SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 17, No. 4, pp. 714–727, 1998
Printed in the USA
0730-7268/98 $6.00 + .00

PREDICTING TOXICITY IN MARINE SEDIMENTS WITH NUMERICAL SEDIMENT QUALITY GUIDELINES

EDWARD R. LONG,*† L. JAY FIELD,‡ and DONALD D. MACDONALD§
†Coastal Monitoring and Bioeffects Assessment Division, ‡Hazardous Materials Response and Assessment Division, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, Washington 98115, USA
§MacDonald Environmental Sciences, Ltd., 1733 Idaho Place, Nanaimo, British Columbia V9X 1C6, Canada

(*Received* 5 *February* 1997; *Accepted* 29 *July* 1997)

Abstract—Matching synoptically collected chemical and laboratory bioassay data (n = 1,068) were compiled from analyses of surficial sediment samples collected during 1990 to 1993 to evaluate the predictive ability of sediment quality guidelines (SQGs), specifically, effects range—low (ERL), effects range—median (ERM), threshold effects level (TEL), and probable effects level (PEL) values. Data were acquired from surveys of sediment quality performed in estuaries along the Atlantic, Pacific, and Gulf of Mexico coasts. Samples were classified as either nontoxic ($p > 0.05$ relative to controls), marginally toxic ($p < 0.05$ only), or highly toxic ($p < 0.05$ and response greater than minimum significant difference relative to controls). This analysis indicated that, when not exceeded, the ERLs and TELs were highly predictive of nontoxicity. The percentages of samples that were highly toxic generally increased with increasing numbers of guidelines (particularly the ERMs and PELs) that were exceeded. Also, the incidence of toxicity increased with increases in concentrations of mixtures of chemicals normalized to (divided by) the SQGs. The ERMs and PELs indicated high predictive ability in samples in which many substances exceeded these concentrations. Suggestions are provided on the uses of these estimates of the predictive ability of sediment guidelines.

Keywords—Sediment quality guidelines Predictive ability Laboratory toxicity tests

INTRODUCTION

Using similar empirical approaches, sediment quality guidelines (SQGs) were prepared for salt water [1–3] and freshwater [4,5] as informal (nonregulatory) benchmarks to aid in the interpretation of sediment chemistry data. For marine sediments, effects range—low (ERL) and effects range—median (ERM) concentrations for 9 trace metals, 3 chlorinated organics, and 13 polynuclear aromatic hydrocarbons (PAHs) were identified [1]. Threshold effects level (TEL) and probable effects level (PEL) concentrations for 9 trace metals, 8 chlorinated organics, 1 phthalate, and 13 PAHs were published [2]. These guidelines were not based upon experiments in which causality was determined. Rather, both sets of marine guidelines were based upon empirical analyses of data compiled from numerous field and laboratory studies performed in many estuaries and bays of North America. These studies included chemistry data and a variety of different types of biological data for numerous taxa derived from either bioassays of field-collected samples, laboratory toxicity tests of clean sediments spiked with specific toxicants, benthic community analyses, or equilibrium-partitioning models.

The objectives of the ERL and TEL values and of the ERM and PEL values were comparable. The ERLs and TELs were intended to represent chemical concentrations toward the low end of the effects ranges, that is, below which adverse biological effects were rarely observed. The ERMs and PELs were intended to represent concentrations toward the middle of the effects ranges and above which effects were more frequently observed. As estimates of reliability, the incidence of adverse effects within concentration ranges defined by these SQGs were determined using data with which they were derived [1,2]. Generally, adverse effects occurred in less than 10% of studies in which concentrations were below the respective ERL or TEL values and were observed in more than 75% or 50% of studies in which concentrations exceeded the ERMs or PELs, respectively.

Since they were published, the guidelines [1,2] have been used as interpretive tools in many sediment assessments throughout North America and elsewhere. Generally, the ERLs and TELs have been used to identify relatively uncontaminated samples that pose a limited risk of toxicity. The ERMs and PELs have been used to identify those samples and areas in which chemical concentrations were sufficiently elevated to warrant further evaluation. Because these guidelines were based upon analyses of large databases, mostly composed of field-collected data in which mixtures of toxicants were encountered, it was assumed [1,2] that the guidelines would provide relatively accurate tools for classifying newly collected samples as potentially toxic or nontoxic. Thus far, however, the accuracy of the two sets of guidelines in predicting nontoxic and toxic conditions correctly has not been evaluated. Therefore, because of the widespread use of these guidelines, we concluded there was a need for analyses of their predictive ability with data independent of those with which the SQGs were derived.

The objectives of this paper are to quantify the frequency with which ERL/ERM and TEL/PEL guidelines correctly classify samples as either nontoxic or toxic; to quantify the incidence of toxicity among samples in which different numbers of SQGs were exceeded; to determine the incidence of toxicity

* To whom correspondence may be addressed

Table 1. Sources of data and the toxicity tests performed in each study

Survey area	Year sampled	No. samples	Bioassays performed: Amphipod survival	Clam embryo survival	Clam embryo development	Microbial bioluminescence	Urchin egg fertilization	Urchin embryo development	Abalone embryo development
Hudson–Raritan estuary	1991	38	X	X	X	X			
Newark Bay	1993	20	X						
Long Island Sound	1991	63	X	X	X	X			
Boston Harbor	1993	30	X			X	X		
Tampa Bay phase 1	1992	16	X			X	X		
Tampa Bay phase 2	1993	45	X				X		
San Diego Bay	1993	121	X					X	
San Pedro Bay	1992	45	X						X
Charleston Harbor	1993	79	X			X	X		
EMAP—Estuaries[a]	1990–1992	611	X						
Total		1,068							

[a] EMAP = Environmental Monitoring and Assessment Program; data from mysid tests not included.

to the SQGs; and to compare the relative predictive ability of the two sets of guidelines. The design followed that of a previous study in freshwater [5] in which type I and type II errors were determined for ERL/ERM and TEL/PEL values. Type I errors (false positives) are those in which toxicity was expected (based upon high chemical concentrations), but was not observed. Type II errors (false negatives) are those in which no toxicity was expected (low chemical concentrations), but was actually observed.

METHODS

Matching, synoptically collected, sediment chemistry and bioassay data for 1,068 samples were compiled from studies performed by the National Oceanic and Atmospheric Administration (NOAA) and U.S. Environmental Protection Agency (U.S. EPA) (Table 1). Regional sediment quality assessments were conducted as a part of NOAA's National Status and Trends Program (NS&TP) and included those performed in (all in the USA) the Hudson–Raritan estuary in New York and New Jersey [6], Newark Bay in New Jersey [6], the bays adjoining Long Island Sound in New York and Connecticut [7], Boston Harbor in Massachusetts [8], Tampa Bay in Florida [9], San Diego Bay [10] and San Pedro Bay [11] in southern California, and Charleston Harbor in South Carolina (unpublished). The U.S. EPA data were generated in Environmental Monitoring and Assessment Program (EMAP) studies of the Virginian and Louisianian estuarine provinces [12–14].

All of these data were generated during surveys performed to quantify the spatial extent, patterns, and severity of adverse biological effects attributable to toxic substances. Samples from the upper 2 to 3 cm of the sediments were collected with grab samplers throughout each survey area to characterize surficial sediment contamination and toxicity. Data for these analyses were selected because they were generated with similar protocols, included matching chemistry and toxicity results, indicated a range in toxicity responses, and represented conditions from all three coastlines.

Sample collection and handling methods, toxicity testing methods, chemical analytical protocols, and raw data are included in the respective technical reports. All analytical laboratories followed the performance-based protocols of the NS&TP and EMAP—Estuaries to ensure comparability among

rials [16,17] for the amphipod survival tests, U.S. National Biological Service [18] for the urchin tests, and U.S. EPA [19] and Schiewe et al. [20] for the Microtox® tests (AZUR Environmental, Carlsbad, CA, USA). All bioassay data were expressed as percent of negative, laboratory controls (not reference samples) to account for variability among studies and laboratories in organism viability.

We considered several different approaches to the classification of samples as either nontoxic or toxic. In an interlaboratory comparison of performance, results of amphipod survival tests were classified as either nontoxic (mean survival 96–96.5%), marginally toxic (mean survival 76.5–83%), clearly toxic (mean survival < 76%), highly toxic (mean survival < 20%) [21]. Swartz et al. [22] classified results of amphipod survival tests as either not toxic (<13% mortality), uncertain (13–24% mortality), or toxic (>24% mortality). Statistical tests were recommended [16] to determine if differences in results of tests of field-collected samples and controls are statistically significant. An alternative approach [23], based upon results of power analyses of amphipod survival tests, recommended the use of minimum significant differences (MSDs) from controls as criteria for classifying samples as toxic.

We chose to use a combination of these approaches to classify samples. Following standardized procedures [16], samples in which test results were not statistically different from negative controls (i.e., $p > 0.05$) were classified as nontoxic and samples in which results were significantly different from controls were classified as toxic. However, to further distinguish differences in degrees of toxicity, sample classifications followed the recommendations of Thursby et al [23]. Samples in which test results were significant relative to controls, but were less than MSDs were labeled as marginally toxic and those in which results were both significant and greater than MSDs were labeled as highly toxic. The highly toxic label does not imply that toxicity was severe; rather, it was used to identify those results for which statistical certainty was greatest. The MSD values calculated and published for *Ampelisca abdita* [23] were used for all amphipod test results. The MSD values for Microtox tests [24], *Arbacia punctulata* fertilization tests [25], and and all other tests were determined empirically with power analyses of the frequency distributions of data from each test.

laboratory replicates is very small [23]. However, these samples could not be classified as nontoxic because they were significantly different from controls. Therefore, we chose to classify them separately as neither nontoxic nor highly toxic. Because of the uncertainty associated with marginally toxic results, this evaluation focuses mainly upon the nontoxic and highly toxic categories.

Following the completion of an electronic database, several analyses were performed to determine the predictive ability of the guidelines. In these analyses, the guidelines for nickel were excluded because of the low degree of reliability determined for these values [1,2]. Also, the sums of low- and high-molecular-weight PAHs and total PAHs were excluded to avoid redundancy with the data for individual compounds. In summations of total polychlorinated biphenyls (PCBs), total dichlorodiphenyltrichloroethanes (DDTs), and total PAHs, concentrations of individual compounds were treated as zeroes when they were below method detection limits (MDLs). The MDLs achieved differed slightly among laboratories; therefore, the use of zeroes minimized inconsistencies in data treatments. In any case the use of either one half of the MDL or zeroes had no effect upon classification of samples relative to the SQGs.

Three data analyses were performed. First, the predictive abilities of individual SQGs were determined. Second, the incidence of toxicity was determined among samples in which none of the substances equaled or exceeded the ERL concentrations; in which one or increasing numbers of substances exceeded ERL concentrations, but none exceeded any ERM; and in which one or increasing numbers of substances exceeded ERM concentrations. The same approach was used to evaluate the predictive ability of the TEL/PELs. We scored samples as exceeding SQGs when a chemical concentration either equaled the value or exceeded it by any amount.

In the third analysis, the incidence of toxicity over ranges in mean SQG quotients [5,25] was determined. The concentrations of individual chemicals were divided by their respective ERMs or PELs and the means of these concentration-to-SQG quotients were determined. The means of these quotients were determined to account for differences among studies in the numbers of chemicals for which analyses were performed. Predictive ability was calculated with samples classified as either nontoxic or highly toxic, excluding the marginally toxic results.

Similar to the criteria used to determine guideline reliability [1], we considered the guidelines to be predictive if the incidence of toxicity was less than 25% when all concentrations were less than the ERLs or TELs and greater than 75% when at least one concentration exceeded an ERM or PEL. Therefore, our target level for both false negatives and false positives was ≤25%.

Data are reported for the results of amphipod survival tests alone and for any one of the battery of two to four tests performed. In the latter analyses, samples were classified as marginally or highly toxic if one or more of the bioassays met the criteria for these classifications.

RESULTS

The database

Data were compiled from 1,068 samples analyzed during EMAP and NOAA studies conducted during 1990 to 1993. Roughly one third of the data were obtained from the NOAA from 20 to 121 (Table 1). The EMAP data from the Atlantic and Gulf coasts constituted the remaining two thirds of the database (n = 611).

Amphipod survival was determined for all samples; one to three additional tests were performed on all samples except those collected in the EMAP and Newark Bay studies (n = 437). The data from bioassays performed with mysids by the EMAP were not used because these tests failed to indicate toxicity. Amphipod survival was determined with *A. abdita* in Atlantic and Gulf coast surveys and with *Rhepoxynius abronius* in California surveys. Other tests included bivalve (*Mulinia lateralis*) embryo survival and development with exposures to elutriates; microbial bioluminescence (Microtox) in exposures to organic solvent extracts; and pore-water tests of echinoderm (*A. punctulata*) fertilization in Gulf and Atlantic coast areas, echinoderm (purple urchin, *Strongylocentrotus purpuratus*) embryo development in San Diego Bay, and embryological development of red abalone (*Haliotis rufescens*) embryos in San Pedro Bay. Insufficient numbers of samples were tested in any of these nonamphipod tests to warrant analyses alone; therefore, the results of these tests were combined.

The chemical data from each survey indicated that samples contained mixtures of contaminants, including trace metals, PAHs, and chlorinated hydrocarbons. The numbers of samples analyzed for each chemical ranged from 399 to 1,060 (Table 2). Analyte concentrations exceeded the MDL in a majority of the samples. The concentrations of most trace metals ranged over two to three orders of magnitude, and those of most organic compounds ranged over four to six orders of magnitude. Concentrations of the PAHs were most often less than the MDL.

None of the samples exceeded the ERM value for arsenic and <1.0% exceeded the ERMs for cadmium and chromium (Table 2). Relatively small proportions of the samples had chemical concentrations that exceeded ERM values, indicating that the data were not skewed toward waste sites with unusually high concentrations. Undoubtedly, some samples contained chemicals that were not quantified or for which there were no SQGs.

Among the different tests performed, 15 to 91% of the samples were at least marginally toxic (Table 3). Bioassay results showed a wide range of response, often from 0 to >100% of mean control responses. In the amphipod tests 36 to 52% of the samples were toxic whereas in the tests of pore water 56 to 91% of samples were toxic.

The frequency distributions of the data from most of the tests were similar, that is, responses in most samples were >80% of control responses (Table 3). Many of the EMAP samples were marginally toxic in amphipod tests. The data from embryological tests with the purple urchin (*S. purpuratus*) and red abalone (*H. rufescens*) indicated similar frequency distributions, both suggesting higher sensitivities to the samples than found in the amphipods. Empirically derived MSDs for each bioassay were very similar, ranging from 80 to 87%.

Incidence of toxicity

Concentrations greater than individual SQGs. Table 4 summarizes the percentages of samples that were not toxic, were marginally toxic, and were highly toxic in the amphipod tests alone and in any of the two to four tests performed when the concentrations of substances equaled or exceeded individ-

Table 2. Ranges in chemical concentrations, numbers of samples in which concentrations were less than or greater than method detection limits (MDLs), and percentages of samples in which effects range—median (ERM) values were exceeded

Chemical[a]	Units	No. samples	% > ERM[b]	No. > MDL	Range in detected concentrations		Range in concn. below detection limits[c]		No. < MDL
					Lowest	Highest	Lowest	Highest	
Arsenic	ppm	920	0.0	913	0.1	41	1.2	1.7	7
Cadmium	ppm	987	0.2	987	0.03	19.8	0.01	0.05	0
Chromium	ppm	1,058	0.5	1,045	1	1,220	1.2	18	13
Copper	ppm	1,057	2.4	1,031	0.7	1,770	0.2	1	26
Lead	ppm	1,052	3.4	1,038	1.4	510	0.3	1.3	14
Mercury	ppm	994	12.7	994	0.01	15	0.001	0.01	0
Nickel	ppm	1,042	2.1	1,006	0.3	136	0.2	1.7	36
Silver	ppm	919	4.4	866	0.01	10.1	0.01	0.7	53
Zinc	ppm	1,060	5.3	1,060	1	1,880	NA	NA	0
2-Methylnaphthalene	ppb	921	1.0	591	0.40	15,557	0.20	10	330
Dibenz[*a,h*]anthracene	ppb	399	11.8	363	0.40	4,534	0.70	10	36
Acenaphthene	ppb	977	3.3	394	0.10	56,338	0.10	80	583
Acenaphthylene	ppb	807	2.8	254	0.40	12,915	0.02	100	553
Anthracene	ppb	997	4.8	521	0.20	89,366	0.03	90	476
Benz[*a*]anthracene	ppb	996	7.2	652	0.30	59,298	0.02	130	344
Benzo[*a*]pyrene	ppb	980	10.0	631	0.20	54,862	0.02	170	349
Chrysene	ppb	997	5.5	688	0.20	60,331	0.10	130	309
Fluoranthene	ppb	1,000	4.2	755	0.30	108,236	0.20	110	245
Fluorene	ppb	945	3.2	530	0.10	54,209	0.10	90	415
Naphthalene	ppb	900	0.9	456	0.70	17,414	0.40	70	444
Phenanthrene	ppb	1,054	5.1	779	0.40	194,343	0.40	90	275
Pyrene	ppb	1,029	8.1	819	0.40	143,132	0.10	120	210
Total LMW PAHs	ppb	956	5.0	956	0.2	552,124	NA	NA	0
Total HMW PAHs	ppb	925	8.2	925	2	461,675	NA	NA	0
Total PAHs	ppb	1,003	1.1	1,003	0.2	1,013,799	NA	NA	0
p,p'-DDE	ppb	789	12.0	741	0.004	2,900	0.03	0.3	48
p,p'-DDD	ppb	742	No ERM	666	0.004	784	0.1	1	76
p,p'-DDT	ppb	656	No ERM	543	0.004	3,517	0.02	1	113
Total DDTs	ppb	813	13.2	813	0.01	4,631	NA	NA	0
Total PCBs	ppb	830	23.4	830	0.1	16,675	NA	NA	0
Dieldrin	ppb	615	No ERM	490	0.002	21.2	0.03	0.5	125
Lindane	ppb	533	No ERM	306	0.01	157	0.05	1	227

[a] LMW = low-molecular-weight, PAH = polynuclear aromatic hydrocarbon, HMW = high-molecular-weight, DDE = dichlorodiphenyldichloroethylene, DDT = dichlorodiphenyltrichloroethane, PCB = polychlorinated biphenyl.
[b] Percent of samples with detectable concentrations.
[c] NA = not applicable for summed concentrations.

results occurred in amphipod tests in 40 to 65% of the samples. The percentages of samples that were highly toxic in amphipod tests ranged from 40% for the ERM value for total PCB to 100% for the cadmium and chromium ERMs. The target percent of false positives (≤25% not toxic) was observed for 13 of the ERMs. The ERMs for six substances correctly classified ≥75% of samples as highly toxic in amphipod tests. Marginally toxic samples contributed relatively little (0–20%) to over-

Table 3. Frequency distribution of toxicity responses (expressed as percent of the total number of samples tested within categories of toxicologic responses), incidence of toxicity, and minimum significant differences (MSDs) for each test

Test medium/species[a]	Endpoint	Duration	*n*	% Control response					% Samples toxic[b]	MSD value
				<20%	20–39.99%	40–59.99%	60–80%	>80%		
Solid phase										
Ampelisca abdita—NOAA	Survival	10 d	289	6.6	4.8	4.5	11.8	72.3	36.3	80
A. abdita—EMAP	Survival	10 d	611	1.4	1.0	2.3	12.1	83.1	38.3	80
Rhepoxynius abronius	Survival	10 d	166	6.0	8.4	6.0	18.7	60.8	51.8	80
Solvent extract										
Photobacterium phosphoreum	Bioluminescence	15 min	224	17.4	12.1	9.8	12.9	47.8	44.6	80
Elutriate										
Mulinea lateralis	Survival	48 h	100	1.0	8.0	12.0	11.0	68.0	29.0	80
M. lateralis	Normal development	48 h	100	7.0	3.0	0.0	1.0	89.0	15.0	80
Porewater										
Arbacia punctulata	Fertilization	1 h	168	24.4	5.9	5.4	5.4	58.9	56.0	87
Strongylocentrotus purpuratus	Normal development	1 h	52	86.5	0.0	3.8	1.9	7.7	90.4	85
Haliotis rufescens	Normal development	48 h	45	71.1	4.4	4.4	6.7	13.3	91.1	85

[a] NOAA = National Oceanic and Atmospheric Administration, EMAP = Environmental Monitoring and Assessment Program.
[b] Marginally + highly toxic ($p < 0.05$, *t* tests).

Table 4. Incidence of toxicity in either amphipod tests alone or any of the two to four tests performed among samples in which individual effects range—median (ERM) values were exceeded

Chemical[a]	Amphipod tests (n = 1,068)					Any test performed[b] (n = 437)				
	No.	% Not toxic	% Marginally toxic	% Highly toxic	% Total toxic	No.	% Not toxic	% Marginally toxic	% Highly toxic	% Total toxic
Metals										
Cadmium	2	0	0	100	100	0	NA	NA	NA	NA
Chromium	5	0	0	100	100	2	0	0	100	100
Copper	25	48	0	52	52	22	18	0	82	82
Lead	35	17	6	77	83	20	5	0	95	95
Mercury	126	34	12	54	66	81	10	6	84	90
Nickel	21	24	14	62	76	5	0	0	100	100
Silver	38	34	18	47	65	22	0	14	86	100
Zinc	56	34	5	61	66	32	13	0	88	88
PAHs										
2-Methylnaphthalene	6	0	17	83	100	4	0	0	100	100
Dibenz[*a*,*h*]anthracene	43	28	2	70	72	31	19	0	81	81
Acenaphthene	13	23	15	62	77	7	0	0	100	100
Acenaphthylene	7	0	14	86	100	6	0	0	100	100
Anthracene	25	24	20	56	76	19	11	0	89	89
Benz[*a*]anthracene	47	23	13	64	77	30	10	0	90	90
Benzo[*a*]pyrene	63	37	8	56	64	46	17	0	83	83
Chrysene	38	32	13	55	68	26	8	0	92	92
Fluoranthene	32	28	13	59	72	21	5	0	95	95
Fluorene	17	29	12	59	71	10	10	0	90	90
Naphthalene	4	25	0	75	75	4	0	25	75	100
Phenanthrene	40	25	15	60	75	25	4	0	96	96
Pyrene	66	33	9	58	67	46	11	0	89	89
Sum LMW PAHs	48	21	13	67	80	31	3	6	90	96
Sum HMW PAHs	76	39	9	51	60	56	16	0	84	84
Sum total PAHs	11	9	18	73	91	6	0	0	100	100
Chlorinated hydrocarbons										
p,p′-DDE	89	45	7	48	55	70	9	6	86	92
Total DDTs	107	38	11	50	61	82	5	9	87	96
Total PCBs	194	49	11	40	51	162	17	6	78	84

[a] PAH = polynuclear aromatic hydrocarbon, LMW = low-molecular weight, HMW = high-molecular-weight, DDE = dichlorodiphenyldichloroethylene, DDT = dichlorodiphenyltrichloroethane, PCB = polychlorinated biphenyl.
[b] Excludes Environmental Monitoring and Assessment Program and Newark Bay samples; NA = not applicable.

all predictive ability. However, based upon sums of the marginally toxic and highly toxic responses, the number of ERMs that correctly predicted toxicity in ≥75% of samples increased from 6 to 13.

Relative to results of the amphipod tests, predictive ability increased considerably when the results were considered for all of the tests performed: ≥75% for all substances that exceeded the ERM concentrations (Table 4). The target percent of false positives (≤25%) was observed for all ERMs and was ≤10% for 18 substances. As with the amphipod data, the marginally toxic results in all tests performed contributed relatively little to overall predictive ability; that is, the samples often were either nontoxic or highly toxic.

Predictive ability observed with the individual PELs was slightly lower than that of equivalent ERMs (Table 5). The percentages of samples exceeding PELs that were highly toxic in amphipod tests ranged from 15% (lindane) to 73% (dieldrin). For 25 of the 31 PELs, highly toxic conditions in amphipod tests occurred in 40 to 65% of the samples. Predictive ability of ≥75% was observed for none of the PELs with only highly toxic responses and with three PELs (cadmium, acenaphthylene, and dieldrin) with marginally plus highly toxic responses combined. The target percent of false positives (≤25%) was observed for the same three PELs. When the results of any of the tests performed were considered, the percent of false positives for the PELs was ≤25% for all except one substance (p,p′-dichlorodiphenyldichloroethylene [p,p′-DDE]) and was ≤10.0% for 15 PELs. For most substances, marginally toxic results contributed 5 to 10% to overall predictive ability in both the amphipod tests alone and in all tests considered. Predictive ability of ≥75% (with highly toxic responses) was observed in any of the tests performed for all PELs except that for p,p′-DDE.

Concentrations above and below all ERL or TEL concentrations. Among the 329 samples in which none of the chemical concentrations exceeded any ERL values, 68% were not toxic, 21% were marginally toxic, and 11% were highly toxic in the amphipod tests (Table 6). Among samples in which multiple bioassays were performed, 46% were not toxic in all tests and 41% were highly toxic in at least one test when all chemical concentrations were less than the ERLs.

Of the samples tested with amphipods, 448 were found in which one or more of the 24 concentrations were greater than or equal to the ERL, but none of the concentrations were greater than or equal to the ERM values; 63% were nontoxic, 20% were marginally toxic, and 18% were highly toxic. A total of 64% of 173 samples was highly toxic in any test performed when one or more ERLs was exceeded and no ERMs were exceeded. The percent of false positives for one or more ERLs exceeded was 63% for amphipod tests alone and 20% for all tests performed.

Generally, the incidence of toxicity increased with the number of chemicals greater than or equal to the ERL concentrations; however, this pattern was variable and inconsistent (Ta-

Table 5. Incidence of toxicity in either amphipod tests alone or any of the two to four tests performed among samples in which individual probable effects levels (PELs) were exceeded

Chemical[a]	Amphipod tests (n = 1,068)					Any test performed[b] (n = 437)				
	No.	% Not toxic	% Marginally toxic	% Highly toxic	% Total toxic	No.	% Not toxic	% Marginally toxic	% Highly toxic	% Total toxic
Metals										
Cadmium	21	19	10	71	81	6	0	0	100	100
Chromium	41	34	7	59	66	24	8	0	92	92
Copper	179	41	11	48	59	146	13	6	81	87
Lead	122	37	11	52	63	85	8	6	86	92
Mercury	127	35	12	54	66	82	11	6	83	89
Nickel	74	34	12	54	66	37	5	5	89	94
Silver	109	41	10	49	59	82	12	11	77	88
Zinc	126	38	10	52	62	87	14	2	84	86
PAHs										
2-Methylnaphthalene	47	28	13	60	73	22	5	9	86	95
Dibenz[*a*,*h*]anthracene	80	36	3	61	64	65	15	2	83	85
Acenaphthene	84	38	8	54	62	56	5	7	88	95
Acenaphthylene	47	23	9	68	77	40	3	8	90	98
Anthracene	131	44	7	49	56	100	11	5	84	89
Benz[*a*]anthracene	116	39	9	52	61	93	12	4	84	88
Benzo[*a*]pyrene	126	41	9	50	59	100	12	3	85	88
Chrysene	116	43	9	47	56	93	12	4	84	88
Fluoranthene	103	42	10	49	59	80	13	5	83	88
Fluorene	74	30	12	58	70	51	6	6	88	94
Naphthalene	38	26	11	63	74	23	0	13	87	100
Phenanthrene	106	40	11	49	60	77	8	5	87	92
Pyrene	117	40	9	51	60	94	11	4	85	89
Sum LMW PAHs	117	36	9	55	64	79	8	5	87	92
Sum HMW PAHs	114	42	7	51	58	90	12	3	84	87
Sum total PAHs	56	32	11	57	68	38	11	0	89	89
Chlorinated hydrocarbons										
p,p'-DDE	3	67	0	33	33	3	33	0	67	67
p,p'-DDD	144	35	11	54	65	115	8	7	85	92
p,p'-DDT	97	33	11	56	67	68	6	7	87	94
Total DDTs	101	36	12	52	64	78	5	9	86	95
Total PCBs	191	50	10	39	49	159	17	6	77	83
Dieldrin	41	20	7	73	80	25	4	0	96	96
Lindane	54	81	4	15	19	50	14	0	86	86

[a] PAH = polynuclear aromatic hydrocarbon, LMW = low-molecular-weight, HMW = high-molecular-weight, DDE = dichlorodiphenyldichloroethylene, DDD = dichlorodiphenyldichloroethane, DDT = dichlorodiphenyltrichloroethane, PCB = polychlorinated biphenyl.
[b] Excludes Environmental Monitoring and Assessment Program and Newark Bay samples.

ble 6). Because of the relatively small numbers of samples in which many ERLs were exceeded, the incidence of toxicity also was calculated for several combined ERL categories. In the amphipod tests (n = 777), the incidence of highly toxic responses was 9% with only 1 ERL exceeded, 13% with 1 to 4 ERLs exceeded, 22% with 5 to 9 ERLs exceeded, and peaked at 67% with 15 to 19 ERLs exceeded.

The proportion of samples that was highly toxic in any test performed was 67% when only one ERL was exceeded (Table 6). The incidence of highly toxic samples increased quickly with the number of ERLs that were exceeded, reaching ≥89% when 10 to 14 concentrations were greater than or equal to the ERLs. With several exceptions (notably one sample in which 22 ERLs were exceeded), generally the proportions of samples that were marginally toxic decreased with increases in the number of concentrations greater than or equal to the ERLs.

Among the 233 samples in which all concentrations were less than the TELs; 65% were not toxic, 26% were marginally toxic, and 9% were highly toxic in amphipod tests (Table 7). A total of 62% of samples (n = 26) were not toxic in all tests performed when all concentrations were less than the TELs. The incidence of toxicity did not increase consistently in either amphipod tests alone or in any tests with increases in the number of TELs exceeded. Sample sizes in which multiple bioassays were performed were relatively small and, partly as a consequence, results were highly variable.

Concentrations above and below all ERM and PEL concentrations. Among the 1,068 samples included in this analysis, 777 and 683 had chemical concentrations less than all ERMs and less than all PELs, respectively (Tables 8 and 9). In amphipod tests, 15 and 13%, respectively, of these samples were highly toxic (false negatives). The incidence of highly toxic responses when one or more concentrations was greater than or equal to the ERM or greater than or equal to the PEL was 39 and 35%, respectively, in amphipod tests and 78 and 77%, respectively, in any test performed. With both the marginally and highly toxic responses combined, the incidence of toxicity in samples with concentrations greater than or equal to one or more ERMs or PELs increased slightly to 52 and 48%, respectively, in the amphipod tests and 86.2 and 86.1%, respectively, in any test.

In both the amphipod tests and any tests performed, the incidence of highly toxic responses generally increased and the incidence of marginally toxic responses markedly decreased with increases in the numbers of ERMs or PELs that were exceeded (Tables 8 and 9). The incidence of highly toxic responses in amphipod tests increased from 23% with only 1

Table 6. Incidence of toxicity in either amphipod tests alone or in any test performed among samples with concentrations of 0 to 24 substances greater than or equal to the effects range—low (ERL) values, but all less than the effects range—median (ERM) values

	Amphipod survival only (n = 777)				Any test performed[a] (n = 212)			
No. ERL values exceeded	No. samples	% Not toxic	% Marginally toxic	% Highly toxic	No. samples	% Not toxic	% Marginally toxic	% Highly toxic
0	329	68	21	11	39	46	13	41
1	143	68	23	9	15	13	20	67
2	66	71	15	14	13	46	8	46
3	37	62	22	16	12	42	17	42
4	43	63	16	21	21	33	14	52
5	30	60	17	23	13	38	8	54
6	33	64	12	24	24	21	13	67
7	20	55	35	10	15	27	20	53
8	15	53	27	20	12	8	33	58
9	8	50	13	38	6	0	33	67
10	9	89	0	11	6	0	0	100
11	12	42	25	33	7	0	14	86
12	9	78	11	11	8	0	13	88
13	2	0	0	100	2	0	0	100
14	4	25	50	25	3	0	33	67
15	2	50	50	0	2	0	50	50
17	1	0	0	100	1	0	0	100
18	2	0	0	100	2	0	0	100
19	1	0	0	100	1	0	0	100
20	4	50	0	50	3	0	0	100
21	4	25	25	50	4	0	0	100
22	1	0	100	0	1	0	100	0
23	1	0	0	100	1	0	0	100
24	1	0	0	100	1	0	0	100
1 or more	448	62.7	19.6	17.6	173	20.2	15.6	64.2
1 to 4	289	67.1	20.1	12.8	61	32.8	14.8	52.5
5 to 9	106	58.5	19.8	21.7	70	21.4	18.6	60.0
10 to 14	36	58.3	16.7	25.0	26	0.0	11.5	88.5
15 to 19	6	16.7	16.7	66.7	6	0.0	16.7	77.8
20 to 24	11	27.3	18.2	54.5	10	0.0	10.0	90.0

[a] Excludes Environmental Monitoring and Assessment Program and Newark Bay data.

ERM exceeded to 32% with 1 to 5 ERMs exceeded, to 52% with 6 to 10 ERMs exceeded, and peaked at 85% with ≥11 ERMs exceeded (Table 8). The lowest percent false positives (10%) occurred among samples with 11 to 20 ERMs exceeded. In samples in which multiple bioassays were performed, incidence of highly toxic responses increased from 70% with only 1 ERM exceeded, to 89% with 6 to 10 ERMs exceeded, and peaked at 100% with ≥11 ERMs exceeded. Results were variable among samples with greater than or equal to eight ERMs exceeded because of the small sample sizes.

The predictive ability of the PELs was somewhat lower than that of the ERMs, but, nevertheless, indicated a similar pattern of increasing incidence of highly toxic responses with increasing numbers of PELs exceeded (Table 9). In the amphipod tests, the incidence of highly toxic responses was 14% with 1 PEL exceeded, 24% with 1 to 5 PELs exceeded, 40% with 6 to 10 PELs exceeded, 50% with 11 to 20 PELs exceeded, and 88% with ≥21 PELs exceeded. The lowest percent false positives (17%) occurred among samples with ≥21 PELs exceeded. The proportion of samples showing highly toxic results was much higher when all bioassays were considered, averaging 80% with 6 to 10 PELs exceeded and peaking at 100% with ≥21 PELs exceeded. Percent false positives in any of the tests performed was <25% when one or more PEL was exceeded.

Over ranges in mean SQG quotients. In the preceeding analyses, the methods did not account for the degree to which the chemical concentrations exceeded the different SQGs. That is, samples in which chemical concentrations exceeded SOGs by very different amounts were scored the same. Given similar sediment characteristics and toxicant bioavailability, the probability of toxicity could increase with increasing concentrations. Therefore, to account for both the actual concentrations of individual substances and the combinations of chemicals occurring as mixtures, the predictive abilities of the mean SQG quotients were determined.

The relationships between the incidence of highly toxic responses in the amphipod tests and mean SQG quotients are illustrated in Figures 1 and 2. To clarify these relationships, the chemical concentrations are shown as medians of 39 SQG quotient intervals, each consisting of at least 25 samples. These relationships were considerably more variable when marginally toxic responses were included; therefore, the plots are shown only for highly toxic responses. The incidence of highly toxic responses was most variable and ranged from 0 to 40% among samples with the lowest mean ERM quotients (0.001–0.02) and PEL quotients (0.006–0.05). A gradual, albeit variable, pattern of increasing incidence of toxicity beginning at mean ERM and PEL quotients of 0.04 and 0.07, respectively, was evident. Among samples with mean ERM or PEL quotients ≥1.0 or ≥1.6, respectively, 60 to 80% were highly toxic in the amphipod tests. Percent false positives decreased to <25% with mean ERM or PEL quotients >1.2 or >2.3, respectively.

Some of the samples with the lowest mean ERM and PEL quotients were highly toxic, as indicated in the left tails of the distributions (Figs. 1 and 2). These samples shared very few of the same characteristics. They were scattered among many

Table 7. Incidence of toxicity in either amphipod tests alone or in any test performed among samples with concentrations of 0 to 27 substances greater than or equal to the threshold effects level (TEL) values, but all less than the probable effects level (PEL) values

No. TEL values exceeded	Amphipod survival only (n = 683)				Any test performed[a] (n = 142)			
	No. samples	% Not toxic	% Marginally toxic	% Highly toxic	No. samples	% Not toxic	% Marginally toxic	% Highly toxic
0	233	65	26	9	26	62	15	23
1	102	74	15	12	9	22	11	67
2	67	67	24	9	5	40	20	40
3	62	69	21	10	10	40	30	30
4	46	65	11	24	5	0	0	100
5	28	61	25	14	7	29	14	57
6	15	53	33	13	4	50	0	50
7	10	70	20	10	5	20	20	60
8	15	73	7	20	6	17	0	83
9	5	60	20	20	3	0	0	100
10	12	67	8	25	6	33	17	50
11	11	27	27	45	6	17	17	67
12	15	67	0	33	11	27	0	73
13	10	90	0	10	5	40	0	60
14	7	71	29	0	4	50	25	25
15	3	33	33	33	2	50	50	0
16	4	75	25	0	1	0	100	0
17	2	50	50	0	1	0	0	100
18	4	75	0	25	2	0	50	50
19	8	63	25	13	3	33	0	67
20	5	40	20	40	4	0	25	75
21	4	0	75	25	4	0	25	75
22	3	33	33	33	3	33	33	33
23	9	33	44	22	7	0	43	57
24	1	0	0	100	1	0	0	100
25	1	0	0	100	1	0	0	100
27	1	0	100	0	1	0	100	0
1 or more	450	65.1	19.1	15.8	116	23.3	17.2	59.5
1 to 5	305	68.9	18.4	12.8	36	27.8	16.7	55.6
6 to 9	45	64.4	20.0	15.6	18	22.2	5.6	72.2
10 to 14	55	63.6	10.9	25.5	32	31.3	9.4	59.4
15 to 19	21	61.9	23.8	14.3	9	22.2	33.3	44.4
20 to 27	24	25.0	41.7	33.3	21	4.8	33.3	61.9

[a] Excludes Environmental Monitoring and Assessment Program and Newark Bay data.

Table 8. Incidence of toxicity in either amphipod tests alone or in any of two to four tests performed among samples with concentrations of 0 to 20 substances greater than or equal to the effects range—median (ERM) concentrations

No. ERM values exceeded	Amphipod survival only (n = 1,068)				Any test performed[a] (n = 437)			
	No. samples	% Not toxic	% Marginally toxic	% Highly toxic	No. samples	% Not toxic	% Marginally toxic	% Highly toxic
0	777	65	20	15	212	25	15	60
1	95	59	18	23	69	17	13	70
2	66	52	12	36	62	11	11	77
3	34	56	21	24	30	17	7	77
4	19	32	5	63	10	20	0	80
5	11	45	0	55	9	11	0	89
6	11	55	9	36	10	20	0	80
7	10	40	0	60	8	13	0	88
8	4	25	25	50	4	0	0	100
9	11	9	9	82	7	0	0	100
10	10	50	20	30	6	17	0	83
11	4	0	25	75	1	0	0	100
12	6	33	0	67	3	0	0	100
13	4	0	0	100	3	0	0	100
14	3	0	0	100	1	0	0	100
15	1	0	0	100	1	0	0	100
17	1	0	0	100	0	0	0	0
20	1	0	0	100	1	0	0	100
1 or more	291	47.8	13.4	38.8	225	13.8	8.0	78.2
1 to 5	225	53.3	14.7	32.0	180	15.0	10.0	75.0
6 to 10	46	37.0	10.9	52.2	35	11.4	0.0	88.6
11 to 20	20	10.0	5.0	85.0	10	0.0	0.0	100.0

[a] Excludes Environmental Monitoring and Assessment Program and Newark Bay data.

Table 9. Incidence of toxicity in either amphipod tests alone or in any of two to four tests performed among samples with concentrations of 0 to 20 substances greater than or equal to the probable effects level (PEL) concentrations

	Amphipod survival only (n = 1,148)				Any test performed[a] (n = 517)			
No. PEL values exceeded	No. samples	% Not toxic	% Marginally toxic	% Highly toxic	No. samples	% Not toxic	% Marginally toxic	% Highly toxic
0	683	65	22	13	142	30	17	53
1	106	73	13	14	79	15	10	75
2	49	45	18	37	42	19	14	67
3	36	53	14	33	25	12	16	72
4	17	47	29	24	11	18	36	45
5	16	56	13	31	13	15	0	85
6	10	30	20	50	7	0	0	100
7	4	25	25	50	3	0	33	67
8	18	56	6	39	16	31	0	69
9	11	55	0	45	7	0	0	100
10	9	67	11	22	7	29	0	71
11	19	53	16	32	17	12	0	88
12	8	38	13	50	7	14	0	86
13	13	54	8	38	11	18	0	82
14	8	38	13	50	8	0	13	88
15	11	36	9	55	10	20	0	80
16	8	25	13	63	7	0	14	86
17	7	43	0	57	5	0	0	100
18	10	40	0	60	8	0	13	88
19	5	0	20	80	2	0	0	100
20	3	0	33	67	3	0	0	100
21	2	0	0	100	1	0	0	100
22	5	0	0	100	1	0	0	100
23	5	20	0	80	1	0	0	100
24	4	25	0	75	3	0	0	100
26	1	0	0	100	1	0	0	100
1 or more	385	51.7	13.0	35.3	295	13.9	8.8	77.3
1 to 5	224	60.3	15.6	24.1	170	15.9	12.9	71.2
6 to 10	52	50.0	9.6	40.4	40	17.5	2.5	80.0
11 to 20	92	39.1	10.9	50.0	78	9.0	3.8	87.2
21 to 26	17	11.8	0.0	88.2	7	0.0	0.0	100.0

[a] Excludes Environmental Monitoring and Assessment Program and Newark Bay data.

of the different NOAA and EMAP study areas. These samples often, but not always, had relatively low organic carbon content (<1.0%) and percent fine-grained materials (<50%) and detectable concentrations of butyl tins, chlorinated pesticides, alkyl-substituted PAHs, ammonia, or other substances not accounted for with the SQGs.

DISCUSSION AND CONCLUSIONS

Sediment quality guidelines [1,2] were based upon empirical analyses of data compiled from many different studies. The SQGs were intended to provide informal (nonregulatory), effects-based benchmarks to aid in the interpretation of sed-

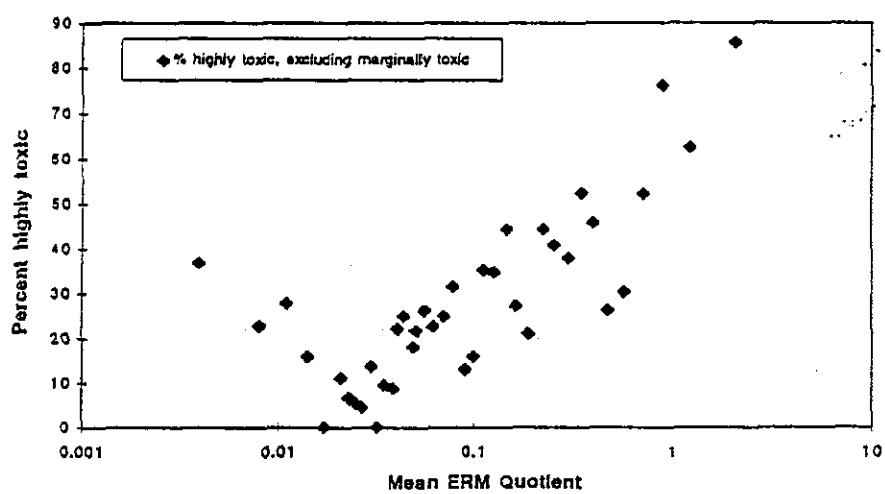

Fig. 1. The relationship between the incidence of toxicity in amphipod survival tests and mean effects range—median (ERM) quotients (plotted as the medians of 39 quotient intervals, each consisting of 25 samples).

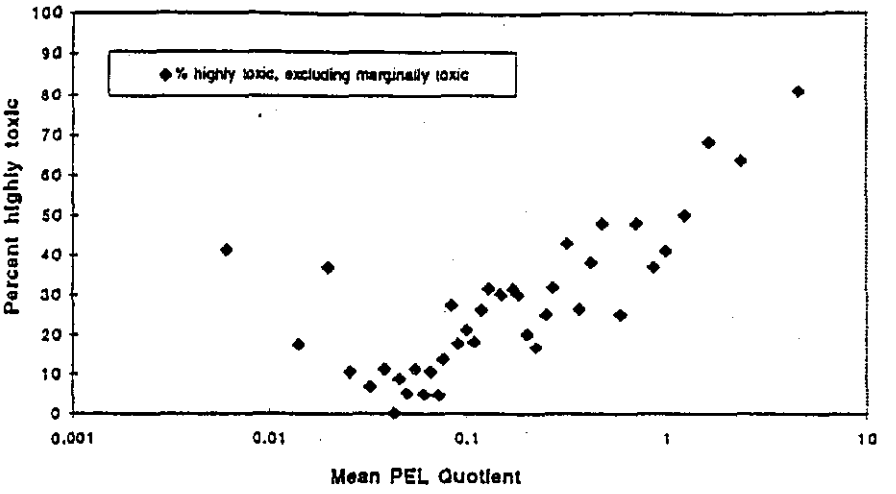

Fig. 2. The relationship between the incidence of toxicity in amphipod survival tests and mean probable effects level (PEL) quotients (plotted as the medians of 39 quotient intervals, each consisting of 25 samples).

iment chemistry data. The ERL and TEL values were intended to represent chemical concentrations below which the probability of toxicity and other effects was minimal. In contrast, the ERM and PEL values were intended to represent mid-range concentrations above which adverse effects were more likely, although not always expected. Intermediate frequencies of effects were expected at chemical concentrations between the ERLs and ERMs and between the TELs and PELs. In this analysis of independent data sets, we attempted to determine if the incidence of toxicity in selected, acute laboratory bioassays would follow the same pattern as observed with multiple measures of effects in the databases used to derive the guidelines.

The majority of the data compiled to develop the guidelines was generated in field studies in which different chemical mixtures were encountered. In these field studies causality could not be determined. The intent of this study was to also use data from surveys of numerous saltwater areas to determine the frequency with which the guidelines correctly predicted nontoxic and toxic conditions.

Unlike SQGs based upon the apparent effects threshold approach [26], the ERL/ERMs and TEL/PELs were not intended to represent concentrations above which adverse effects were always observed. Because the ERLs and TELs were intended to represent conservative concentrations below which toxicity was not frequently expected, we estimated the frequency of false negatives as the incidence of toxicity among samples in which all concentrations were lower than these values. Earlier [1,2], as a measure of reliability, we reported that the frequency of false negatives among the data sets used to derive the guidelines was ≤25% for most chemicals and ≤10% for many chemicals. Specifically, at concentrations below the individual ERL and TEL values for nine trace metals, the incidence of effects ranged from 1.9 to 9.4% and from 2.7 to 9.0%, respectively. For organic compounds, the incidence of effects was more variable, ranging from 5.0 to 27.3% for 19 ERLs and from 0.0 to 47.6% for 25 TELs when concentrations were below these levels.

The same criterion (≤25% false positives) previously used for estimates of reliability was used as the target for estimates of predictive ability in this analysis. Based upon the highly toxic responses, the ERLs and TELs indicated 11 and 9% false negatives (toxicity observed when not expected), respectively, in the tests of amphipod survival, thus bettering the target of ≤25%. The incidence of false negatives also was relatively low (41 and 23% for the ERLs and TELs, respectively) in any one of the two to four tests performed. Based again upon the highly toxic responses, the incidences of false negatives in amphipod tests were, as expected, slightly higher (15 and 13%, respectively) for the ERMs and PELs than for the ERLs and TELs. Therefore, the probabilities of highly toxic responses in amphipod survival tests are relatively low (≤16%) among samples in which all chemical concentrations are lower than both sets of SQGs. However, the incidences of false negatives among any of the tests performed were 60 and 53% (highly toxic responses) for the ERMs and PELs, respectively. These data suggest that there remains a moderate probability of toxicity among samples with all chemical concentrations less than the ERMs or less than the PELs if a battery of relatively sensitive, sublethal bioassays is considered.

In the amphipod tests, the incidences of highly toxic responses and total toxic responses were 18 to 20% and 16 to 19%, respectively, when one or more chemicals exceeded the ERLs and/or TELs. These results agreed well with the original intent of the ERLs and TELs as indicators of the lower end of the possible effects range. These results also agreed very well with the estimates of reliability (calculated with the database used to derive the SQGs) for most ERLs and TELs (30–50% effects) [1,2]. However, when predictive ability was estimated with data from more sensitive sublethal tests, toxicity was observed much more frequently than in the amphipod tests alone.

The ERMs and PELs were derived as mid-range points within the distributions of effects data for each chemical. The ERMs were calculated as the medians (50th percentiles) of chemical concentrations associated with measures of adverse effects. The derivation of the PELs incorporated both the no-effects data along with effects data into the calculations of mid-range concentrations. Neither set of guidelines was intended as a toxicity threshold above which effects were always expected. The incidence of highly significant toxicity in the amphipod survival tests among samples that exceeded individual ERMs and PELs generally agreed with the intent of these values (i.e., as mid-range values). That is, 40 to 65% of

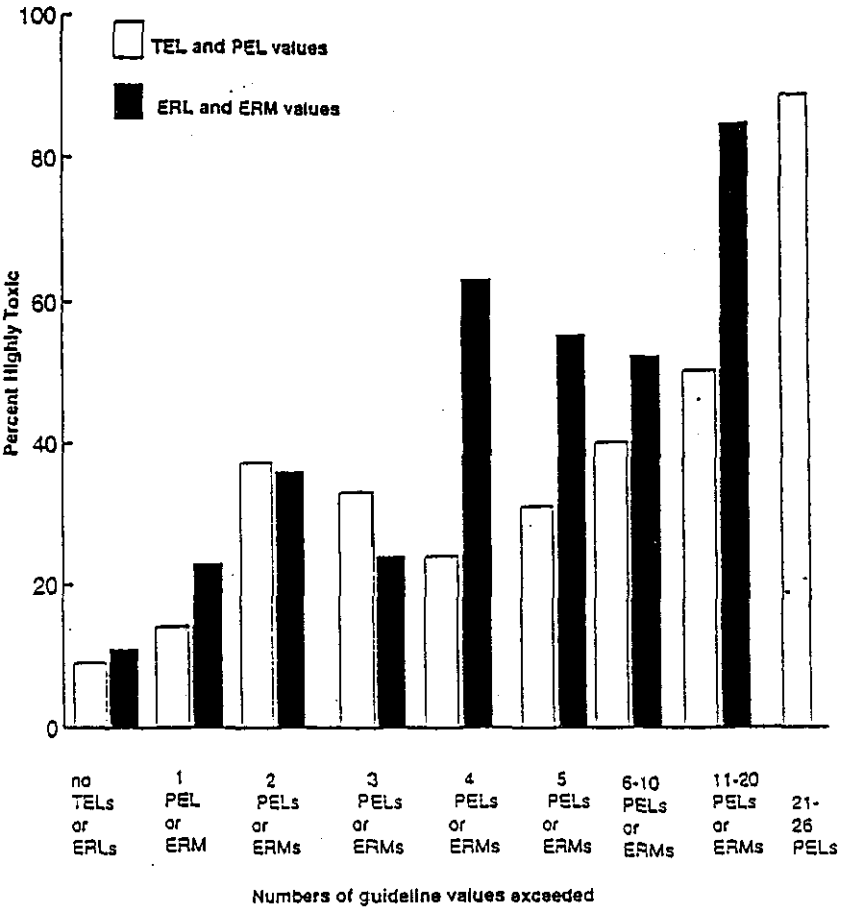

Fig. 3. Summary of the predictive ability of threshold effects level/probable effects level (TEL/PEL) values and effects range—low/effects range—median (ERL/ERM) values in amphipod survival tests (as percent highly toxic among the total numbers of samples).

the samples were highly toxic in amphipod tests at concentrations above most of these individual values. Also, the incidence of total toxicity (marginally + highly toxic) was 52 and 48% when the concentrations of one or more chemicals exceeded ERMs and PELs, respectively. When results from any one of a battery of bioassays were considered, the percentages of samples that were highly toxic increased remarkably to ≥85% for 19 of the ERMs and for 19 of the PELs and to 77 to 78% when one or more ERMs and/or PELs were exceeded.

In all analyses performed on the predictive ability of the SQGs, the percentages of samples demonstrating toxicity were lowest when either no chemicals or the least number of chemicals exceeded the lower range guidelines and increased with increases in the numbers of mid-range guidelines that were exceeded (Fig. 3). Results were variable at intermediate concentrations, but, nevertheless, the data indicated an overall pattern of increasing incidence of toxicity with increasing numbers of ERMs and PELs exceeded. Percent false positives in amphipod tests (no toxicity observed when toxicity was expected) dropped to <25% among samples in which 11 to 20 ERMs (n = 20) and 21 to 26 PELs (n = 17) were exceeded.

Because the two sets of SQGs were derived with slightly different procedures, one objective of this evaluation was to compare their predictive ability. The results indicated that the two sets of SQGs were very similar in predicting toxicity (Fig. 3). The percentages of false negatives for the ERLs and TELs were 11 and 9%, respectively, in the amphipod tests. The incidences of highly toxic responses in amphipod tests were slightly higher for the PELs than for the ERMs among samples in which two or three chemicals exceeded the guideline concentrations. Otherwise, the incidence of toxicity often was higher when chemical concentrations exceeded the ERMs as compared to when the concentrations exceeded the PELs.

Based upon these data, users of the SQGs can identify the probability that their samples would be toxic by comparing the chemical concentrations in their samples to the appropriate SQGs and then to the incidence of toxicity shown in this paper. For example, highly toxic responses would be expected in amphipod survival tests in only approximately 9 to 11% of the samples when all chemical concentrations are below the TELs or ERLs (Fig. 3). Among samples in which only one ERL or TEL value is exceeded and no other chemicals exceeded any other ERL/ERMs or TEL/PELs, toxicity in amphipod tests would be expected in only 9 and 12% of the samples, respectively.

The probability of toxicity in amphipod survival tests is not very high (23 and 14%, respectively) among samples in which only one ERM or only one PEL value is exceeded (Fig. 3). However, the probabilities of toxicity increase with the number of ERMs and PELs exceeded. Based upon the results of this evaluation (n = 1,068), users can expect toxicity in a large majority of samples, that is, in >85% of the samples in amphipod tests (n = 20, n = 17) and in 100% of samples in any one of a battery of sensitive bioassays (n = 9 or 6) when 11 or more ERMs or 21 or more PELs are exceeded. Therefore,

Table 10. Incidence of toxicity in amphipod tests only within three ranges in mean sediment quality guideline quotients

	No. samples	% Not toxic	% Marginally toxic	% Highly toxic
Mean effects range—median quotients				
<0.1	653	67.8	20.5	11.6
0.11 to 1.0	364	51.6	16.5	31.9
>1.0	51	23.5	5.9	70.6
Mean probable effects level quotients				
<0.1	481	67.6	22.0	10.4
0.11 to 1.0	474	58.6	17.1	24.3
>1.0	113	35.4	8.8	55.8

the probability of incorrectly classifying samples as toxic would be 15 and 0%, respectively, in these highly contaminated samples.

The data from the analyses of the mean SQG quotients suggest that the probability of observing toxicity was a function of not only the number of guidelines exceeded but the degree to which they were exceeded. Therefore, the probabilities of highly toxic responses would be relatively low (<12% in amphipod tests) among samples with mean SQG quotients <0.1 (Table 10). The probabilities of toxicity increase to 32 and 24%, respectively, with mean ERM and PEL quotients of 0.11 to 1.0 and increase again to 71 and 56%, respectively, with quotients >1.0.

Despite the selection of high-quality data sets from NS&TP and EMAP—Estuaries studies, the analyses of predictive ability had a number of limitations or potential sources of error. Different results may have been obtained if other data had been used in this evaluation of predictive ability.

The core bioassay upon which these analyses focused was the amphipod survival test. This bioassay has become the most widely applied sediment toxicity test in North America and provides important information for many research, monitoring, and management programs. Amphipod survival tests have been used in both the derivation and field validation of various guidelines [22,26]. However, because different taxa have different sensitivities to toxicants, the use of a battery of toxicity tests is widely accepted and highly recommended in sediment quality assessments [27]. Furthermore, the use of multiple tests increases the number of surrogates of sediment-dwelling taxa. Considerable gains in predictive ability were attained by the addition of data from other tests to those from the amphipod tests. Because only one, two, or three (not, say, 10) tests accompanied the amphipod bioassays, we attribute the gains in predictive ability not to the number of tests performed, but, rather, to the greater sensitivity of the tests to the chemicals in the sediments.

Tests of invertebrate gametes and embryos exposed to pore waters and bioluminescent bacteria exposed to solvent extracts have been used widely in U.S. estuaries [24] and generally are more sensitive than are test with amphipods to the same samples. The large differences in sensitivity between the amphipod survival tests and the other tests performed is reflected in the data that were analyzed. The probabilities of observing toxicity in the more sensitive sublethal tests would be much higher than in the amphipod tests. Users are advised to consider the data from both categories of bioassays when using the guidelines, especially because highly sensitive tests such as those teria [28] have shown strong associations with chemical concentrations.

Sediment quality guidelines were not available for many substances that were measured in the samples. Some substances may have occurred at concentrations above toxicologic thresholds. Other substances that were not measured probably occurred in many or all samples. Also, some samples may have had high concentrations of ammonia and hydrogen sulfide that covaried with anthropogenic substances and contributed to toxicity. Together, the effects of these substances may have contributed to the false negatives observed. However, our nationwide experience indicates that toxicants often covary with each other to a large degree [7,25] and the quantified substances for which SQGs were available should have served as reasonable surrogates for the covariates. Furthermore, our experience in assessments of surficial sediments suggests that ammonia and sulfides occur in either pore water or overlying water in test chambers at toxicologically significant concentrations in <10% of the samples. Nevertheless, the contribution of all potentially toxic substances in the samples could not be accounted for.

Although standardized and widely accepted methods and protocols were used, some interlaboratory and interstudy differences in methods may have occurred. Some variability in results may have been attributable to merging data from different studies and geographic areas. For example, data were compiled from tests performed with two species of amphipods to increase the sample size and to include data from all three coastlines. Differences in sensitivity between these two amphipod species may have contributed to variability in the results. Also, variability may have been increased by merging data from different species of urchins and molluscs along with data from the Microtox tests into one category.

Most of the samples were not collected within hazardous waste sites and most were not highly contaminated (zero to five SQGs exceeded). The relatively small numbers of highly contaminated samples appeared to contribute to variability in results. Additional data from highly contaminated sites would be useful in further clarification of predictive ability.

Despite these potential limitations of this study, the predictive ability estimated with these data often matched their previously reported reliability. Also, the results of this analysis agreed relatively well with the the estimates of reliability reported [5] for freshwater sediment effects concentrations. The results of this analysis [5] determined type I (false positive) and type II (false negative) errors for freshwater ERL/ERM and TEL/PEL values based upon data from individual samples from numerous studies. For most substances, the errors ranged from 5 to 30%. The paired sets of values, however, differed somewhat in absolute concentrations and error rates.

The toxicity/chemistry relationships observed in this study may not apply in all situations, especially in sediments in which contaminants are found in forms such as copper slag [29] or coal pitch in organically enriched mud [30]. The guidelines are most useful when applied to fine-grained, sedimentary deposits such as those sampled during the NOAA and EMAP—Estuaries studies.

In conclusion, the results of these analyses indicate the following: the probabilities of highly toxic responses occurring in amphipod survival tests among samples in which all chemical concentrations are less than ERLs and/or TELs are 9 to 11%; the probabilities of highly toxic responses occurring in

quotients are <0.1 are 10 to 12%; the probabilities of highly toxic responses occurring when one or more ERLs or TELs are exceeded and no ERMs or PELs are exceeded are 16 to 18% in amphipod tests alone and 60 to 64% in any one of a battery of sensitive tests performed; the probabilities of either marginally or highly toxic responses occurring are 48 to 52% in amphipod tests and 86% in any one of a battery of sensitive tests performed when concentrations exceed one or more ERMs or PELs; consistent with their original intent, the ERMs and PELs are considerably better at predicting toxicity than are the ERLs and TELs. Furthermore, the probabilities of toxicity occurring generally increase with increasing numbers of chemicals that exceed the ERM and PEL concentrations; the probabilities of toxicity occurring generally increase with increasing mean SQG quotients; and the incidence of false negatives is slightly lower for the TELs than for the ERLs, but the incidence of false positives is generally higher for the PELs than for the ERMs; however, there is good overall agreement in the predictive ability of the TEL/PELs and the ERL/ERMs.

Based upon these analyses of predictive ability and previous analyses of reliability, it appears that the SQGs provide reasonably accurate estimates of chemical concentrations that are either nontoxic or toxic in laboratory bioassays. However, we urge that all SQGs should be used with caution, because, as observed in this analysis, they are not perfect predictors of toxicity. Especially among samples with intermediate chemical concentrations, the SQGs are most useful when accompanied by data from in situ biological analyses, other toxicologic assays, other interpretive tools such as metals : aluminum ratios, and other guidelines derived either from empirical approaches and/or cause–effects studies.

Acknowledgement—Funding for much of the NOAA data collection was provided by the Coastal Ocean Program and the National Status and Trends Program, both of NOAA. Helpful comments on an initial version of the mansucript were provided by Chris Ingersoll, Jeff Hyland, R. Scott Carr, Rick Swartz, and Tom O'Connor. The EMAP data were provided by the U.S. EPA; database development was provided by the Office of Research and Development, U.S. EPA, and the Hazardous Materials Response and Assessment Division, NOAA. Funding for data analyses was provided by the Hazardous Materials Response and Assessment Division, and Coastal Monitoring and Bioeffects Assessment Division, NOAA. Corinne Severn and Carolyn Hong provided valuable assistance in data management.

REFERENCES

1. Long ER, MacDonald DD, Smith SL, Calder FD. 1995. Incidence of adverse biological effects within ranges of chemical concentrations in marine and estuarine sediments. *Environ Manage* 19: 81–97.
2. MacDonald DD, Carr RS, Calder FD, Long ER. 1996. Development and evaluation of sediment quality guidelines for Florida coastal waters. *Ecotoxicology* 5:253–278.
3. Long ER, MacDonald DD. 1992. National Status and Trends Program approach. Sediment classification methods compendium. 823-R-92-006. U.S. Environmental Protection Agency, Washington, DC.
4. Smith SL, MacDonald DD, Keenleyside KA, Ingersoll CG, Field LJ. 1996. A preliminary evaluation of sediment quality assessment values for freshwater sediments. *J Great Lakes Res* 22:624–638.
5. Ingersoll CG, et al. 1996. Calculation and evaluation of sediment effect concentrations for the amphipod *Hyalella azteca* and the midge *Chironomus riparius*. *J Great Lakes Res* 22:602–623.
6. Long ER, Wolfe DA, Scott KJ, Thursby GB, Stern EA, Peven C, Schwartz T. 1995. Magnitude and extent of sediment toxicity in the Hudson–Raritan estuary. NOAA Technical Memorandum
7. Wolfe DA, Bricker SB, Long ER, Scott KJ, Thursby GB. 1994. Biological effects of toxic contaminants in sediments from Long Island Sound and environs. NOAA Technical Memorandum NOS ORCA 80. National Oceanic and Atmospheric Administration, Silver Spring, MD, USA.
8. Long ER, Sloane GM, Carr RS, Scott KJ, Thursby GB, Wade TL. 1994. Sediment toxicity in Boston Harbor: Magnitude, extent, and relationships with chemical toxicants. NOAA Technical Memorandum NOS ORCA 96. National Oceanic and Atmospheric Administration, Silver Spring, MD, USA.
9. Long ER, et al. 1994. Magnitude and extent of sediment toxicity in Tampa Bay, Florida. NOAA Technical Memorandum NOS ORCA 78. National Oceanic and Atmospheric Administration, Silver Spring, MD, USA.
10. Fairey R, et al. 1996. Chemistry, toxicity and benthic community conditions in sediments of the San Diego Bay region. Final Report. California State Water Resources Control Board, Sacramento, CA, USA.
11. Sapudar RA, et al. 1994. Sediment chemistry and toxicity in the vicinity of the Los Angeles and Long Beach harbors. Project Report. California State Water Resources Control Board, Sacramento, CA, USA.
12. Schimmel SC, Melzian BD, Campbell DE, Strobel CJ, Benyi SJ, Rosen JS, Buffum HW. 1994. Statistical summary. EMAP—Estuaries. Virginian Province—1991. EPA 620/R-94/005. U.S. Environmental Protection Agency, Washington, DC.
13. Strobel CJ, Buffum HW, Benyi SJ, Petrocelli EA, Reifsteck DR, Keith DJ. 1995. Statistical summary. EMAP—Estuaries. Virginian Province—1990 to 1993. EPA 620/R-94/-26. U.S. Environmental Protection Agency, Narragansett, RI.
14. Summers JK, Macauley JM. 1993. Statistical summary: EMAP—Estuaries Louisianian Province—1991. EPA Report 600/R-93-001. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.
15. Lauenstein GG, Cantillo AY. 1993. Sampling and analytical methods of the National Status and Trends Program. National Benthic Surveillance and Mussel Watch projects. 1984–1992. Vol 1. Overview and summary of methods. NOAA Technical Memorandum NOS ORCA 71. National Oceanic and Atmospheric Administration, Silver Spring, MD, USA.
16. American Society for Testing and Materials. 1990. Standard guide for conducting solid phase, 10-day, static sediment toxicity tests with marine and estuarine infaunal amphipods. E 1367-90. In *Annual Book of ASTM Standards*. Philadelphia, PA.
17. American Society for Testing and Materials. 1993. Standard guide for conducting solid phase, 10-day, static sediment toxicity tests with marine and estuarine infaunal amphipods. E 1367-92. In *Annual Book of ASTM Standards*. Philadelphia, PA.
18. Carr RS, Chapman DC. 1992. Comparison of solid-phase and pore water approaches for assessing the quality of marine and estuarine sediments. *Chem Ecol* 7:19–30.
19. U.S. Environmental Protection Agency. 1990. Recommended protocols for conducting laboratory bioassays on Puget Sound sediments. Revised. Puget Sound Estuary Program, Seattle, WA.
20. Schiewe MH, Hawk EG, Actor DI, Krahn MM. 1985. Use of a bacterial bioluminescence assay to assess toxicity of contaminated marine sediments. *Can J Fish Aquat Sci* 42:1244–1248.
21. Mearns AJ, Swartz RC, Cummins JM, Dinnel PA, Plesha P, Chapman PM. 1986. Inter-laboratory comparison of a sediment toxicity test using the marine amphipod, *Rhepoxynius abronius*. *Mar Environ Res* 19:13–37.
22. Swartz RC, Schults DW, Ozretich RJ, Lamberson JO, Cole FA, DeWitt TH, Redmond MS, Ferraro SP. 1995. SigmaPAH: A model to predict the toxicity of polynuclear aromatic hydrocarbon mixtures in field-collected sediments. *Environ Toxicol Chem* 14: 1977–1987.
23. Thursby GB, Heltshe J, Scott KJ. 1997. Revised approach to toxicity test acceptability criteria using a statistical performance assessment. *Environ Toxicol Chem* 16:1322–1329.
24. Long ER, Robertson A, Wolfe DA, Hameedi J, Sloane GM. 1996. Estimates of the spatial extent of sediment toxicity in major US estuaries. *Environ Sci Technol* 30:3585–3592.
25. Carr RS, Long ER, Windom HL, Chapman DC, Thursby G, Sloane GM, Wolfe DA. 1996. Sediment quality assessment studies

iment quality values refinement: 1988 update and evaluation of Puget Sound AET. Vol 1. PTI Environmental Services, Bellevue, WA, USA

27. Ingersoll CG, et al. 1997. Workgroup summary report on uncertainty evaluation of measurment endpoints used in sediment ecological risk assessment. In Ingersoll CG, Dillon T, Biddinger GR, eds, *Ecological Risk Assessment of Contaminated Sediment*. SETAC, Pensacola, FL, USA.
28. Johnson BJ, Long ER. 1998. Rapid toxicity assessment of sediments from large estuarine ecosystems: A new tandem in vitro testing approach. *Environ Toxicol Chem* (in press).
29. Lee GF, Jones RA. 1991. Overview: Aquatic life risk assessment in copper-contaminated sediments at National City marine terminal. Woodward-Clyde Consultants, San Diego, CA, USA.
30. Paine MD, Chapman PM, Allard PJ, Murdoch MH, Minifie D. 1996. Limited bioavailability of sediment PAH near an aluminum smelter: Contamination does not equal effects. *Environ Toxicol Chem* 15:2003–2018.