Comparison of national and regional sediment quality guidelines for classifying sediment toxicity in California

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

Abstract

A number of sediment quality guidelines (SQGs) have been developed for relating chemical concentrations in sediment to their potential for biological effects, but there have been few studies evaluating the relative effectiveness of different SQG approaches. Here we apply six SQG approaches to assess how well they predict toxicity in California sediments. Four of the SQG approaches were nationally derived indices that were established in previous studies: Effects Range Median (ERM), Logistic Regression Model (LRM), Sediment Quality Guideline Quotient 1 (SQGQ1), and Consensus. Two approaches were variations of nationally derived approaches that were recalibrated to California-specific data (CA LRM and CA ERM). Each SQG approach was applied to a standardized set of matched chemistry and toxicity data for California and an index of the aggregate magnitude of contamination (e.g., mean SQG quotient or maximum probability of toxicity) was calculated. A set of three thresholds for classification of the results into four categories of predicted toxicity was established for each SQG approach using a statistical optimization procedure. The performance of each SQG approach was evaluated in terms of correlation and categorical classification accuracy. Each SQG index was significantly correlated with toxicity and able to correctly classify the level of toxicity for up to 40% of samples. The CA LRM had the best overall performance, but the magnitude of differences in classification accuracy among the SQG approaches was relatively small. Recalibration of the indices using California data improved performance of the LRM, but not the ERM. The LRM approach is more amenable to revision than other national SQGs, which is a desirable attribute for use in programs where the ability to incorporate new information or chemicals of concern is important. The use of a consistent threshold development approach appeared to be a more important factor than type of SQG approach in determining SQG performance. The relatively small change in classification accuracy obtained with regional calibration of these SQG approaches suggests that further calibration and normalization efforts are likely to have limited success in improving classification accuracy associated with biological effects.

Introduction

Many monitoring programs are conducted to evaluate chemical contamination effects on sediment quality, but interpreting these data is often difficult (Wenning *et al*. 2005). The biological availability of chemicals in sediments is complex and not entirely understood. Moreover, the chemicals are often present in complex mixtures for which joint effect is difficult to predict.

A number of SQGs for relating chemical concentrations to potential for biological effects have been developed, generally falling into two classes. The first is a mechanistic approach, which models the chemical and biological processes that affect contaminant bioavailability. Current mechanistic SQGs are based on equilibrium partitioning theory and apply to selected classes of contaminants, primarily divalent metals and several types of nonionic organics United States Environmental Protection Agency (USEPA; 2003, 2005a, 2008). While these models are useful for describing contaminant bioavailability, mechanistic SQGs address the question of what may be causing sediment toxicity, not whether or not a sediment will be toxic. In addition, some of the

[1] *National Oceanic and Atmospheric Administration, Seattle, WA*

parameters needed to apply these guidelines (e.g., sediment acid volatile sulfides and simultaneously extracted metals) are rarely collected in current routine monitoring programs.

Second, the more widely used empirical SQGs are derived from statistical association of matched sediment chemistry and biological effects data. Multiple collections of empirical SQGs that are based on different statistical approaches have been developed. Examples of empirical SQG approaches for the marine environment include ERM, Probable Effects Level (PEL), Apparent Effects Threshold (AET), SQGQ1, and LRM (Barrick *et al*. 1988, Fairey *et al*. 2001, Field *et al*. 2002, Long *et al*. 1995, MacDonald *et al*. 1996). Consensus guidelines, which aggregate several different SQGs having a similar narrative intent (e.g., median effect), are an evolution of the empirical approach. Marine consensus SQGs have been developed for some constituents, including metals, polychlorinated biphenyls (PCBs), and polycyclic aromatic hydrocarbons (PAHs: MacDonald *et al*. 2000, Swartz 1999, Vidal and Bay 2005).

It is unclear which empirical SQG approach is most effective for describing the potential for biological effects associated with chemical contamination. Numerous studies have shown that each SQG approach has predictive ability with respect to biological effects, but most studies have generally been limited to examination of just one or two approaches and often use variable methods to measure performance (Wenning *et al*. 2005). Long *et al*. (2000) applied ERMs and PELs to several data sets and observed different patterns in predictive ability. Vidal and Bay (2005) compared five SQG approaches using a common data set and found large differences in predictive ability among some approaches, however their study did not include the LRM approach. Vidal and Bay (2005) also observed that comparisons of SQG performance can be strongly influenced by the selection of thresholds used to classify the results. Existing studies are inadequate for comparing the performance of empirical SQGs because of their limited scope, lack of comparability in methods, and lack of thresholds derived using a consistent methodology.

It is also unclear whether performance of SQGs is improved when they are calibrated to local conditions. The predictive ability of SQGs to biological effects has been shown to vary when the same guidelines are applied to data from different regions (Fairey *et al*. 2001, Long *et al*. 1998, Long *et al*. 2006, O'Connor *et al*. 1998, Vidal and Bay 2005). These variations in performance may be due to differences in the nature of the chemical mixtures between sites or regions, variations in bioavailability due to geochemical factors, or differences in the sensitivity of methods used to measure biological effects. Variation in SQG performance among studies creates uncertainty in determining the threshold of SQG exceedance associated with adverse impacts on sediment quality. The use of SQGs and interpretation thresholds that are derived or calibrated relative to site-specific conditions has been recommended as a way to reduce the uncertainty of SQG interpretation (Fairey *et al*. 2001, Long *et al*. 2006, Vidal and Bay 2005).

This study applied six SQG approaches to a large California data set of paired chemistry and toxicity measurements to assess: 1) which national SQG approach best classifies the toxicity of California sediments, 2) whether the relationship of national SQGs to sediment toxicity is improved when the SQGs are recalibrated to California data, and 3) if performance further improves when the SQGs are further recalibrated to two subregions within California.

METHODS

The study assessed the performance of six SQG approaches by applying them to matched chemistry and toxicity data for California, calculating an index of overall contamination based on the mean SQG quotient or the maximum probability of toxicity, and determining the correlation and categorical classification accuracy (Figure 1). Four of the SQG approaches were derived in previous national studies (ERM, LRM, SQGQ1, Consensus) and two were variations of nationally derived SQGs that were recalibrated to California-specific data (CA LRM and CA ERM). Thresholds relating each SQG index to toxicity response categories were derived using a standardized statistical approach. Each SQG index was evaluated by determining three measures of association between the calculated effect categories and the observed toxicity response: correlation, weighted kappa, and percent agreement. SQG calibration and performance evaluations were conducted at two scales in order to investigate the influence of regional variations in sediment characteristics: statewide (all California data) and regional (separate northern and southern California data sets).

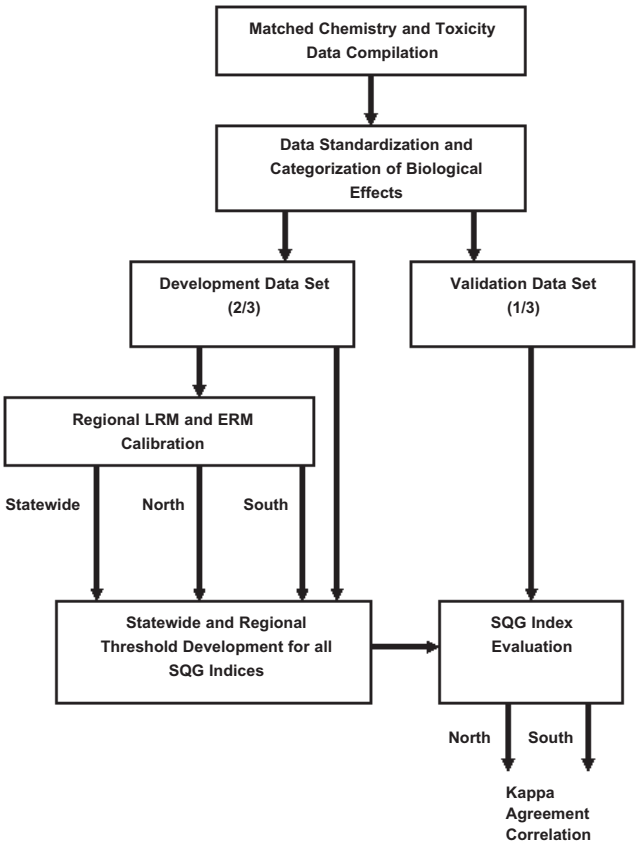

Figure 1. Schematic of data analyses.

Data

Paired chemistry and sediment toxicity measurements from California marine embayments were compiled from 151 dredging, monitoring, and research studies conducted between 1984 and 2004. The database included stations from marine and estuarine embayments located from 41.94°N (Del Norte County, CA, USA) to 31.75°N (USA-Mexico international border). More information on the studies used to populate this database can be found at http://www.sccwrp.org/view.php?id=519.

The data were screened to select information that was of high quality and comparable. All stations were from locations in enclosed bays or harbors at subtidal depths and only data from surficial sediment (top 30 cm or less) were selected. Toxicity data were limited to information from solid phase 10-day amphipod survival tests using *Rhepoxynius abronius* or *Eohaustorius estuarius* and conducted using standardized methods (USEPA 1994). Toxicity data were further screened to ensure mean control survival was >85% and overlying water ammonia concentrations were below species-specific criteria (USEPA 1994). Screening steps to select chemistry data for analysis included a review of the data quality assessment from 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, dichloro diphenyl trichlorethanes (DDTs), PCBs, and chlordanes were calculated using a consistent methodology for all samples. Low molecular weight PAHs (LMW PAH) 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 (HMW PAH) 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 LMW PAH and HMW PAH values. Total PCBs was calculated from the sum of congeners 8, 18, 028, 44, 52, 66, 101, 105, 110, 118, 128, 138, 153, 180, 187, and 195. The congener list was a subset of that used by the NOAA Status and Trends program; the sum was multiplied by a correction factor of 1.72 to approximate the value obtained using the larger NOAA list. 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, Cary, NC) 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. Estimated values were constrained to always be less than the study reporting limit. The imputation method was also used to estimate total organic carbon (TOC) for the purposes of calculating the

SQGQI and Consensus quotients. Estimated values were not used in calculations for any other analytes missing in the data sets, except when needed to calculate standardized sums of PAHs, PCBs, or pesticides. For example, a value for phenanthrene was estimated for a sample that contained data for other PAHs in order to use the standardized method to calculate the PAH sums, but the estimated phenanthrene value was not used individually to calculate summary SQG values for that sample.

The standardized data set was divided into two groups to facilitate investigation of regional differences in chemical contamination on SQG performance: northern California embayments north of Point Conception and southern California embayments south of Point Conception. Each regional data set was further divided into two portions: a calibration subset used for index development and threshold calibration, and an independent validation subset used for the analysis of SQG performance. Approximately one third of the data were used for validation. The validation samples were selected by first grouping the data into one of eight subregions based on latitude to ensure even spatial representation. The samples within each subregion were then ranked by the mean Effects Range Median quotient (mERMq) and one third of the samples systematically sampled from throughout the mERMq distribution. Additional validation data were obtained from recent monitoring studies that were not included in the initial data compilation effort. The north and south validation data sets contained 146 and 249 samples, respectively.

National SQGs

The ERM values used in the analyses were obtained from Long *et al*. (1995). The 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 is shown in Table 1 and is 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 1.

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 1) and averaging the individual quotients.

The LRM approach was based on the statistical analysis of paired chemistry and amphipod toxicity data from studies throughout the United States (US; 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 2). The maximum probability (P_{max}) for each sample was used as the index of overall contamination.

Regional SQGs

Regionally calibrated versions were developed for two of the national SQG approaches: ERM and LRM. Regional versions were not developed for the other national SQG approaches (SQGQ1 and Consensus) because these approaches are based on the inclusion of SQG values from other sources and cannot be easily recalibrated with new data. Three versions of each regional SQG approach were developed: a statewide version that was calibrated to data from throughout California (CA ERM or CA LRM), and two region-specific versions. The region-specific versions were calibrated separately for the northern California (NorCA ERM or NorCA LRM) and southern California (SoCA ERM or SoCA LRM) data sets.

For the CA ERM variations, local calibration involved calculation of new individual chemical ERM values. The data were screened to select toxic samples (>20% mortality) with chemical concentrations >2x median concentration of nontoxic samples. A separate screening process was used for each chemical. After screening, the data were sorted in ascending order and the median concentration of

Table 1. 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	CA ERM	SoCA ERM	NorCA ERM	SQGQ1	Consensus
Arsenic	mg/kg	70.0	19.2	19.1			55.0
Cadmium	mg/kg	9.6	1.0	1.2	0.6	4.2	5.9
Chromium	mg/kg	370.0	154.0	110.0	291.0		224.9
Copper	mg/kg	270.0	151.0	208.0	91.2	270.0	225.0
Lead	mg/kg	218.0	87.4	94.5	56.4	112.2	222.3
Mercury	mg/kg	0.7	0.8	0.8	0.7		0.6
Nickel	mg/kg	51.6	83.5	42.0			67.6
Silver	mg/kg	3.7	0.9	1.1	0.4	1.8	3.4
Zinc	mg/kg	410.0	332.5	406.9	214.5	410.0	357.1
2-Methylnaphthalene	µg/kg	670.0	22.2	23.6	20.2		
Acenaphthene	µg/kg	500.0	23.0	24.5	19.0		
Acenaphthylene	µg/kg	640.0	26.0	47.0	19.8		
Anthracene	µg/kg	1100.0	130.0	215.5	60.8		
Benzo(a)anthracene	µg/kg	1600.0	356.6	540.0	169.5		
Benzo(a)pyrene	µg/kg	1600.0	405.5	630.0	225.3		
Chrysene	µg/kg	2800.0	577.0	739.9	239.0		
Dibenz(a,h)anthracene	µg/kg	260.0	94.4	130.0	23.4		
Dieldrin	µg/kg	8.0	2.0	2.0	0.8	8.0	7.0
Fluoranthene	µg/kg	5100.0	432.3	723.0	410.9		
Fluorene	µg/kg	540.0	30.7	46.2			
Naphthalene	µg/kg	2100.0	34.4	33.4	42.5		
p,p'-DDE	µg/kg		25.9	38.3	3.8		
Phenanthrene	µg/kg	1500.0	267.5	275.9	310.6		
Pyrene	µg/kg	2600.0	534.8	1000.0	480.0		
Chlordane, total	µg/kg		17.2	23.1	4.0	6.0	
DDTs, total	µg/kg	46.1	49.3	60.0	13.1		25.4
PAH, total	mg/kg					1,800.0*	1,800.0*
PCB, total	µg/kg	180.0	111.5	125.4	21.3	400.0	0.5
Tributyltin	µg/kg		202.0	308.0	30.0		

* µg/g organic carbon basis

each chemical was selected as the region-specific ERM value. ERM values were calculated for all chemicals having >10 records in the screened data set. This resulted in calculating CA ERM and SoCA ERM values for 27 chemicals, and NorCA ERM values for 25 chemicals (Table 1).

California logistic regression models for individual chemicals were developed for the statewide and regional California data sets using the methods described in USEPA (2005b). These models were applied to the California calibration data using <80% control adjusted amphipod survival as the definition of a toxic sample. The specific models included in the CA LRM, SoCA LRM, and NorCA LRM approaches were selected from a library of candidate models that included national models, as well as models derived using the California data sets. The selected models were chosen based on the suitability of fit with the observed probability of toxicity (Table 2). Models with high false positive rates were not included.

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.

Three thresholds, defining four ranges of SQG

Table 2. Logistic Regression parameters for the regional and national models compared in this study. National values were taken from Field *et al*. (2002). B0 = intercept; B1 = slope; T50 is the calculated concentration corresponding to a toxicity probability of 0.5. Concentrations are on a dry weight basis.

Chemical	Units	LRM			CA LRM			SoCA LRM			NorCA LRM		
		B0	B1	T50	B0	B1	T50	B0	B1	T50	B0	B1	T50
Cadmium	mg/kg	-0.3	2.5	1.4	0.3	3.2	0.8	0.3	3.2	0.8	1.5	3.4	0.4
Copper	mg/kg				-5.6	2.6	145.0	-6.8	2.8	268.0	-6.6	3.8	51.0
Lead	mg/kg	-5.5	2.8	94.0	-4.7	2.8	46.0	-8.6	4.8	62.0			
Mercury	mg/kg				-0.1	2.7	1.1				1.7	3.1	0.3
Nickel	mg/kg							-8.5	5.7	30.0			
Zinc	mg/kg	-8.0	3.3	245.0	-5.1	2.4	132.0	-10.0	4.2	234.0	-13.8	6.9	100.0
1-Methylnaphthalene	µg/kg	-4.1	2.1	94.0									
1-Methylphenanthrene	µg/kg	-3.6	1.8	112.0									
2,6-Dimethylnaphthalene	µg/kg	-4.1	1.9	133.0									
2-Methylnaphthalene	µg/kg	-3.8	1.8	128.0									
Acenaphthene	µg/kg	-3.6	1.8	116.0									
Acenaphthylene	µg/kg	-3.0	1.4	140.0									
Benzo(a)pyrene	µg/kg										-2.3	1.2	80.0
Benzo(b)fluoranthene	µg/kg	-4.5	1.5	1107.0							-4.6	2.3	90.0
Biphenyl	µg/kg	-4.1	2.2	73.0									
alpha-Chlordane	µg/kg				-3.4	4.5	5.8	-3.4	4.5	5.8			
gama-Chlordane	µg/kg							-3.6	4.2	7.4			
Chrysene	µg/kg										-2.5	1.3	95.0
Dieldrin	µg/kg	-1.2	2.6	2.9	-1.8	2.6	5.1	-1.2	4.3	2.0			
Fluoranthene	µg/kg	-4.5	1.5	1034.0									
Fluorene	µg/kg	-3.7	1.8	114.0									
HMW PAH	µg/kg				-8.2	2.0	12506.0	-8.2	2.0	12506.0	-4.3	1.5	785.2
LMW PAH	µg/kg				-6.8	1.9	4127.0	-6.8	1.9	4127.0	-3.4	1.5	185.2
Naphthalene	µg/kg	-3.8	1.6	217.0									
trans-Nonachlor	µg/kg				-4.3	5.3	6.3	-4.3	5.3	6.3			
o,p'-DDD	µg/kg							-2.0	3.3	4.1	1.1	2.0	0.3
p,p'-DDD	µg/kg	-1.9	1.5	19.0				-1.8	2.0	7.6	-0.8	2.5	2.0
p,p'-DDT	µg/kg				-3.6	3.3	12.0	-1.5	1.6	8.1	-0.6	3.3	1.5
Phenanthrene	µg/kg	-4.5	1.7	455.0									
DDTs, total	µg/kg										-1.3	2.8	3.0
PCB, total	µg/kg	-3.5	1.4	368.0	-4.4	1.5	945.0	-4.4	1.5	945.0	-4.4	1.5	945.0

index results, were established for each SQG approach. Each SQG index range corresponded to 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*. 2007). The categories for *E. estuarius* were: Nontoxic (≥90% survival), Low Toxicity (82 - 89%), Moderate Toxicity (59 - 81%) and High Toxicity (<59%). The categories for *R. abronius* were: Nontoxic (>90% survival), Low Toxicity (83 - 89%), Moderate Toxicity (70 - 82%) and High Toxicity (<70%).

The thresholds were selected using a statistical optimization procedure based on maximizing overall agreement between the SQG index and toxicity categories in subsets of the calibration data set. The optimal set of thresholds was selected by computing the percent agreement for a large set of possible candidates. These candidates were selected by choosing all permutations of three thresholds, taken at 5% increments of the range of index values. In addition, distances between individual guidelines within each set were constrained to be no less than 10% of the range of index values. For each data subset, the set of three guidelines that yielded the highest percent agreement was selected.

The threshold optimization procedure was repeated (i.e., bootstrapping) 50 times on subsets of the data that contained an even distribution of samples across toxicity categories. This step was included in order to minimize the influence on the optimization results of the skewed sample distribution in the calibration data set, which contained a higher

proportion of nontoxic and low toxicity samples. Each threshold selection data set contained 30 randomly chosen calibration samples from each toxicity category. The median of the thresholds across all subsets was selected as the final thresholds for that SQG approach.

Evaluation of SQG Performance

SQG performance was evaluated by quantifying the strength of association between chemistry and toxicity 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 or P_{max}) and percent amphipod mortality. Analyses of categorical classification accuracy were based on the frequency with which the SQG index category (determined by applying the thresholds derived from the calibration data set) correctly predicted the measured toxicity response category. All analyses were conducted using an independent validation data set that was not used for threshold development.

Two measures of classification accuracy were calculated: percent agreement and weighted kappa. Percent agreement is the number of samples that are correctly classified, calculated as:

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

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

The weighted kappa statistic (Cohen 1960, 1968) is also a measure of agreement between the SQG predictions and toxicity, but differs in that a correction for chance is applied and partial credit is given according to the magnitude of disagreement. Kappa weights were based on the linear weighting scheme of Cicchetti-Allsion (1971); a weight of 1 was assigned to cases of perfect agreement and weights of 1/3, 1/6, and 0 assigned to disagreements of one, two, or three toxicity categories, respectively. SAS PROC FREQ (SAS Institute Inc, Cary, NC) was used to calculate the weighted kappa (Stokes *et al*. 2000).

A bootstrap resampling approach similar to that used for threshold development was also used in calculation of the correlation, percent agreement, and weighted kappa values. The reported correlation and classification accuracy values are the median of 50 resamples. The 90th percentile confidence limits of the bootstrapped results were used to identify the best performing SQG approaches with respect to correlation and classification accuracy. The approach having the highest values for both correlation and classification accuracy was selected as the best performing SQG. Correlation results were given greater weight when the rankings were variable among the performance measures in order to minimize the influence of threshold selection.

Results

Different patterns of sediment contamination were apparent between the northern and southern California data sets (Table 3), reflecting different anthropogenic inputs and geology. Median concentrations of most PAH compounds, chromium, and nickel were greatest in the north, while the south data set contained higher concentrations of chlordane, copper, DDTs, PCBs, and zinc. The southern California data set usually contained the highest concentrations of each contaminant, which may reflect the larger south data set. An exception was the presence of higher chromium and nickel concentrations in the north data set, which was likely due to higher naturally occurring concentrations of these elements in northern California soils.

There was a similar range and distribution of sediment toxicity between the northern and southern California data sets (Figure 2). The distribution of the data was skewed towards low toxicity; approximately 60% of the samples in each region had less than 20% mortality and less than 10% had greater than 60% mortality.

There were large differences in the number of chemicals and their threshold concentrations included in the different SQG indices (Tables 1 and 2). The number of chemicals varied from 9 for the SQGQ1 to 25 for the mERMQ. Individual chemical concentrations for the ERM, SQGQ1, and Consensus SQGs were similar because these values were often derived from similar sources. There were often large differences in individual chemical concentration between the national and region-specific versions of the ERM. This was especially evident for PAH compounds, where the national ERM values were 1 - 2 orders of magnitude greater than the CA ERMs (Table 1).

Table 3. Cumulative distribution of sediment chemistry data for the California samples used in the analysis.

Chemical	Units	Northern California			Southern California		
		N	50th Percentile	90th Percentile	N	50th Percentile	90th Percentile
2-Methylnaphthalene	µg/kg	367	10.6	27.2	713	9.6	49.1
Acenaphthene	µg/kg	407	6.0	21.2	674	5.1	46.0
Acenaphthylene	µg/kg	398	8.2	24.3	671	6.2	79.0
Anthracene	µg/kg	422	20.2	91.1	771	18.0	370.0
Arsenic	mg/kg	393	8.5	12.9	828	8.6	17.3
Benz(a)anthracene	µg/kg	427	63.8	189.0	838	44.9	720.0
Benzo(a)pyrene	µg/kg	430	95.7	289.0	845	65.9	1100.0
Cadmium	mg/kg	420	0.2	0.4	850	0.4	1.4
Chlordanes, total	µg/kg	404	0.8	3.3	816	7.1	34.3
Chromium	mg/kg	329	122.0	245.0	851	56.0	95.0
Chrysene	µg/kg	427	72.0	229.0	847	64.0	1090.0
Copper	mg/kg	405	40.1	65.5	851	76.5	252.0
DDTs, total	µg/kg	404	3.6	12.4	816	21.4	112.0
Dibenz(a,h)anthracene	µg/kg	412	12.1	32.5	787	19.1	230.0
Dieldrin	µg/kg	368	0.2	0.9	297	1.0	3.4
Fluoranthene	µg/kg	425	151.0	423.0	849	89.9	1320.0
Fluorene	µg/kg	414	9.3	34.4	708	6.9	77.5
Lead	mg/kg	409	21.2	37.8	851	35.9	101.0
Mercury	mg/kg	430	0.3	0.4	843	0.2	0.9
Naphthalene	µg/kg	365	20.9	51.2	733	9.4	44.3
Nickel	mg/kg	399	84.0	114.6	838	20.7	36.6
PCB, total	µg/kg	351	7.9	32.0	851	24.8	196.2
Phenanthrene	µg/kg	392	75.4	242.0	815	39.8	429.0
Pyrene	µg/kg	427	190.0	520.0	850	102.0	1500.0
Silver	mg/kg	418	0.2	0.5	839	0.4	1.4
PAH, total	µg/kg	431	945.0	2492.0	851	619.0	8573.0
Zinc	mg/kg	409	110.0	164.0	851	180.0	369.0

The categorization thresholds for the SQGs varied geographically (e.g., statewide, north, south). The highest thresholds were usually obtained for southern California data, but the differences were typically small (Table 4). The SQGQ1 was an exception, having nearly a three-fold difference between thresholds derived using northern and southern California data.

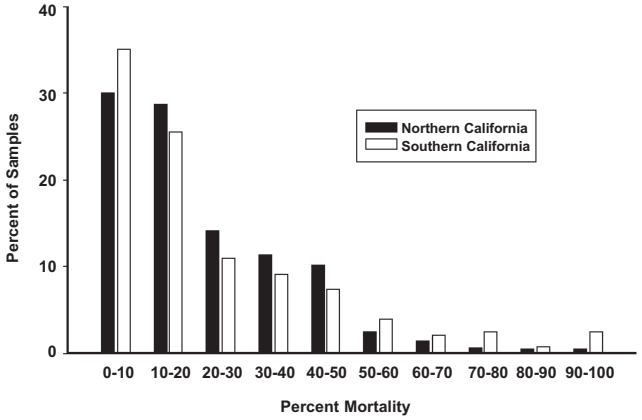

Figure 2. Distribution of sediment toxicity data (10-day amphipod mortality) for the California samples used in the analysis.

Each of the statewide-calibrated SQG approaches correlated significantly with amphipod survival when applied to statewide validation data. Spearman correlation coefficients ranged from 0.35 to 0.16 (Table 5), with the CA LRM having the highest correlation. Correlations generally increased when the indices were evaluated using the separate north and south data sets, though CA LRM performed best in both habitats (Table 6).

The CA LRM (Table 5) also performed best with respect to classification accuracy, when the indices were applied to the statewide data set. Very little improvement in classification accuracy was obtained using the CA ERM approach, relative to the national ERM approach. While both measures of classification accuracy ranked the SQG approaches similarly, the weighted kappa statistic provided a greater degree of discrimination among

Table 4. Thresholds used for evaluations of SGQ index classification accuracy. Nontoxic: <Low threshold; Low Toxicity: Low threshold to <Moderate threshold; Moderate Toxicity: Moderate threshold to <High threshold; High toxicity: >High threshold.

SQG Approach	Index	Low Threshold			Moderate Threshold			High Threshold		
		North	South	State	North	South	State	North	South	State
National ERM	Mean Quotient	0.08	0.06	0.07	0.15	0.12	0.13	0.29	0.38	0.33
National LRM	Maximum Probability	0.17	0.23	0.20	0.26	0.44	0.35	0.50	0.61	0.55
Consensus	Mean Quotient	0.15	0.14	0.14	0.23	0.26	0.25	0.51	0.60	0.55
SQGQ1	Mean Quotient	0.06	0.16	0.10	0.11	0.34	0.19	0.33	0.80	0.52
CA LRM	Maximum Probability	0.25	0.42	0.34	0.42	0.58	0.50	0.62	0.72	0.67
CA ERM	Mean Quotient	0.15	0.14	0.15	0.23	0.25	0.24	0.68	1.28	0.93

approaches than did percent agreement.

When the SQG indices and statewide thresholds were evaluated relative to the regional data sets, the CA LRM was the only approach with consistently high classification accuracy and correlations (Table 6). The CA ERM also had relatively high classification accuracy for northern California data and high classification accuracy was also obtained for the ERM and Consensus for southern California.

Developing thresholds on a regional basis had little effect overall. Percent agreement scores across indices were almost identical between thresholds developed using statewide and regional data sets (Table 6). However, classification accuracy (weighted kappa) was improved for the worst performing SQG approaches, such as SQGQ1 in the south and national LRM in the north.

Increased classification accuracy was obtained for the region-calibrated SQGs in the north (NorCA LRM and NorCA ERM) compared to statewide-calibrated versions (Table 6). However, no improvement was measured for the approaches that were calibrated to southern California data (SoCA LRM and SoCA ERM).

Table 5. Nonparametric Spearman correlation (r) and classification accuracy of statewide SQG approaches with amphipod mortality. Values in the shaded cells are within the 90th percentile of the highest median value for the bootstrapped analyses. Analyses were conducted on the combined data for the north and south validation data sets and used thresholds developed using the statewide data set.

Region	Approach	Weighted Kappa	% Agreement	r
State	CA LRM	0.23	37	0.35
State	National ERM	0.17	32	0.25
State	Consensus	0.17	31	0.25
State	National LRM	0.15	35	0.22
State	CA ERM	0.17	33	0.20
State	SQGQ1	0.12	32	0.16

Discussion

While the P_{max}, based on the CA LRM, was the best-performing SQG index, there was relatively little difference in performance among many of the indices. This differs from the findings of Vidal and Bay (2005) and probably results from using thresholds that were selected using a consistent methodology and calibration data set. The standardized thresholds allowed each SQG approach to be evaluated on a level playing field, so that differences in performance could be compared without the confounding effect of differences in threshold selection.

Two of the SQG approaches were recalibrated using California data, which had mixed effects. For the CA LRM, there was a substantive improvement in performance, but performance of the mean quotients based on the CA ERM, was comparable to that of the national mERMQ. This may have resulted from differences in the SQG calibration process. The CA ERMs consisted of entirely of new values that were derived from the California data set. All available CA ERMs were used in the quotient calculations. In contrast, for the CA LRM, the set of

Table 6. Classification accuracy and Spearman correlation of SQG approaches applied to data from each region separately. Values in the shaded cells are within the 90th percentile of the highest median value of the bootstrapped analyses. Analyses were conducted separately using thresholds developed with statewide and region-specific data sets.

Approach	Northern California			Southern California		
	Weighted Kappa	% Agreement	r	Weighted Kappa	% Agreement	r
Statewide Thresholds						
CA LRM	**0.20**	**38**	**0.39**	**0.25**	**35**	**0.42**
National ERM	0.12	27	**0.31**	**0.21**	**38**	0.28
Consensus	0.12	28	0.23	**0.22**	**36**	**0.31**
National LRM	0.11	**35**	0.18	0.18	**34**	**0.33**
CA ERM	**0.21**	**33**	0.22	0.15	**34**	0.18
SQGQ1	0.13	**35**	0.25	0.10	28	0.26
Region-specific Thresholds						
CA LRM	**0.16**	27	**0.39**	**0.28**	**40**	**0.42**
National ERM	**0.17**	**30**	**0.31**	**0.22**	**38**	0.28
Consensus	**0.15**	**29**	0.23	**0.25**	**39**	**0.31**
National LRM	**0.20**	**33**	0.15	**0.22**	**36**	**0.33**
CA ERM	**0.21**	**33**	0.22	0.13	33	0.18
SQGQ1	**0.21**	**33**	0.25	0.18	33	0.26
Nor/SoCA LRM	**0.21**	**33**	**0.27**	**0.22**	**36**	**0.37**
Nor/SoCA ERM	**0.20**	**35**	**0.22**	0.18	**35**	0.18

models used for evaluation was selected from a combination of national and California derived models. This selection process was based on increasing model goodness of fit and reducing false positives. It is possible that this additional selection step improved the predictive ability of the CA LRM. A similar selection process was not used for the CA ERM because of differences in derivation methodology compared to the national ERMs, which were based on multiple types of toxicity tests and other biological response values (Long *et al*. 1995).

The improved performance of the CA LRM may also have been due to differences in the composition, magnitude, and bioavailability of sediment contamination in the California data, relative to the data used for national LRM development. Regional differences in contamination and geochemistry have been identified as important factors affecting the predictive accuracy of SQGs (Long *et al*. 2000, Wenning *et al*. 2005). Since the values used in empirical SQG approaches are derived from chemistry-toxicity relationships in the development data set, regionally calibrated approaches would be expected to have greater predictive accuracy.

The regional SQG results suggest that further improvement in SQG performance could be obtained through further site-specific normalization or the use of mechanistic SQGs. Normalization of the organics data to total organic carbon and metals data to a reference element (iron) and use of USEPA equilibrium partitioning sediment benchmarks were also evaluated in preliminary phases of this study, but did not result in any improvement in correlation or classification accuracy.

Use of thresholds calibrated to the north and south subregions produced only small increases in performance relative to the statewide thresholds. The relatively small differences in regional performance are probably related to the heterogeneous nature of sediment contamination. Even though there are differences in overall pattern and magnitude of contamination in the northern and southern California data sets, contamination patterns within each region is highly diverse due to the presence of multiple waterbodies and contaminant inputs from a multitude of sources. The limited improvement in classification accuracy obtained with regional calibration suggests that further calibration and normalization

efforts are likely to have limited success in improving the association of empirical SQG indices with biological effects. This limitation of SQGs for interpreting the biological significance of sediment contamination is well known and underscores the importance of using these approaches in combination with biological measures in a multiple lines of evidence approach (Wenning *et al*. 2005).

Because the performance difference among SQG indices was small, characteristics such as history of use, ease of application, types of chemicals included in the constituent array, and feasibility for revision should be considered when selecting the SQG approach to be used. For instance, the Consensus and SQGQ1 approaches incorporate a lesser number of chemicals than the other approaches and it is difficult to add new contaminants of concern because the SQGs are dependent on the availability of values from other sources. Local calibration is also not feasible for these approaches for the same reason.

The best performing index, CA LRM, is highly amenable to revision as demonstrated by this study. But LRM approaches are also the most difficult to apply and interpret because a complex set of regressions must be used to determine probabilities of toxicity, rather than comparing chemistry data to a simple table of SQG values. These difficulties can be overcome by incorporating the regression calculations into spreadsheets or other data analysis tools and establishing thresholds for interpreting the P_{max} values.

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. Bellvue, WA.

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

Cicchetti, D.V. and T. Allison. 1971. A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology* 11:101-109.

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.

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

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

MacDonald, D.D., R.S. Carr, F.D. Calder, E.R. Long and C.G. Ingersoll. 1996. Development and evaluation of sediment quality guidelines for Florida coastal waters. *Ecotoxicology* 5:253-278.

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.

O'Connor, T.P., K.D. Daskalakis, J.L. Hyland, J.F. Paul and J.K. Summers. 1998. Comparisons of sediment toxicity with predictions based on chemical guidelines. *Environmental Toxicology and Chemistry* 17:468-471.

Stokes, M.E., C.S. Davis and G.C. Koch. 2000. Categorical Data Analysis using the SAS system. Second Edition. SAS Institute Inc. Cary, NC.

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

United States Environmental Protection Agency (USEPA). 1994. Methods for assessing the toxicity of sediment-associated contaminants with estuarine and marine amphipods. EPA 600-R94-025. USEPA, Office of Research and Development. Washington, DC.

USEPA. 2003. Procedures for the derivation of equilibrium partitioning sediment benchmarks (ESBs) for the protection of benthic organisms: PAH mixtures. EPA-600-R-02-013. USEPA, Office of Research and Development. Washington, DC.

USEPA. 2005a. Procedures for the derivation of equilibrium partitioning sediment benchmarks (ESBs) for the protection of benthic organisms: Metal mixtures (cadmium, copper, lead, nickel, silver, and zinc). EPA-600-R-02-011. USEPA, Office of Research and Development. Washington, DC.

USEPA. 2005b. Predicting toxicity to amphipods from sediment chemistry (Final Report). EPA/600/R-04/030. USEPA, ORD National Center for Environmental Assessment. Washington, DC.

USEPA. 2008. Procedures for the derivation of equilibrium partitioning sediment benchmarks (ESBs) for the protection of benthic organisms: Compendium of Tier 2 values for nonionic organics. EPA-600-R-02-016. USEPA, Office of Research and Development. Washington, DC.

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. Society of Environmental Toxicology and Chemistry. Pensacola, FL.

Acknowledgements

The authors 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 compiled and standardized the data sets. Jeff Brown, Diana Young, and Darrin Greenstein assisted with data compilation and statistical analysis. The authors 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.