Comparison of sediment quality guideline performance for predicting sediment toxicity in southern California

Doris E. Vidal and Steven M. Bay

ABSTRACT - Several types of sediment quality guidelines (SQGs) are used by multiple agencies in southern California to interpret sediment chemistry data, yet little information is available to identify the best approaches to use. The objective of this study was to evaluate the performance of five SQGs to predict sediment toxicity in southern California: the effects range-median (ERM), consensus, mean sediment quality guideline quotient (SQGQ1), apparent effects threshold (AET), and equilibrium partitioning (EqP) for organics. Large differences in performance among the SQGs were obtained when each approach was applied to the same southern California dataset. SQG approaches that performed well in identifying nontoxic samples were not necessarily the best predictors of toxicity. In general, the ERMq, SQGQ1, and consensus approaches had a better overall performance than the AET and EqP approaches. The results indicate the need to predetermine management objectives and evaluate SQG performance using regional data so that the most appropriate SQG can be identified for specific applications.

INTRODUCTION

Contaminated sediments have been identified as an important cause of impairment to coastal water bodies, and diverse methods have been developed for their assessment and management. SQGs describe the level of contaminants in a sediment associated with various categories of adverse effects and are often used to interpret sediment chemistry data (U.S. EPA 2000). SQGs have been used by regulatory agencies, research institutions, and environmental organizations throughout the United States to identify contamination hotspots, characterize the suitability of dredge material for disposal, and establish goals for sediment cleanup and source control.

SQGs can be classified in two main categories based on the approach used to derive their values: empirical and mechanistic. Empirical SQG approaches are based on the statistical analysis of large databases of matched sediment chemistry and toxicity data to identify chemical concentrations associated with various levels of biological effects. Examples of this type of SQG include the effects range-low and effects range-median (ERL and ERM), which are concentrations corresponding to the 10th and 50th percentile of the distribution observed in toxic samples, respectively (Long *et al.* 1995). Variations in chemical speciation, bioavailability, or the mixture of chemicals present in the sample are not directly addressed in empirical SQGs; such effects are indirectly incorporated into the guidelines through the use of a database containing samples from diverse locations and sediment types. Empirical SQGs have two major practical advantages: they can be calculated for a large number of contaminants, and only routine chemical analysis data is needed for their application.

The second principal type of SQG includes values based on mechanistic models that incorporate factors that affect the bioavailability of chemicals in the sediment. Mechanistic SQGs may incorporate the effects of sediment organic carbon or sulfides (for metals) on the equilibrium partitioning of contaminants and also use laboratory dose response models to account for the effects of multiple contaminants (Pavlou 1987, Di Toro *et al.* 1991 and 1999). SQGs based on equilibrium partitioning (EqP) have been developed for selected pesticides and organics (U.S. EPA 2001, 2003a, b). By accounting for variations in bioavailability and mixture effects, mechanistic SQGs have a greater ability relative to empirical SQGs to determine the specific contaminants

responsible for toxicity. Mechanistic SQGs often require more extensive chemical data and published values are not available for many contaminants, relative to empirical SQGs.

The selection of an SQG type for use in a region such as southern California is hindered by a lack of information regarding the most suitable approach for sediment and contamination characteristics. Previous studies have shown that the performance (predictive ability) of particular SQGs is variable when applied to large-scale datasets (O'Conner *et al.* 1998, Long *et al*. 1998, MacDonald *et al.* 2000 b, Fairey *et al.* 2001). Variations in predictive ability among SQG approaches have also been observed when regional datasets have been analyzed (Avocet Consulting 2002, Crane *et al.* 2002). Regional differences in sediment characteristics (e.g., grain size or total organic carbon, TOC) or in the concentrations or types of contaminants may influence the predictive ability of SQGs (Field *et al.* 2002). The effects of unusual regional geochemical properties have been documented in other studies (Long *et al.* 2000, Hyland *et al.* 1999). The assessment of SQG predictive ability and possible adjustment for regional anthropogenic activities has been recommended to improve the use of SQGs for sediment assessment (Fairey *et al.* 2001). No systematic study of SQG performance for southern California sediments has been conducted. It cannot be determined whether the differences in SQG performance noted in previous studies are relevant for the sediment contamination characteristics present in this region.

The objective of this study was to evaluate the performance of selected empirical and mechanistic SQGs to predict sediment toxicity in southern California. SQG performance was evaluated on the capability of each guideline to identify both toxic and nontoxic samples in a dataset comprised of studies from southern California bays, harbors, and coastal areas.

Methods

A dataset consisting of paired chemistry and toxicity (10-d survival for marine amphipods) measurements for southern California sediment samples was used in this study. The data were obtained from a database created by the Los Angeles Contaminated Sediments Task Force (CSTF), which contained sediment quality data from 117 dredging, monitoring, and research studies conducted in the Southern California Bight between 1984 and 2001 (Myre *et al*. 2003). These studies included stations from offshore areas, bays, and harbors located from 35.4°N (San Luis Obispo County, CA, USA) to 31.75°N (U.S.-Mexico international border) (Figure 1). More information on these studies and the data used in the present analyses can be found at the SCCWRP website at www.SCCWRP.org/tools/CSTF.html.

A series of toxicity and chemistry screening criteria were used to select data from the CSTF database for analysis. Toxicity data were limited to those from solid-phase, 10-d amphipod survival tests using one of four marine amphipod species: *Ampelisca abdita, Rhepoxynius abronius*, *Eohaustorius estuarius,* and *Grandidierella japonica*. Each study was reviewed to verify that conventional toxicity data quality assurance criteria were met (ASTM 1997, Long *et al.* 1995, U.S. EPA 1994). The studies incorporated into the dataset had a mean control survival > 85% and < 0.4 mg/L unionized ammonia in the interstitial or overlying water. Samples with a mean survival < 80% of the control that were not statistically different from the control were excluded; this screening step eliminated toxicity data that were highly variable and thus unreliable. The screening steps removed a total of 3% of the amphipod toxicity samples originally available for the dataset.

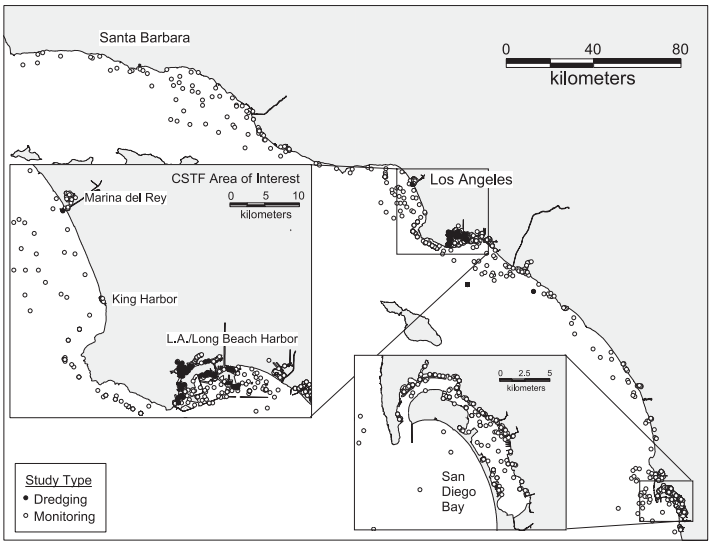

Figure 1. Location of southern California amphipod toxicity and matching chemistry samples used for the evaluation of SQG performance.

Table 1. Distribution of sediment chemistry and amphipod toxicity (control adjusted survival) data for southern California samples used in the analysis.

Chemical Name	Units	No. of Samples	10th Percentile	50th Percentile	90th Percentile
1-Methylnaphthalene	µg/kg	509	5.1	18.8	33.3
1-Methylphenanthrene	µg/kg	611	5.1	14.5	94.8
p,p'-DDD	µg/kg	959	0.5	2.5	25.1
p,p'-DDE	µg/kg	1,016	1.0	12.2	187.4
p,p'-DDT	µg/kg	782	0.25	1.9	21.4
2,6-Dimethylnaphthalene	µg/kg	473	5.2	19.4	42.0
2-Methylnaphthalene	µg/kg	658	5.6	19.5	58.9
Acenaphthene	µg/kg	624	3.5	21.0	70.7
Acenaphthylene	µg/kg	631	4.8	12.5	85.3
Anthracene	µg/kg	774	5.0	25.0	447.7
Antimony	mg/kg	823	0.17	0.95	5.00
Arsenic	mg/kg	931	2.8	7.6	18.0
Benz (a) anthracene	µg/kg	875	6.2	33.3	721.2
Benzo (e) pyrene	µg/kg	738	9.0	58.9	991.9
Benzo (g, h, I) perylene	µg/kg	773	6.6	50.0	528.2
Biphenyl	µg/kg	455	2.1	18.2	28.0
Cadmium	mg/kg	1,020	0.10	0.36	1.82
Chromium	mg/kg	1,020	21.0	56.0	111.0
Chrysene	µg/kg	890	7.9	47.7	1,000.0
Copper	mg/kg	988	8.0	50.0	213.7
Dibenz (a, h) anthracene	µg/kg	804	6.0	25.0	210.0
Dieldrin	µg/kg	331	0.20	1.08	4.77
Fluoranthene	µg/kg	895	10.0	66.9	1,368.0
Fluorene	µg/kg	680	4.2	23.0	110.0
Indeno (1,2,3-c, d) pyrene	µg/kg	753	6.5	50.0	579.6
Lead	mg/kg	1,022	6.4	27.4	99.3
Mercury	mg/kg	938	0.03	0.17	0.72
Naphthalene	µg/kg	709	5.3	15.8	50.8
Nickel	mg/kg	967	9.1	21.0	40.0
Total PCBs	µg/kg	577	7.0	77.2	608.3
Perylene	µg/kg	708	9.0	33.3	475.3
Phenanthrene	µg/kg	849	7.8	30.1	439.2
Pyrene	µg/kg	901	11.3	83.1	1,550.0
Silver	mg/kg	958	0.10	0.30	1.50
Zinc	mg/kg	963	40.0	130.0	339.8
TOC	%	944	0.31	1.13	2.92
Ampelisca abdita	%	166	82	97	104
Eohaustorius estuarius	%	329	60	91	101
Grandidierella japonica	%	41	60	90	100
Rhepoxynius abronius	%	562	40	81	97

The chemistry data were screened on the basis of completeness and detection limits. Screening for completeness required each study to contain data for a minimum suite of contaminants that included metals and PAHs. In addition, most studies contained data for pesticides, chlorinated hydrocarbons, and TOC (Table 1). The chemistry data were screened also to exclude data with high detection limits so that the presence of nondetectable values would not bias the calculations. A maximum acceptable detection limit was established for each chemical that was usually based on 20% of the lowest SQG for the empirical approaches examined. For chemicals exclusively included in the EqP approach, 20% of the Chronic EqP value was used. Nondetectable data having a detection limit above maximum criteria were eliminated on a chemical-specific basis. One-half of the detection limit was substituted for the nondetectable values that satisfied the screening criteria. The detection limit screening process resulted in the exclusion of 1.8% of the records (individual analytes measured for a sediment sample) from the available chemistry data.

Data Analyses

Four empirical SQG approaches and one mechanistic approach were compared in this study. The empirical approaches included effects range-median (ERM), which is the 50th percentile (median) of the distribution of chemical values in toxic data (Long *et al.* 1995); consensus medium effect concentration (MEC), which is the mean of diverse types of SQGs with a similar narrative intent (Swartz 1999, MacDonald *et al*. 2000a); mean sediment quality guideline quotient (SQGQ1), which is a subset of guidelines from various sources (Fairey *et al*. 2001); and the apparent effects threshold (AET) method, which represents the highest nontoxic concentration of a chemical in data from Washington (Tetra Tech 1986, Barrick *et al.* 1988). The mechanistic SQG approach was based on equilibrium partitioning (EqP) for narcotic organics (U.S. EPA 2001). Metals data were not included in the EqP analyses because the dataset did not contain information on acid volatile sulfides, which is required for application of the metals EqP approach.

The chemical-specific SQG values used in the analyses were obtained from published values for each approach (Table 2). Published consensus MECs were available for only polychlorinated biphenyls (PCBs) and polyaromatic hydrocarbons (PAHs), so new consensus values for 1,1,1-Trichloro-2,2-bis-(4'-chlorophenyl)ethane (DDTs), dieldrin, arsenic, cadmium, chromium, copper, lead, mercury, nickel, silver, and zinc were calculated for this study. The new MEC values were calculated as the geometric mean of the ERM, lowest AET, and probable effect concentration (PEL) for each chemical.

Chemical mixtures present in each sediment sample were addressed in three of the empirical approaches by normalizing each chemical to its respective SQG value and then summarizing the values as the mean quotient (e.g., mean ERMq, mean SQGQ1q, mean MECq). The data were normalized in a similar manner for the EqP approach by calculating the toxic units (TUs) represented by each chemical and then summarized by calculating the sum of the TUs. These normalization and summarization steps were recommended by developers of each SQG approach as a way to quantify both the number and magnitude of SQGs exceeded. The AET results were not normalized.

Two sets of mean ERM quotients were calculated and compared including and excluding total DDT values. The reliability of the DDT ERM is regarded as poor (Long *et al.* 1995), and this comparison was intended to determine the influence of the elevated DDT concentrations found in some areas of southern California on the overall performance of the ERM approach.

Thresholds specific to each guideline were taken from the literature to represent two levels of application, Levels I and II (Figure 2). Level I thresholds establish a concentration below which samples are expected to be nontoxic. Samples above a Level II threshold are expected to be toxic. The two application thresholds for the ERMq, SQGQ1, and consensus MECq consisted of different mean quotient values associated with either a low probability of toxicity (Level I) or a high probability of toxicity (Level II), as shown in Table 3. Comparison of the AET for Level I and II applications was accomplished by using two sets of SQG values; the lowest AET for any test (LAET) for Level I and the highest AET for any test (HAET) for Level II. For the EQP approach, the Level I and II applications were conducted using SQG values based on chronic or acute toxicity, respectively.

The performance of each SQG application threshold was assessed based on its ability to correctly classify a sediment sample as either a hit (chemistry $\geq$ threshold; amphipod toxicity expected) or a

Table 2. Chemical values used for Level I and II data analyses. Values for the effects range median (ERM) analysis were taken from Long et al. (2000). High AET (HAET) and low AET (LAET) values were taken from Barrick et al. (1988). Equilibrium partitioning for organics values was taken from the NSI inventory (U.S. EPA 2001). Concentrations were expressed on an organic carbon (OC) normalized basis for EqP acute and chronic guidelines, and on a dry weight (d wt) basis for all other guidelines.

Chemical Name	Units d wt	ERM	SQGQ1	LAET	HAET	Consensus	EqP Units OC	EqP Acute	EqP Chronic
1,1,1-Trichloroethane							µg/g	170	17
1,1,2,2-Tetrachloroethane							µg/g	830	160
1,2,4-Trichlorobenzene	µg/kg			30	64		µg/g	6,100	920
1,2-Dichlorobenzene	µg/kg			35	110		µg/g	610	34
1,3-Dichlorobenzene	µg/kg			170	170		µg/g	1,500	170
1,4-Dichlorobenzene	µg/kg			110	120		µg/g	420	35
2,4-Dimethylphenol	µg/kg			29	210				
2-methylnaphthalene	µg/kg	670		670	1,900				
4,4'-DDD	µg/kg			16	43				
4,4'-DDE	µg/kg			9	15				
4,4'-DDT	µg/kg			3.9	11				
4-Bromophenyl Phenyl Ether							µg/g	2,300	130
Acenaphthene	µg/kg	500		500	2,000		µg/g	2,043	491
Acenaphthylene	µg/kg	640		560	1,300		µg/g	1880	452
Aldrin	µg/kg			10					
Anthracene	µg/kg	1,100		960	13,000		µg/g	2,471	594
Antimony	mg/kg			150	200				
Arsenic	mg/kg	70		57	700	55			
Benzene							µg/g	100	5.7
Benzo(a)anthracene	µg/kg	1,600		1,300	5,100		µg/g	3,499	841
Benzo(a)pyrene	µg/kg	1,600		1,600	3,600		µg/g	4,014	965
Benzo(g,h,i)perylene	µg/kg			670	2,600				
Benzoic Acid	µg/kg			650	760				
Benzyl Alcohol	µg/kg			57	870				
Biphenyl							µg/kg	850	110
Bis(2-ethylhexyl) Phthalate				1,300	3,100				
Bromoform							µg/g	460	65
Butylbenzyl Phthalate	µg/kg			63	900		µg/g	15,000	1,100
Cadmium	mg/kg	9.6	4.2	5.1	9.6	5.9			
Chlordane	µg/kg	6	6	10					
Chlorobenzene							µg/g	1,500	82
Chromium	mg/kg	370		260	270	248.8			
Chrysene	µg/kg	2,800		1,400	9,200		µg/g	3,571	844
Copper	mg/kg	270		390	1,300	224.9			
DDT	µg/kg	46.1		6.9	69	25.4			
delta-BHC							µg/g	230	13
Diazinon (pesticide)							µg/g	0.73	0.19
Dibenz(a,h)anthracene	µg/kg	260		230	970				
Dibenzofuran	µg/kg			540	1,700		µg/g	3,700	200
Dieldrin	µg/kg	8	8	10		7	µg/g	55	13

Table 2 continued.

Chemical Name	Units D wt	ERM	SQGQ1	LAET	HAET	Consensus	EqP Units OC	EqP Acute	EqP Chronic
Diethyl phthalate	µg/kg			48	1,200		µg/g	1,100	63
Dimethyl phthalate	µg/kg			71	1,411				
Di-n-octyl Phthalate	µg/kg			420	6,200				
Di-n-butyl Phthalate	µg/kg			1,400	5,100		µg/g	8,000	1,100
Endrin							µg/g	17	5.5
Ethylbenzene	µg/kg			33	50		µg/g	8,500	480
Fluoranthene	µg/kg	5100		1,700	30,000		µg/g	2,941	707
Fluorene	µg/kg	540		540	3,600		µg/g	2,238	538
Heptachlor	µg/kg			10					
Hexachlorobenzene	µg/kg			22	230				
Hexachlorobutadiene	µg/kg			11	270				
Hexachloroethane	µg/kg			1,400	14,000		µg/g	1,800	100
HPAH	µg/kg			1,200	6,900				
Indeno(1,2,3-c,d)pyrene	µg/kg			600	2,600				
Lead	mg/kg	218	112	450	660	222.3			
Lindane	µg/kg			10			µg/g	8.8	0.37
LPAH	µg/kg			5,200	24,000				
Malathion (pesticide)							µg/g	0.62	0.067
Mercury	mg/kg	0.71		0.41	2.1	0.6			
Methoxychlor							µg/g	9.5	1.9
Naphthalene	mg/kg	2100		2,100	2,700		µg/g	1,602	385
Nickel	mg/kg	51.6		140	140	67.6			
N-Nitrosodiphenylamine	µg/kg			28	130				
PAHs			1800*			1800*	µg/g		
PCBs	mg/kg	180	400	130	3,100	0.47			
Pentachlorobenzene							µg/g	1,200	69
Pentachlorophenol	µg/kg			140	690				
Phenanthrene	µg/kg	1500		1,500	6,900		µg/g	2,479	596
Phenol	µg/kg			420	1200				
Pyrene	µg/kg	2600		2,600	16,000		µg/g	2,900	679
Silver	mg/kg	3.7	1.8	6.1	8.4	3.4			
Tetrachloroethylene	µg/kg			160	1,600		µg/g	420	53
Tetrachloromethane							µg/g	2,100	120
Toluene							µg/g	1,600	89
Toxaphene							µg/g	490	10
Xylene	µg/kg			40	160				
Zinc	mg/kg	410	410	410	1,600	357.1			
Endosulfan I							µg/g	0.74	0.29
Endosulfan II							µg/g	3.5	1.4
Trichloroethene							µg/g	2,000	210
m-Xylene							µg/g	45	2.5

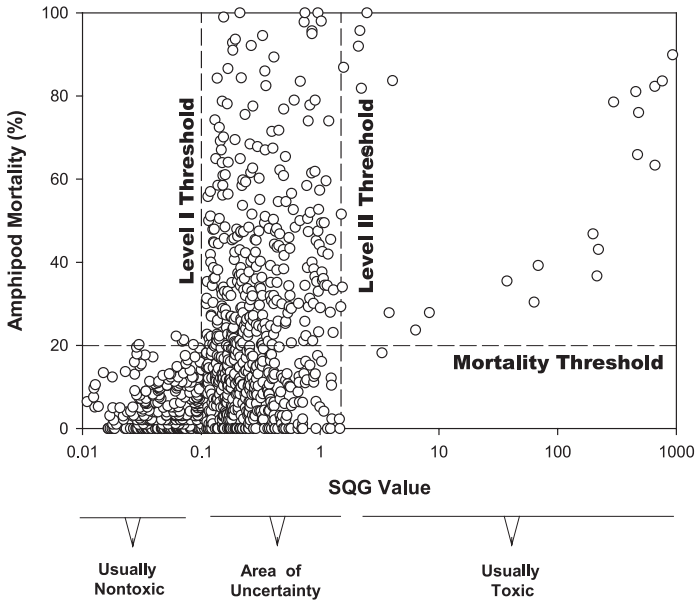

Figure 2. Example of sediment toxicity data distribution showing high uncertainty within the middle range of overall contamination between the Level I and II SQG application thresholds.

no hit (chemistry < threshold; amphipod toxicity not expected). A sample was considered toxic when its amphipod survival was significantly different from a control or reference, and the survival was < 80% of the control (Thursby *et al.* 1997). The outcome predicted by the guideline and the measured amphipod toxicity results then were used to calculate four performance measures (Figure 3):

Nontoxicity Efficiency: Percentage of the samples predicted to be nontoxic (no hits), which were in fact nontoxic to amphipods.

% Nontoxicity Efficiency = (True Negative /(False Negative + True Negative)) x 100

Where:

False Negative = Number of toxic samples incorrectly predicted as nontoxic (no hits)

True Negative = Number of nontoxic samples correctly predicted as nontoxic

Nontoxicity Specificity: Percentage of all nontoxic samples in the dataset that were correctly predicted to be nontoxic by the SQG.

% Nontoxicity Specificity = (True Negative /(True Negative + False Positive)) x 100

Where:

False Positive = Number of nontoxic samples incorrectly predicted as toxic (hits).

Toxicity Efficiency: Percentage of the samples predicted to be toxic (hits), which were in fact toxic to amphipods.

% Toxicity Efficiency = (True Positive /(True Positive + False Positive)) x 100

Where:

True Positive = Number of toxic samples correctly predicted as toxic (hits).

Toxicity Sensitivity: Percentage of all toxic samples in the dataset that were correctly classified as toxic by the SQG.

% Toxicity Sensitivity = (True Positive /(True Positive + False Negative)) x 100

RESULTS

The dataset used to evaluate the empirical approaches contained 1,098 samples from 703 stations. The stations represented the continental shelf off of Southern California as well as all of the developed bays and harbors (Figure 1). A total of 933 records from 52 studies were used for analyses involving the EqP approach, because not all the studies had TOC data available. The dataset included 82 chemicals that were used to evaluate the various SQG approaches (Table 2). The number of chemicals reported for each study varied because of differences in study design; 50% of the samples contained data for at least 33 chemicals (Table 1). Thirty-one percent of the samples in the dataset were toxic to amphipods. The majority of the toxicity data was represented by two amphipod species: *Rhepoxynius abronius* (51%) and *Eohaustorius estuarius* (30%).

Similar toxicity patterns were observed when the ERM, consensus, and SQGQ1 mean quotients were calculated from the data (Figure 4A-D). A large amount of data (64%) were distributed between the Level I and II thresholds for each SQG and showed no apparent relationship between amphipod mortality and the mean quotient values. The data distribution showed a pattern of higher amphipod mortality at the

Table 3. Thresholds used for evaluating SQG performance at Level I and Level II applications.

Guideline Metric	Level I	Level II	Source
ERM Mean Quotient	0.1	1.5	Long *et al*. (2000)
AET	> any LAET	> any HAET	Barrick, R. *et al*. (1988)
EqP Organics Sum TU	Sum Chronic TU= 1	Sum Acute TU= 1	U.S.EPA (2001)
Consensus MEC Mean Quotient	0.1	1.5	MacDonald *et al.* (2000b) Swartz (1999); This study
SQGQ1 Mean Quotient	0.1	2.0	Fairey *et al*. (2001)

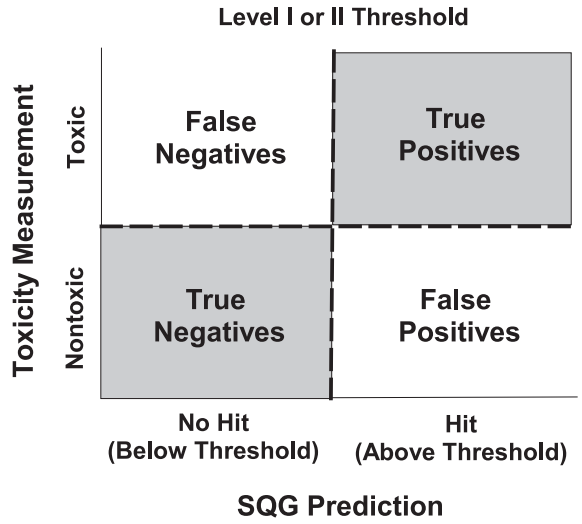

Figure 3. Classification of sediment quality predictions based on application of a SQG threshold and toxicity test results. Shaded regions represent cases of agreement (i.e., prediction agrees with toxicity test result).

highest mean quotients and samples that were located below the Level I threshold showed a clear pattern of low amphipod mortality.

The data distributions for the AET and EqP were different from the other approaches. No particular trends in amphipod mortality relative to the guideline value were observed (Figure 4E-H). The low and high AETs showed no relationship between the number of guideline exceedances and amphipod mortality. Many samples with high mortality did not exceed any of the guideline values. Similarly, for EqP for organics, there was no relationship between amphipod mortality and the number of chronic or acute toxic units.

Nontoxicity Prediction Performance

The SQGs were highly variable in their ability to correctly predict nontoxic samples (nontoxicity efficiency) using the Level I thresholds. Nontoxicity efficiency ranged from 64% to 95%, with the Consensus MEC and SQGQ1 approaches showing the highest efficiency (Table 4). The lowest nontoxicity efficiency was obtained using the EqP Level I threshold (chronic toxicity). The entire dataset contained 69% nontoxic samples, so Level I SQG approaches with less than 69% nontoxicity efficiency did not provide an increased ability to predict the likelihood of a sample being nontoxic. Most of the SQGs were able to identify only a small percentage of the truly nontoxic samples as a no hit using the Level I threshold (nontoxicity specificity of 24% to 70%).

The Level I performances of the mean ERMq with and without DDT were comparable between the two guidelines (Table 4). Nontoxicity efficiency varied by less than three percentage points and nontoxicity specificity varied by less than eight percentage points.

Toxicity Prediction Performance

Similar to the nontoxicity prediction results, the ability of the SQGs to predict the occurrence of toxicity was highly variable (Table 4). Toxicity efficiency for the Level II thresholds ranged from 36% (AET) to 85% (ERM without DDT). Thirty-one percent of all of the samples in the database were toxic

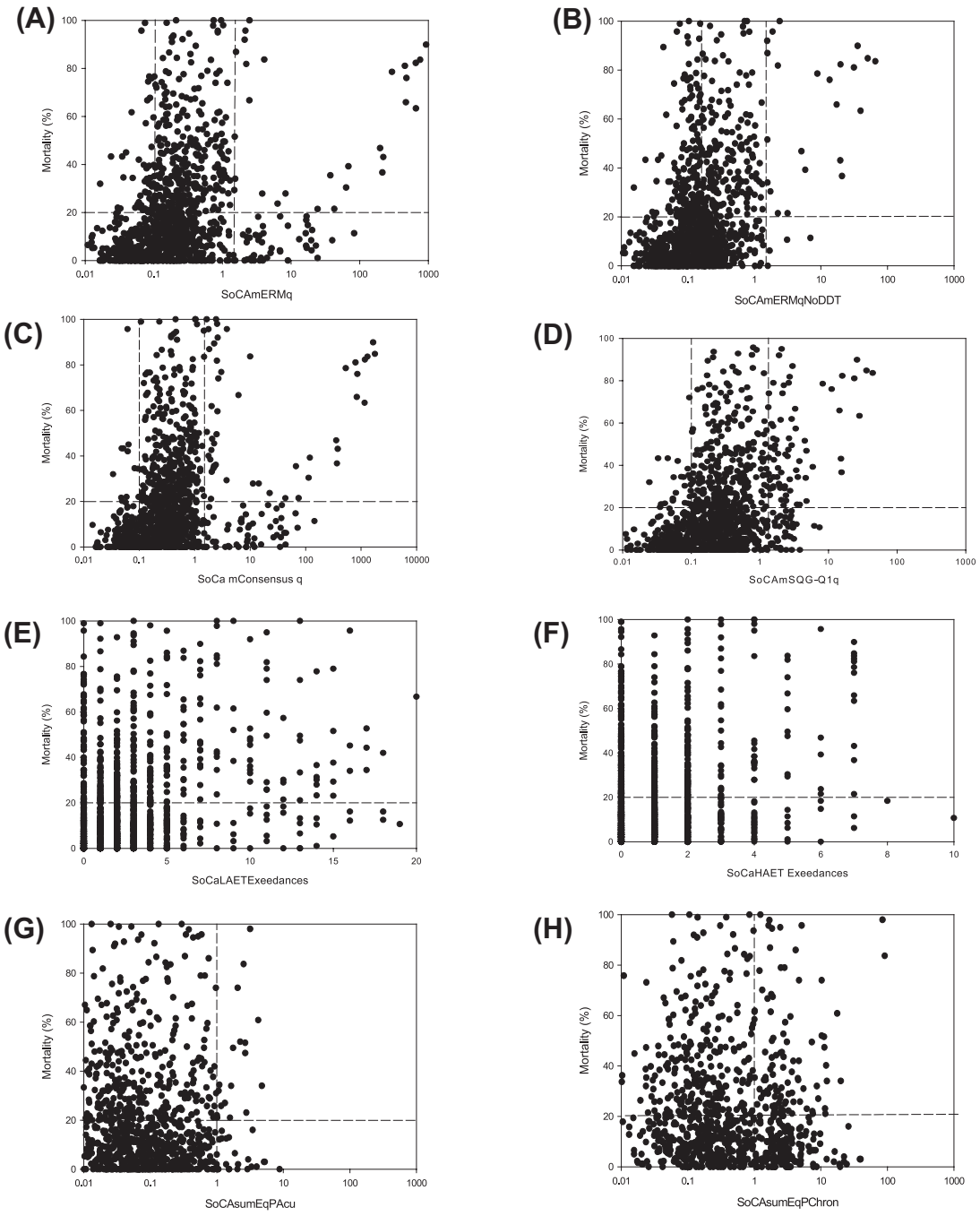

Figure 4. Toxicity data distributions for the SQG approaches evaluated in this study. Amphipod mortality is plotted relative to the mean SQG quotient, number of AET exceedances, or sum of EqP toxic units. A: mean ERMq with DDT; B: mean ERMq excluding DDT; C: mean consensus MECq; D: mean SQGQ1; E: number of LAET exceedances (Level I); F: number of HAET exceedances (Level II); G: sum of EqP chronic toxic units (Level I); H: sum of EqP acute toxic units (Level II). Vertical dashed lines indicate Level I or II thresholds and the horizontal dashed line indicates the boundary between toxic and nontoxic samples.

Table 4. Performance measurements for the SQG approaches.

Performance Measurement	ERMq No DDT	ERMq	AET	EqP	SQGQ1	Consensus
Level I						
Nontoxicity Efficiency	83.6%	85.9%	76.1%	63.9%	93.0%	95.2%
Nontoxicity Specificity	47.2%	39.7%	26.0%	69.4%	34.1%	24.4%
Level II						
Toxicity Efficiency	85.2%	38.4%	36.3%	45.0%	64.3%	46.2%
Toxicity Sensitivity	6.3%	7.7%	61.2%	5.4%	12.4%	13.2%

to amphipods, so SQG efficiency values < 40% represented very little ability to predict the likelihood of toxicity. Toxicity sensitivity, the percentage of all toxic samples above the Level II threshold, was very low (< 14%) for every SQG except the AET (61%).

Inclusion of DDT in the mean ERMq calculations resulted in a large decrease in the ability to correctly predict toxicity using the Level II threshold. Toxicity efficiency declined from 85% (without DDT) to 38% when DDT was included in the calculations (Table 4). Toxicity sensitivity was low for the mean ERMq regardless of whether DDT was included (6% to 7%).

Discussion

The results showed substantial differences in the abilities of various SQGs to discriminate between toxic and nontoxic samples when applied to southern California data. None of the analyzed guidelines had the highest performance in all categories, but the mean ERMq (excluding DDT) showed relatively high efficiency to accurately identify both toxic and nontoxic samples (84% to 85%). Several other SQG approaches also showed high efficiency for predicting nontoxic samples, indicating that multiple approaches are useful for identifying samples with low potential for adverse effects.

Each of the guidelines showed a tradeoff in performance relative to toxicity or nontoxicity efficiency and the corresponding measure of sensitivity or specificity. For example, the SQGQ1 and consensus guidelines had the highest accuracy to predict nontoxic samples (high nontoxicity efficiency) but these two guidelines also had the lowest ability to identify all of the nontoxic samples (nontoxicity specificity). The significance of this dichotomy is that a large number of sediment samples in the database with mid-range chemical concentrations and variable toxicity ranges (uncertainty middle range or "gray zone") were improperly classified. This tradeoff in performance appears to be unavoidable with the currently available approaches and indicates that the objective of an SQG application (e.g., high efficiency or high sensitivity) must be determined before the optimal SQG can be selected.

The AET and EqP approaches had the lowest overall performance when applied to the southern California dataset. The low performance of the AETs may be related to their regional basis. AETs were developed using data from the Puget Sound region of Washington and were intended for use only as regional guidelines. Differences in the pattern and magnitude of contamination between Puget Sound and southern California may have contributed to the moderate and low performances obtained in this study. The EqP approach also had a low ability to identify toxic samples. Since the EqP guidelines for organics incorporate adjustments for variations in bioavailability and toxic potency, it was expected that this approach would have relatively high toxicity efficiency. The low toxicity efficiency of the Level II EqP guidelines suggests that PAHs and other organics included in the guidelines may not be the principal causes of toxicity to amphipods in southern California. Alternatively, the EqP model may not adequately describe the bioavailability of these compounds in southern California sediments.

The performance of some SQGs when applied to the southern California dataset differed from that reported for other datasets. Application of the mean ERMq approach (>1.5) to datasets from California bays, Biscayne Bay (Florida), or nationwide sites by Fairey *et al*. (2001) produced Level II toxicity efficiencies that were variable (46% to 67%) and greater

than that reported here (38%). Application of the SQGQ1 approach (>1.5) by Fairey *et al*. to the same three datasets produced toxicity efficiencies of 76%, 100%, and 88%, values that also were greater than the toxicity efficiency obtained in the present study (64%). Barrick *et al.* (1988) analyzed data from various areas of Puget Sound (Washington) using AET values and obtained a wide range of toxicity efficiencies (33% to 100%) and toxicity sensitivities (20% to 100%). These studies, together with the present investigation, demonstrate that regional differences in contamination or toxicity response can have a large influence on the SQG performance.

The observation that various empirical SQGs using different chemical values have a similar ability to discriminate among toxic and nontoxic samples suggests that exceedances of individual empirical chemical guidelines are unreliable indicators of toxicity and do not necessarily indicate the cause of toxicity. For example, the mean SQGQ1 and mean ERMq had similar nontoxicity efficiency and specificity values, yet the SQGQ1 uses only 9 chemicals in comparison to the 24 used for the ERMq. The presence of many contaminants in a sediment samples and the high degree of correlation among them indicates that most empirical SQG values should not be used in isolation, but rather should be used in combination to provide an overall indication of the potential for adverse effects (e.g., likely to be toxic or nontoxic). Other studies have also suggested caution in the use of individual chemical SQG values when assessing sediment quality (Fairey *et al.* 2001, Long *et al.* 2000).

The improvement in mean ERMq performance that resulted from excluding the DDT data provides an example of the potential for erroneous conclusions based on the use of individual chemical SQGs. The negative influence of the DDT ERM on guideline performance demonstrates the potential influence of regional differences in contamination. Sediments from the Palos Verdes shelf contain some of the world's highest concentrations of DDT. Relatively little data from Palos Verdes was used in deriving the original ERM values, and the calculated ERM value may have been unduly influenced by the inclusion of data where DDT was not the principal cause of toxicity. Fairey *et al*. (2001) applied the mean ERMq to data from California and also obtained a Level II toxicity efficiency value (46%) that was much lower than the efficiency obtained using data from a national dataset (64%). These results confirm the caution provided by Long *et al*. (1995) regarding the low accuracy of the DDT ERM. The effects on performance caused by the exclusion of the DDT data were not investigated for the other approaches, but a similar change in toxicity efficiency would probably have occurred.

Chemical-specific SGQ values based on equilibrium partitioning theory are more appropriate for use in determining the cause of sediment toxicity, as these guidelines attempt to address changes in contaminant bioavailability due to sediment characteristics and the effect of chemical mixtures. Application of the EqP approach for organics was not successful in this study, however, indicating that model revisions are needed before this approach can be applied with confidence to southern California sediments. The performance of the EqP approach would likely be improved by the inclusion of values for additional organics; data for PCBs and pesticides were not used in these analyses because acute and chronic SQG values were not available. The inclusion of metals data also may have improved the performance of the EqP approach. Application of the metals EqP approach requires data on the acid volatile sulfides (AVS) concentration of the sediment, which was not available in the dataset.

Patterns of sediment contamination in southern California probably reduced the performance of many of the SQGs. The southern California dataset contained a relatively small number of samples that were either highly contaminated or contained very low levels of contamination. Consequently, most of the samples contained moderate contamination levels that cannot be reliably classified using empirical SQGs. Similar results were reported in a study conducted in the Saint Louis River (Crane *et al.* 2002). Crane *et al.* found that their dataset contained few highly contaminated samples, which resulted in low toxicity efficiency and sensitivity.

While the results presented in this report indicate the relative performance of the analyzed SQG approaches for southern California data, they are not definitive. The Level I and II thresholds evaluated reflect recommendations based on other datasets and may not be the optimal ones for use in southern California. An advantage of the mean quotient approach using empirical SQGs is that the application threshold can be adjusted to provide better performance under local conditions. For example, a change Level II threshold for the mean ERMq (no DDT) from 2.0 to 1.2 results in a three-fold increase in toxicity sensitivity without a substantial change in toxicity efficiency for the southern California data.

Mechanistic models such as the EqP approach for organics are based upon generalized geochemical and toxicological relationships that are more difficult to adjust to local conditions. The four performance measures described in this report, along with a regional database, should be used to evaluate the performance of SQG thresholds and guide the development of alternative values before they are used to assess sediment quality.

Literature Cited

American Society for Testing Materials (ASTM). 1997. Standard guide for conducting solid phase, 10 day, static sediment test with marine and estuarine infaunal amphipod. pp. 732-757 *in:* Annual Book of Standards. Vol. 11.05. E 1367-92. Philadelphia, PA.

Avocet Consulting. 2002. Development of Freshwater Sediment Quality Values for use in Washington State. Phase 1, Task 6 Final Report. Publication Number: 02-09-050. Washington Department of Ecology Toxics Cleanup Program. Olympia, WA.

Barrick, R., S. Becker, R. Pastorok, L. Brown and H. Beller. 1988. Sediment quality values refinement: 1988 update evaluation of Puget Sound AET, Vol. 1. Prepared for the Puget Sound Estuary Program, Office of Puget Sound. Bellevue, WA.

Crane, J.L., D.D MacDonald, C.G. Ingersoll, D.E. Smorong, R.A. Lindskoog, C.G. Severn, Berger and L.J. Field. 2002. Evaluation of numerical sediment quality targets for the St. Louis River area of concern. *Archives of Environmental Toxicology and Chemistry* 43: 1-10.

Di Toro, D.M., C.S. Zarba, D.J. Hansen, W.J. Berry, R.C. Swartz, C.E. Cowan., S.P. Pavlou, H.E. Allen, N.A. Thomas and P.R. Paquin. 1991. Technical basis for establishing sediment quality criteria for non-ionic organic chemicals using equilibrium partitioning. *Environmental Toxicology and Chemistry* 10: 1541-1583.

Di Toro, D.M., J.A. McGrath and D.J. Hansen. 1999. Technical basis for narcotic chemicals and polycyclic aromatic hydrocarbon criteria. I. Water and tissue. *Environmental Toxicology and Chemistry* 19: 1951-1970.

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.D. Mac Donald, S.B. Norton and C.G. Ingersoll. 2002. Predicting amphipod toxicity from sediment chemistry using logistic regression models. *Environmental Toxicology and Chemistry* 21: 1993-2005.

Hyland, J.L., R.F. Van Dolah and T.R. Snoots. 1999. Predicting stress in benthic communities of southeastern U.S. estuaries in relation to chemical contamination of sediments. *Environmental Toxicology and Chemistry* 18: 2557-2564.

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.

MacDonald, D.D., L.M. Di Pinto, J. Field, C.G. Ingersoll, E.R. Long, and R.C. Swartz. 2000a. 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 and T.A. Berger. 2000b. Development and evaluation of consensus-based sediment quality guidelines for fresh water ecosystems. *Archives of Environmental Toxicology and Chemistry* 39: 20-31.

Myre, P.L., S.M. Bay and D.E. Vidal. 2003. Sediment Quality Database User Guide. Prepared for the Los Angeles Basin Contaminated Sediments Task Force. Los Angeles, CA.

O'Conner, 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.

Pavlou, S.P. 1987. The use of equilibrium partitioning approach in determining safe levels of contaminants in marine sediments *in:* K.L Dickson, A.W. Maki and W.A. Brungs (eds), Fate and Effects of Sediment-bound Chemicals in Aquatic Systems. Proceedings of the Sixth Pellston Workshop, Florissant, Colorado, August 12-17, 1984. Pergamon Press. New York, NY.

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

Tetra Tech Inc. 1986. Evaluation of statistical relationships among chemicals and biological variables using pattern recognition techniques. Prepared for the Puget Sound Dredged Disposal Analysis and Puget Sound Estuary Program. Bellevue, WA.

Thursby, G.B., J. Heltshe and K.J. Scott. 1997. Revised approach to toxicity test acceptability criteria using a statistical performance assessment. *Environmental Toxicology and Chemistry* 16: 1322-1329.

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

U.S. EPA (United States Environmental Protection Agency). 2000. Development of a framework for evaluating numerical sediment quality targets and sediment contamination in the Saint Luis River Area of concern. EPA 905-R-00-008. Great Lakes National Program Office, Region V. Chicago, IL.

U.S. EPA (United States Environmental Protection Agency). 2001. The incidence and severity of sediment contamination in surface waters of the United States. National Sediment Quality Survey: Second Edition. EPA 823-R-01-01. Office of Science and Technology Standards and Health Protection Division. Washington, DC.

U.S. EPA (United States Environmental Protection Agency). 2003a. Procedures for the derivation of equilibrium partitioning sediment benchmarks (ESBs) for the protection of benthic organisms: Dieldrin. EPA 600-R-02-010. Office of Research and Development. Washington, DC.

U.S. EPA (United States Environmental Protection Agency). 2003b. Procedures for the derivation of equilibrium partitioning sediment benchmarks (ESBs) for the protection of benthic organisms: Endrin. EPA 600-R-02-009. U.S. Environmental Protection Agency. Office of Research and Development. Washington, DC.

Acknowledgements

The authors would like to acknowledge the invaluable assistance of D. Greenstein, J. Brown, R. Hagstrom, L. Cooper, S. Moore, and B. Pauley for their help with the analyses and preparation of the report. The authors also thank P. Myre and T. Gries for their assistance in conducting the analyses and compiling information. This project could not have been completed without the assistance of individuals, representing both public agencies and private companies, working in collaboration through the Los Angeles Basin Contaminated Sediments Task Force. The authors especially thank P. Johansen, R. Cameron, and K. Anderson for their assistance.