to the benthos may improve confidence in the interpretation of the chemistry data since the benthic community is one of the primary receptors of concern in most aquatic ecological risk assessments.

LITERATURE CITED

Barrick, R., S. Becker, L. Brown, H. Beller and R. Pastorok. 1988. Sediment quality values refinement: 1988 update and evaluation of Puget Sound AET, Volume 1. PTI Environmental Services. Bellevue, WA.

Bay, S., D. Greenstein and D. Young. 2007a. Evaluation of methods for measuring sediment toxicity in California bays and estuaries. Technical Report 503. Southern California Coastal Water Research Project. Costa Mesa, CA.

Bay, S.M., K.J. Ritter, D.E. Vidal-Dorsch and L.J. Field. 2007b. Comparison of national and regional sediment quality guidelines for predicting sediment toxicity in California. Southern California Coastal Water Research Project. Costa Mesa, CA

Canran, L., P.M. Berry, T.P. Dawson and R.G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28: 385-393.

Chapman, P.M., B.G. McDonald and G.S. Lawrence. 2002. Weight-of-evidence issues and frameworks for sediment quality (and other) assessments. *Human and Ecological Risk Assessment* 8:1489-515.

Cicchetti, D.V. and A.R. Feinstein. 1990. High agreement but low kappa: II: Resolving the paradoxes. *Journal of Clinical Epidemiology* 43: 551-558.

Cohen J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37-46.

Cohen, J. 1968. Weighted Kappa nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin* 70:213-220.

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

Field L.J., D. MacDonald, S.B. Norton, C.G. Severn and C.G. Ingersoll. 1999. Evaluating sediment chemistry and toxicity data using logistic regression modeling. *Environmental Toxicology and Chemistry* 18:1311-1322.

Field, L.J., D.D. MacDonald, S.B. Norton, C.G. Ingersoll, C.G. Severn, D. Smorong and R. Lindskoog. 2002. Predicting amphipod toxicity from sediments using Logistic Regression Models. *Environmental Toxicology and Chemistry* 9:1993-2005.

Fienstein, A.R. and D.V. Cicchetti. 1990. High Agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43:6:543 –549.

Helsel, D. 2005. More than obvious: better methods for interpreting nondetect data. *Environmental Science and Technology* 39:419A-423A.

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. 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.

Hyland, J.L., W.L. Balthis, V.D. Engle, E.R. Long, J.F. Paul, J.K. Summers and R.F. Van Dolah. 2003. Incidence of stress in benthic communities along the U.S. Atlantic and Gulf of Mexico coasts within different ranges of sediment contamination from chemical mixtures. *Environmental Monitoring and Assessment* 81:149-161.

Kraemer, H.C. and D.A. Bloch. 1988. Kappa coefficients in epidemiology: an appraisal of a reappraisal. *Journal of Clinical Epidemeiology* 41:959-968.

Lantz, C.A. and E. Nebenzahl. 1996. Behavior and interpretation of the kappa statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology* 49:431-434.

Leung, K.M.Y., A. Bjorgesater, J.S. Gray, W.K. Li, G.C.S. Lui, U. Wang and P.K.S. Lam. 2005. Deriving sediment quality guidelines from field-based species sensitivity distributions. *Environmental Science and Technology* 39:5148-5156.

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

Long, E.R., L.J. Field and D.D. MacDonald. 1998. Predicting toxicity in marine sediments with numerical sediment quality guidelines. *Environmental Toxicology and Chemistry* 17:714-727.

Long, E.R., D.D. MacDonald, C.G. Severn and C.B. Hong. 2000. Classifying the probabilities of acute toxicity in marine sediments with empirically-derived sediment quality guidelines. *Environmental Toxicology and Chemistry* 19:2598-2601.

Long, E.R., M. Dutch, S. Aasen, K. Welch and M.J. Hameedi. 2005. Spatial extent of degraded sediment quality in Puget Sound (Washington state, USA) based upon measures of the sediment quality triad. *Environmental Monitoring and Assessment* 111:173-222.

Long, E.R., C.G. Ingersoll and D.D. MacDonald. 2006. Calculation and uses of mean sediment quality guideline quotients: A critical review. *Environmental Science and Technology* 40:1726-1736.

MacDonald, D.D., L.M. Di Pinto, L.J. Field, C.G. Ingersoll, E.R. Long and R.C. Swartz. 2000. Development and evaluation of consensus-based sediment effect concentrations for polychlorinated biphenyls (PCB). *Environmental Toxicology and Chemistry* 19:1403-1413.

MacDonald, D.D., C.G. Ingersoll, D.E. Smorong, H. Greening, R. Pribble, T. Janicki, S. Janicki, S. Grabe, B. Sloane, D. Eckenrod, E.R. Long and R.S. Carr. 2004. Development of an ecosystem-based framework for assessing and managing sediment quality conditions in Tampa Bay, Florida. *Archives of Environmental Contamination and Toxicology* 46:147-161.

McDonald, B.G., A.M.H. deBruyn, B.G. Wernick, J. Patterson, N. Pellerin and P. Chapman. 2007. Design and application of a transparent and scalable weight-of-evidence framework: An example from Wabamun Lake, Alberta, Canada. *Integrated Environmental Assessment and Management* 3:476-483.

Neff, J.M., D.J. Bean, B.W. Cornaby, R.M. Vaga, T.C. Gulbransen and J.A. Scanlon. 1986. Sediment quality criteria methodology validation: Calculation of screening level concentrations from field data. Report to US Environmental Protection Agency, Office of Water Regulation and Standards. Washington, DC.

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. 2007. Evaluation of five indicators of benthic community condition in two California bay and estuary habitats. Technical Report 524. Southern California Coastal Water Research Project. Costa Mesa, CA.

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 Program. Westminster, CA.

Stokes, M.E, C.S. Davis and G.G. Koch. 2000. *Categorical Data Analysis using the SAS system*. 2d ed. Cary NC: SAS Institute Inc.

Swartz, R.C. 1999. Consensus sediment quality guidelines for PAH mixtures. *Environmental Toxicology and Chemistry* 18:780-787.

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.

[USEPA] U.S. Environmental Protection Agency. 1994. Methods for assessing the toxicity of sediment-associated contaminants with estuarine and marine amphipods. Washington DC: Office of Research and Development. EPA 600-R94-025.

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.

Vidal, D.E. and S.M. Bay. 2005. Comparative sediment guideline performance for predicting sediment toxicity in southern California, USA. *Environmental Toxicology and Chemistry* 24:3173-3182.

Wenning, R.J., G.E. Batley, C.G. Ingersoll and D.W. Moore, eds. 2005. Use of sediment quality guidelines (SQGs) and related tools for the assessment of contaminated sediments. Pensacola (FL): Society of Environmental Toxicology and Chemistry.

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.

Table 1. Distribution of chemical contaminants in the southern California data set.

Chemical	Units	Percentile 10th	50th	90th	N
Cadmium	mg/kg	0.10	0.34	1.13	441
Copper	mg/kg	14.30	68.8	211	441
Lead	mg/kg	10.4	34.4	101	441
Mercury	mg/kg	0.04	0.16	0.78	441
Zinc	mg/kg	61.7	164	315	441
DDD, total	µg/kg	0.14	22.1	1.93	408
DDE, total	µg/kg	0.17	6.07	144	408
DDT, total	µg/kg	0.13	1.00	9.09	406
alpha-Chlordane	µg/kg	0.30	1.10	4.20	404
gamma-Chlordane	µg/kg	0.43	1.43	6.20	380
HPAH	µg/kg	77.3	339	3723	441
LPAH	µg/kg	19.1	77.1	525	441
PCB, total	µg/kg	6.11	21.1	116	441

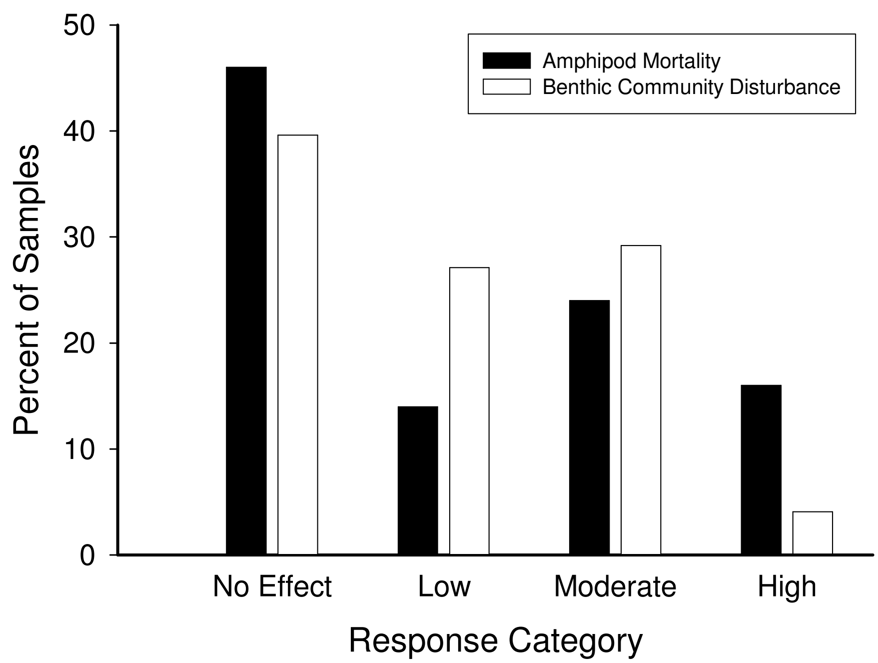

Figure 1. Distribution of sediment toxicity and benthic community disturbance categories in the data set. N=441.

Table 2. Chemical values for individual sediment quality guidelines used for data analyses. Values for the effects range median (ERM) were taken from Long *et al*. 1995. Mean sediment quality guideline quotient (SQGQ1) values taken from Fairey *et al.* 2001. Consensus midpoint effect concentration values taken from Swartz, 1999; MacDonald *et al.* 2000; and Vidal and Bay 2005. Concentrations are on a dry weight basis except where noted.

Chemical	Units	ERM	SQGQ1	Consensus
Arsenic	mg/kg	70.0		55.0
Cadmium	mg/kg	9.6	4.2	5.9
Chromium	mg/kg	370.0		224.9
Copper	mg/kg	270.0	270.0	225.0
Lead	mg/kg	218.0	112.2	222.3
Mercury	mg/kg	0.71		0.6
Nickel	mg/kg	51.6		67.6
Silver	mg/kg	3.7	1.8	3.4
Zinc	mg/kg	410.0	410.0	357.1
2-Methylnaphthalene	µg/kg	670.0		
Acenaphthene	µg/kg	500.0		
Acenaphthylene	µg/kg	640.0		
Anthracene	µg/kg	1,100.0		
Benzo(a)anthracene	µg/kg	1,600.0		
Benzo(a)pyrene	µg/kg	1,600.0		
Chrysene	µg/kg	2,800.0		
Dibenz(a,h)anthracene	µg/kg	260.0		
Dieldrin	µg/kg	8.0	8.0	7.0
Fluoranthene	µg/kg	5,100.0		
Fluorene	µg/kg	540.0		
Naphthalene	µg/kg	2,100.0		
p,p'-DDE	µg/kg			
Phenanthrene	µg/kg	1,500.0		
Pyrene	µg/kg	2,600.0		
Total Chlordane	µg/kg	NA	6.0	
Total DDTs	µg/kg	46.1		25.4
Total PAHs	µg/kg		1,800.0*	1,800.0*
Total PCBs	µg/kg	180.0	400.0	0.47
Tributyltin	µg/kg			

* µg/g organic carbon basis

Table 3. Logistic regression parameters for the regional and national approaches compared in this study. National values are taken from Field *et al.* (2002), CA LRM values are from Bay *et al.* (2007b). B0=intercept; B1=slope; T50=calculated concentration corresponding to a toxicity probability of 0.5. Concentrations are on a dry weight basis.

		LRM			CA LRM		
Chemical	**Units**	**B0**	**B1**	**T50**	**B0**	**B1**	**T50**
Cadmium	mg/kg	-0.34	2.51	1.4	0.29	3.18	0.8
Copper	mg/kg				-5.59	2.59	145
Lead	mg/kg	-5.45	2.77	94	-4.72	2.84	46
Mercury	mg/kg				-0.06	2.68	1.1
Nickel	mg/kg						
Zinc	mg/kg	-7.98	3.34	245	-5.13	2.42	132
1-Methylnaphthalene	µg/kg	-4.14	2.10	94			
1-Methylphenanthrene	µg/kg	-3.59	1.75	112			
2,6-Dimethylnaphthalene	µg/kg	-4.05	1.90	133			
2-Methylnaphthalene	µg/kg	-3.76	1.78	128			
Acenaphthene	µg/kg	-3.62	1.75	116			
Acenaphthylene	µg/kg	-2.96	1.38	140			
Benzo(a)pyrene	µg/kg						
Benzo(b)fluoranthene	µg/kg	-4.54	1.49	1107			
Biphenyl	µg/kg	-4.11	2.21	73			
Chlordane, alpha-	µg/kg				-3.41	4.46	5.8
Chlordane, gamma-	µg/kg						
Chrysene	µg/kg						
Dieldrin	µg/kg	-1.17	2.56	2.9	-1.83	2.59	5.1
Fluoranthene	µg/kg	-4.46	1.48	1034			
Fluorene	µg/kg	-3.71	1.81	114			
HMW PAH	µg/kg				-8.19	2.00	12506
LMW PAH	µg/kg				-6.81	1.88	4127
Naphthalene	µg/kg	-3.78	1.62	217			
Nonachlor trans	µg/kg				-4.26	5.31	6.3
o,p'-DDD	µg/kg						
p,p'-DDD	µg/kg	-1.90	1.49	19			
p,p'-DDT	µg/kg				-3.55	3.26	12
Phenanthrene	µg/kg	-4.46	1.68	455			
Total DDTs	µg/kg						
Total PCBs	µg/kg	-3.46	1.35	368	-4.41	1.48	945

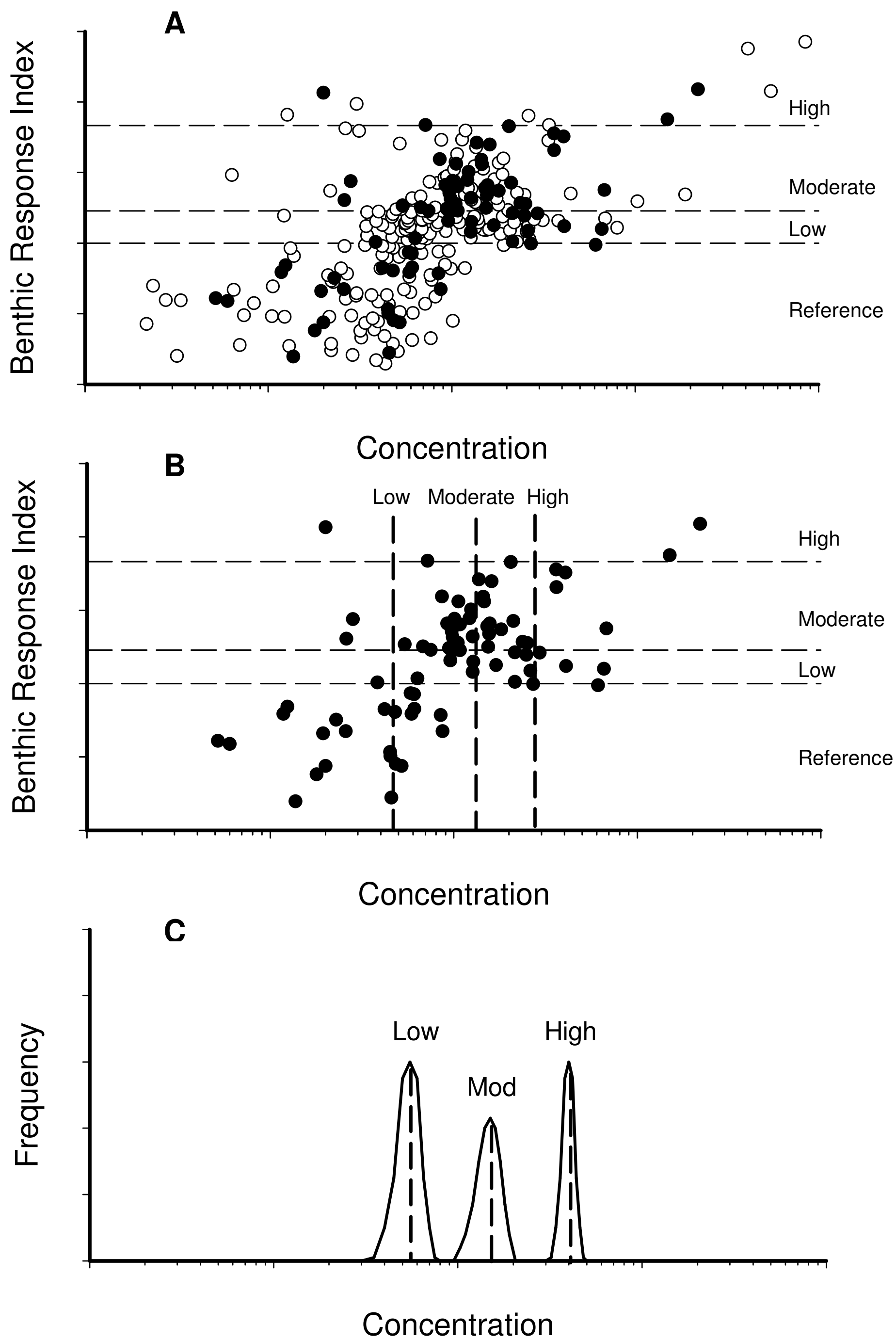

Figure 2. Key steps in the CSI chemical guideline development procedure. A. A random subset of data having an even response distribution (solid circles) is selected for analysis from the development data set. B. The optimum set of guidelines for the subset is determined based on maximizing the kappa statistic. C. The final guidelines are the median of the values from 50 bootstrapped analyses.

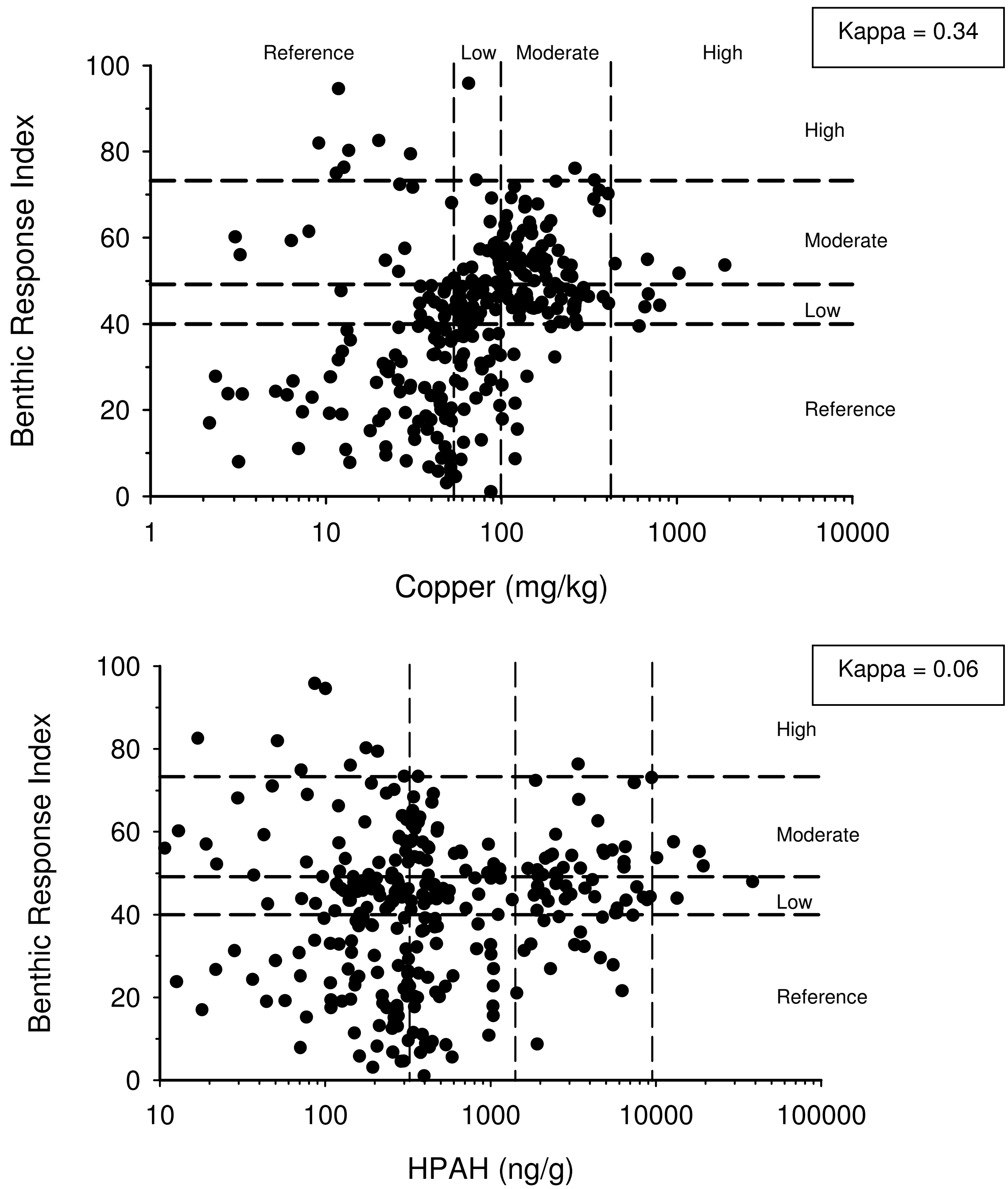

Figure 3. BRI data distributions and CSI guideline values for copper and high molecular weight PAHs.

Table 4. Guideline values and weighting factors used in the calculation of the mean CSI score. Weights are based on the kappa obtained during threshold optimization for each chemical, normalized to the largest kappa value.

		Guideline				
Chemical	Units	Low	Moderate	High	Kappa	Weight
Cadmium	mg/kg	0.09	0.22	1.66	0.13	38
Copper	mg/kg	52.8	96.5	406	0.34	100
Lead	mg/kg	26.4	60.8	154	0.30	88
Mercury	mg/kg	0.09	0.45	2.18	0.10	30
Zinc	mg/kg	112	200	629	0.33	98
DDD, total	µg/kg	0.38	2.69	117	0.16	46
DDE, total	µg/kg	0.11	4.15	153	0.11	31
DDT, total	µg/kg	0.42	1.52	89.3	0.06	16
Alpha Chlordane	µg/kg	0.50	1.23	11.1	0.19	55
Gamma Chlordane	µg/kg	0.54	1.45	14.5	0.20	58
HPAH	µg/kg	312	1325	9320	0.05	16
LPAH	µg/kg	85.4	312	2471	0.02	5
PCB, total	µg/kg	11.9	24.7	288	0.19	55

Table 5. Correlation and classification accuracy of SQG index values with respect to toxicity or benthic community condition. Cells marked with an asterisk are within the 90^{th} percentile of the largest value across 50 bootstrap samplings.

	Toxicity		Benthos	
SQG Approach	Correlation r	Agreement %	Correlation r	Agreement %
CSI[1]	0.54*	50*	0.49*	53*
CA LRM	0.55*	44	0.52*	49*
National LRM	0.45	46*	0.53*	49*
National ERM	0.45	49*	0.47*	40
Consensus	0.53*	47*	0.46	44
SQGQ1	0.42	40	0.47*	36

[1] All thresholds for CSI were based on benthic community condition.

Table 6. SQG index thresholds based on toxicity or benthic community condition. na: not applicable; CSI thresholds were not developed for toxicity.

SQG Approach	Index	Toxicity			Benthos		
		Low	Moderate	High	Low	Moderate	High
CSI	Weighted Mean Score	na	na	na	1.69	2.34	3.00
CA LRM	Maximum Probability	0.42	0.58	0.72	0.43	0.60	0.78
National LRM	Maximum Probability	0.23	0.44	0.61	0.22	0.43	0.65
National ERM	Mean Quotient	0.06	0.12	0.38	0.06	0.13	0.40
Consensus	Mean Quotient	0.14	0.26	0.60	0.15	0.30	0.68
SQGQ1	Mean Quotient	0.16	0.34	0.80	0.13	0.37	1.26