6/12/99

Sediment Quality Guidelines developed for the National Status and Trends Program

Background and Intended Uses

Through its National Status and Trends (NS&T) Program, NOAA generates considerable amounts of chemical data on sediments. Without national criteria or other widely-applicable numerical tools, NOAA scientists found it difficult to estimate the possible toxicological significance of chemical concentrations in sediments. Thus, numerical sediment quality guidelines (SQGs) were developed as informal, interpretive tools for the NS&T Program.

The SQGs were initially intended for use by NOAA scientists in ranking areas that warranted further detailed study on the actual occurrence of adverse effects such as toxicity. Also, they were intended for use in ranking chemicals that might be of potential concern. In many regional surveys of sediment toxicity performed throughout North America, NOAA has used the guidelines to compare the degree of contamination among sub-regions, and to identify chemicals elevated in concentration above the guidelines that were also associated with measures of adverse effects.

The SQGs were not promulgated as regulatory criteria or standards. They were not intended as cleanup or remediation targets, nor as discharge attainment targets. Nor were they intended as pass-fail criteria for dredged material disposal decisions or any other regulatory purpose. Rather, they were intended as informal (non-regulatory) guidelines for use in interpreting chemical data from analyses of sediments.

Derivation

SQGs were needed relatively quickly for use in interpreting data from the ongoing NS&T Program studies; thus, existing data were used in their derivation, rather than data from tedious and expensive laboratory tests or modeling approaches. SQGs were needed that could be applied nationwide in the NS&T Program; therefore, data from studies performed throughout North America were assembled and compiled into a database to ensure broad applicability of the guidelines. Because guidelines were needed that were based on measures of biological effects associated with toxicants, data were compiled that included both chemical measures and biological effects. SQGs based on a weight of evidence from numerous studies were expected to be more useful nationwide than values based upon only limited amounts of data. SQGs were needed for a variety of different substances commonly measured in the NS&T Program; accordingly, guidelines were developed for as many chemicals as the data would warrant. SQGs were needed that would estimate the “safe” concentrations, i.e., concentrations below which effects were not likely. Also, guidelines were needed above which adverse effects were more likely. Therefore, two values were derived for each substance.

SQGs were derived initially using a database compiled from studies performed in both saltwater and freshwater and published in NOAA Technical Memorandum NOS OMA 52 (Long and Morgan 1990). A larger database compiled from many studies performed by numerous investigators in only saltwater was used to revise and update the SQGs (Long et al. 1995). Data from freshwater studies and/or of marginal quality used in 1990 were removed from the database in 1995, and a considerable amount of higher quality data were added to the database. Data from each study were arranged in order of ascending concentrations. Study endpoints in which adverse effects were reported were identified. From the ascending data tables, the 10th percentile and the 50th percentile (median) of the effects database were identified for each substance. The 10th percentile values were named the “Effects Range-Low” (ERL), indicative of concentrations below which adverse effects rarely occur. The 50th percentiles were named the “Effects Range-Median” (ERM) values, representative of concentrations above which effects frequently occur.

An example of the derivation method is shown in **Figure 1** in which the data for phenanthrene are arranged in ascending order. Green symbols indicate study endpoints in which no adverse effects were observed, such as in reference area samples. Red symbols indicate those study endpoints at which an adverse effect was observed. In the case of phenanthrene, there were 53 study endpoints indicating adverse effects. The 10th percentile of this data distribution was the 6th value, equivalent to 240 ppb phenanthrene. The 50th percentile was the 27th value, equivalent to 1500 ppb. As was apparent in the data for phenanthrene, the percentages of study endpoints indicating toxicity increased with increasing concentrations of most chemicals. The measures of reliability discussed below were calculated from the data available within the three concentration ranges defined by the ERL and ERM.

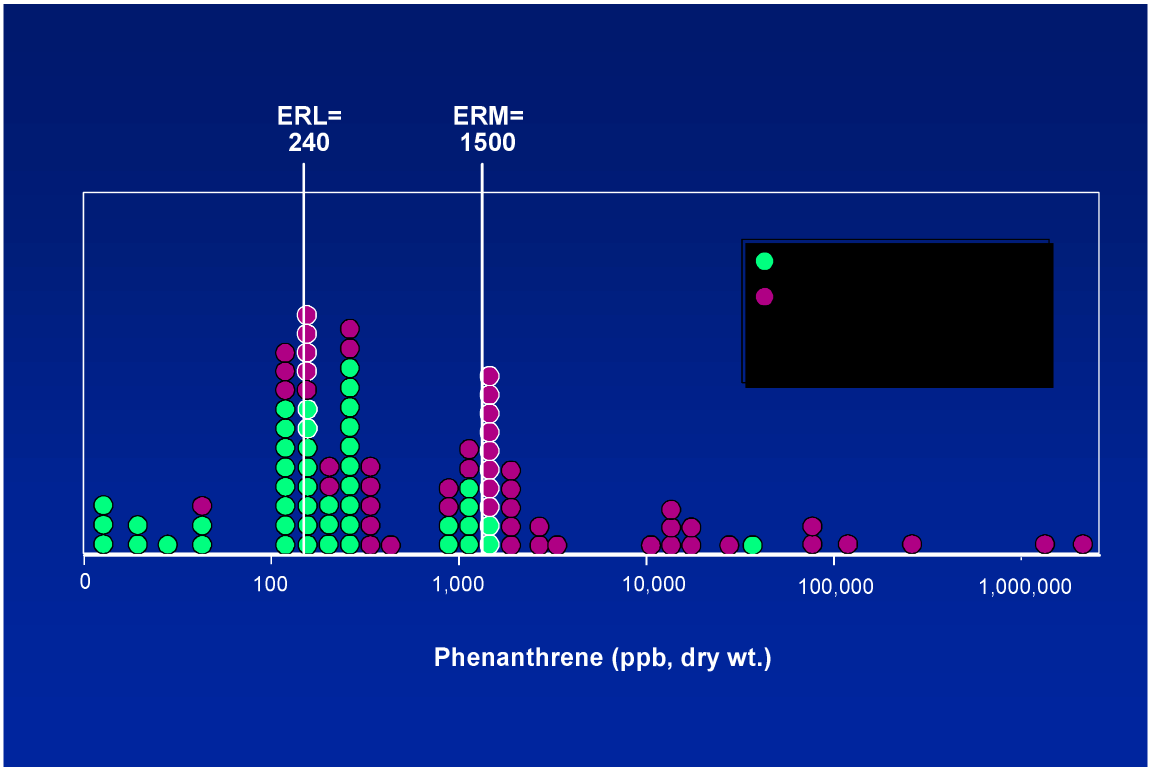

Figure 1. An example of the database used to derive the SQGs. Data for phenanthrene in which no adverse effects were observed are indicated by green symbols and those in which toxicity or some other measure of adverse effects were observed are indicated by red symbols. ERL= Effects Range-Low; ERM= Effects Range-Median.

The sediment quality guidelines

Based on the database assembled by Long et al. (1995), ERL and ERM values were calculated for 9 trace metals, 13 individual PAHs, 3 classes of PAHs, and 3 classes of chlorinated organic hydrocarbons **(Tables 1 and 2)**. There were insufficient amounts of reliable data available to perform similar calculations for other substances, including a few previously reported by Long and Morgan (1990).

The amount and quality of data used to derive the SQGs differed among the substances. Therefore, to provide a measure of the reliability of the SQGs, the percentages of study endpoints indicating adverse effects were calculated for the chemical ranges defined by the ERLs and ERMs **(Tables 1 and 2)**. Because the ERLs were intended to represent concentrations below which effects were rarely observed, low percentages of studies were expected to indicate effects within the ranges below the ERLs. Indeed, for all trace metals the percent of studies indicating adverse effects was less than 10% when concentrations were below the ERL values. For most organics, the incidence of effects was less than 25% when concentrations were below the ERLs.

Table 1. ERL and ERM guideline values for trace metals (ppm, dry wt.) and percent incidence of biological effects in concentration ranges defined by the two values (from Long et al., 1995). ERL= Effects Range-Low; ERM= Effects Range-Median.

	Guidelines		Percent incidence of effects*		
Chemical	ERL	ERM	<ERL	ERL - ERM	>ERM
Arsenic	8.2	70	5.0	11.1	63.0
Cadmium	1.2	9.6	6.6	36.6	65.7
Chromium	81	370	2.9	21.1	95.0
Copper	34	270	9.4	29.1	83.7
Lead	46.7	218	8.0	35.8	90.2
Mercury	0.15	0.71	8.3	23.5	42.3
Nickel	20.9	51.6	1.9	16.7	16.9
Silver	1.0	3.7	2.6	32.3	92.8
Zinc	150	410	6.1	47.0	69.8

*Number of data entries within each concentration range in which biological effects were observed divided by the total number of entries within each range.

The incidence of effects increased to 20% to 30% for most trace metals and 40% to 60% for most organics when concentrations exceeded ERL values but were lower than the ERM values. When concentrations exceeded the ERM values, the incidence of adverse effects increased to 60% to 90% for most trace metals and 80% to 100% for most organics. However, the reliabilities of the ERMs for nickel, mercury, DDE, total DDTs, and total PCBs were much lower than those for other substances. Therefore, the probabilities that the ERM values for these substances would accurately predict adverse effects are much lower than those for most chemicals.

Table 2. ERL and ERM guideline values for organic compounds (ppb, dry wt.) and percent incidence of biological effects in concentration ranges defined by the two values (from Long et al. 1995). ERL= Effects Range-Low; ERM= Effects Range-Median.

	Guidelines		Percent incidence of effects*		
Chemical	ERL	ERM	<ERL	ERL--ERM	>ERM
Acenaphthene	16	500	20.0	32.4	84.2
Acenaphthylene	44	640	14.3	17.9	100
Anthracene	85.3	1100	25.0	44.2	85.2
Fluorene	19	540	27.3	36.5	86.7
2-methyl naphthalene	70	670	12.5	73.3	100
Naphthalene	160	2100	16.0	41.0	88.9
Phenanthrene	240	1500	18.5	46.2	90.3
Sum LPAH	552	3160	13.0	48.1	100
Benz(a)anthracene	261	1600	21.1	43.8	92.6
Benzo(a)pyrene	430	1600	10.3	63.0	80.0
Chrysene	384	2800	19.0	45.0	88.5
Dibenzo (a,h) anthracene	63.4	260	11.5	54.5	66.7
Fluoranthene	600	5100	20.6	63.6	92.3
Pyrene	665	2600	17.2	53.1	87.5
Sum HPAH	1700	9600	10.5	40.0	81.2
Sum of total PAH	4022	44792	14.3	36.1	85.0
p,p'-DDE	2.2	27	5.0	50.0	50.0
Sum total DDTs	1.58	46.1	20.0	75.0	53.6
Total PCBs	22.7	180	18.5	40.8	51.0

*Number of data entries within each concentration range in which biological effects were observed divided by the total number of entries within each range.

Interpretation

Two guideline values were generated for each chemical: the ERL and the ERM. It is important to understand that these values were not derived as toxicity thresholds. That is, there is no assurance that there will be a total lack of toxicity when chemical concentrations are less than the ERL values. Similarly, there is no assurance that samples in which ERM values are exceeded will be toxic. Toxicity, or a lack thereof, must be confirmed with empirical data from toxicity tests.

The ERL values were not intended as concentrations that are always predictive of toxicity. Rather, they were intended and should be used primarily as estimates of the concentrations below which toxicity is least likely. As shown in Tables 1 and 2, the incidence of effects was usually higher when concentrations exceeded the ERLs than when concentrations were below the ERLs. However, the ERM values are better indicators of concentrations associated with effects than the ERLs.

Uses

The guidelines are commonly used in North America both to rank and prioritize sites of concern and chemicals of concern. That is, samples or study areas in which many chemicals exceed the ERM values and exceed them by a large degree may be considered as more contaminated than those in which none of the SQGs are exceeded. Samples in which ERL concentrations are exceeded, but no ERM values are exceeded, might be given intermediate ranks. Similarly, chemicals at concentrations well above the ERM values might be given a higher priority than those at concentrations below the ERLs. Chemicals at intermediate concentrations may qualify as a moderate priority. However, caution should be used when prioritizing sites or chemicals where only the concentrations of nickel, mercury, DDE, total DDTs, or total PCBs are elevated.

In studies performed by NOAA of toxicity and contamination of sediments in specific estuaries and bays, the SQGs have been used to rank and prioritize both sites and chemicals of potential concern. In these studies the chemical data were compared with the SQGs to identify spatial patterns in contamination, to estimate the spatial scales in contamination, and to rank sampling sites. The data also were compared with the SQGs to (1) identify which chemicals, if any, exceeded the ERL and ERM values, (2) tally the number of samples in which the SQGs were exceeded, (3) calculate the degrees to which the SQGs were exceeded, and (4) to identify which chemicals were most associated with measures of toxicity. For each regional assessment of bioeffects, the SQGs were used along with the results of toxicity tests to estimate the relative quality of sediments throughout the study area.

Field validation of predictive ability

To provide quantitative information on how well the SQGs correctly predict toxicity in actual field conditions, an analysis was conducted (Long et al. 1998a) with existing data compiled from many regional assessments conducted by NOAA and EPA. Matching chemistry and toxicity data from 1,068 samples from the Atlantic, Gulf of Mexico, and Pacific coasts were compiled into a database. Data were available from acute amphipod survival tests for all 1,068 samples; data from one or two additional tests in which sublethal responses were recorded were available for 437 samples. Several analyses were conducted with the data to investigate the predictive ability of the SQGs.

In the first analysis, the percentages of samples that were highly toxic were determined when individual ERM values were equaled or exceeded. That is, samples were identified in which the ERM value was equaled/exceeded for a particular substance. The percentages of those samples that were highly toxic in either the amphipod survival tests alone or in a battery of 2 to 4 tests (including those with amphipods) were then determined. Statistical analyses were used to classify samples as either non-toxic ($p>0.05$), marginally toxic ($p<0.05$), or highly toxic ($p<0.05$ and sample means exceed minimum significant differences) relative to controls in the laboratory tests. The predictive abilities of 28 sets of ERLs/ERMs were determined.

For most substances, 40% to 60% of samples in which chemical concentrations exceeded individual ERMs were highly toxic in the amphipod tests **(Figure 2)**.

For example, among the samples in which copper concentrations exceeded the ERM value (n=25), 52% were highly toxic in the amphipod survival tests. More than 75% of samples were highly toxic in which the ERMs for lead, 2-methylnaphthalene, and acenaphthylene were exceeded. For most substances, an increase in predictive ability of approximately 20% to 30% occurred when the data from the sublethal tests were included along with the amphipod data. Therefore, for most substances, 80% to 90% of samples were highly toxic in at least one of the tests performed when concentrations exceeded individual ERMs.

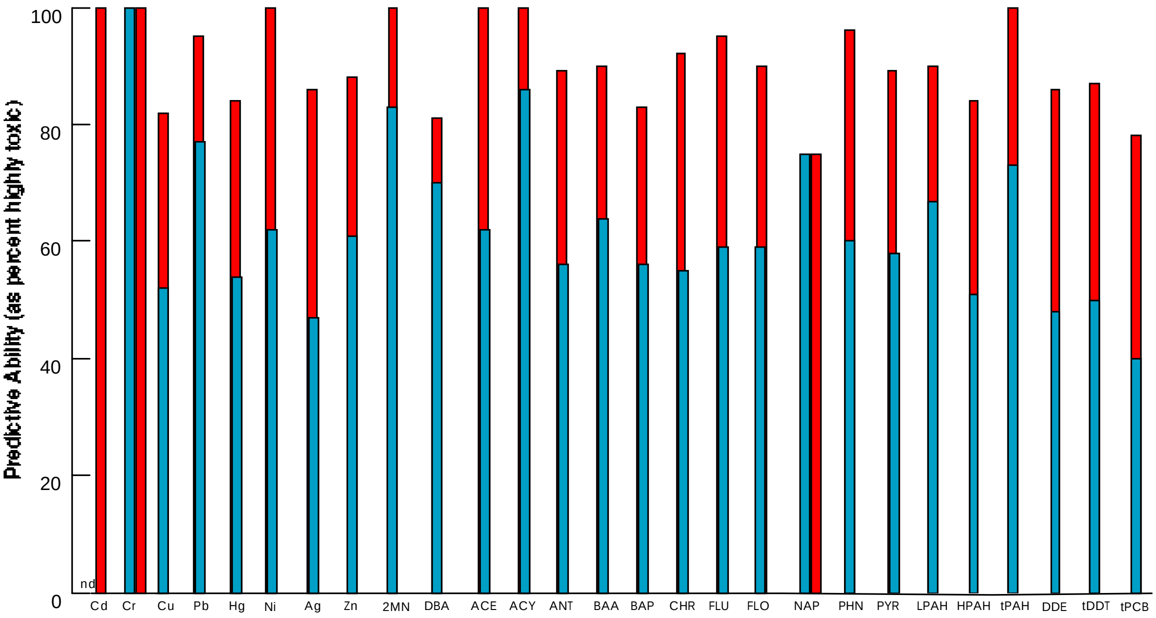

fig.2

Figure 2. Percent of sediment samples in field validation database in which highly significant toxicity was observed in amphipod survival tests alone (blue bars) and in any of 2 to 4 tests performed (red bars) when chemical concentrations equaled or exceeded individual ERM values (from Long et al., 1998a). ERL= Effects Range-Low; ERM= Effects Range-Median.

These data suggest that individual ERM values were reasonably predictive of toxicity. Given that the ERMs were derived as median values (not toxicity thresholds) in the effects database, predictive abilities of roughly 50% might be expected. Indeed, in the amphipod tests, 40% to 60% of samples were highly toxic when individual ERMs were exceeded. However, chemicals often occur in complex mixtures in environmental samples and toxicity in these tests could not be necessarily attributable to the substance which occurred at concentrations greater than the ERM values.

Therefore, a second series of analyses was conducted to estimate the effects of additivity of toxicants upon measures of toxicity. In these analyses the percentages of samples were calculated for several categories of samples: (1) no SQGs exceeded, (2) only ERLs exceeded, no ERMs equaled/exceeded, (3) increasing numbers of ERMs exceeded.

Table 3. Percentages of samples in which no significant toxicity, marginal toxicity, and highly significant toxicity was observed in amphipod survival tests (from Long et al., 1998a). ERL= Effects Range-Low; ERM= Effects Range-Median.

Chemical category	Number of samples	Percent not toxic	Percent marginally toxic	Percent highly toxic
no ERLs exceeded	329	68	21	11
1 or more ERLs exceeded	448	63	20	18
1 or more ERMs exceeded	291	48	13	39
1 to 5 ERMs exceeded	225	53	15	32
6 to 10 ERMs exceeded	46	37	11	52
11 to 20 ERMs exceeded	20	10	05	85

Only 11% of the 329 samples were highly toxic in the amphipod tests when none of the ERLs were exceeded **(Table 3)**. In this category, 21% of the samples were marginally toxic and 68% were not significantly toxic in this category. These data suggest that the ERLs were reasonably predictive of non-toxic conditions.

Given that the ERLs were calculated as the 10th percentiles of effects data, roughly equivalent predictive abilities (i.e., about 10%) were expected in this field validation study. The data, however, indicated that 18% of samples in which one or more ERLs (but, no ERMs) were exceeded were highly toxic. The incidence of toxicity increased with increases in the numbers of ERLs exceeded, peaking at 67% when 15 to 19 ERLs were exceeded (Long et al. 1998a; data not shown).

Given that the ERMs were derived as 50th percentile values in the effects databases, roughly equivalent predictive abilities (i.e., about 50%) were expected. There were 291 samples in which at least 1 ERM was exceeded by any amount **(Table 3)**. Among these samples, 13% were marginally toxic and 39% were highly toxic. As the numbers of chemicals exceeding the ERMs increased, there was an increase in the percentages of samples that were highly toxic, peaking at 85% when 11 to 20 ERMs were exceeded.

Mean ERM quotients

Chemicals often occur in saltwater sediments as complex mixtures. To provide a tool useful in assessing the potential toxicological significance of the presence of mixtures, mean ERM quotients were calculated for all 1068 samples used in the field validation study (Long et al. 1998a). These indices were derived as the average of the 25 quotients obtained by dividing the individual chemical concentrations by their respective ERM values. The percentages of samples that were not toxic, marginally toxic, and highly toxic were determined within ranges in the quotients. The data suggested a relatively consistent dose-response relationship: as the mean ERM quotients increased, the incidence of highly toxic responses increased (Long et al. 1998a). As more experience is gained with this tool, it may be useful in assessing the potential significance of chemical mixtures in sediment samples.

Probabilities of toxicity

The data from the study of predictive ability were compiled for both the sets of ERL and ERM values (from Long et al. 1995) and the comparable TEL (Threshold Effects Levels) and PEL (Probable Effects Levels) values from MacDonald et al. (1996) to provide a synopsis of the likelihood of significant toxicity in amphipod survival tests (Long and MacDonald 1998). This is an attempt to estimate the likelihood that samples with certain chemical characteristics would be toxic.

Table 4 lists the chemical characteristics that equate to different probabilities of amphipod toxicity based on the data from Long et al. (1998a). Data used to derive **Table 4** were compiled from Long et al. (1998a) in which there were 1086 samples and merged with more recent data from Biscayne Bay (FL) (n=226) and Pearl Harbor (HI) n=219), giving a total data set of 1513 samples. These samples were collected in various studies performed on the Atlantic, Pacific and Gulf of Mexico coasts.

The percent incidence of highly toxic responses and the average survival of the amphipods in all samples within each cateogry are shown in **Table 4**. Four chemical indices calibrated to the SQGs are shown for each of four categories. In category 1, sediments least likely to be toxic were actually toxic in only 8-9% of the samples. Average amphipod survival in these samples was 92-93%, indicating that survival, on average, was not decreased appreciably from what would be expected in clean reference sediments. As the numbers of SQGs exceeded increases and as the mean SQG quotients increase, the incidence of toxicity increases and the average survival rate decreases.

Samples with chemical characteristics equivalent to Category 2 have the most uncertainty as to toxicity. Average survival approximates the critical threshold of 80% of controls, whereas in the other categories, average survival is clearly greater than or less than 80%. In category 3, about 50% of samples were toxic and average survival was about 60-70%. In category 4, about 73-83% of samples were toxic and average survival dropped to about 40%, indicating high probabilities of toxic conditions.

These data may be useful in determining the need for additional testing and analyses of sediments. For example, the probability of incorrectly classifying a site as non-toxic when all chemical concentrations are below all SQGs and either of the mean SQG quotients is less than 0.1 is about 10% and the probability of a site being toxic is about 75% or greater when chemical data match the characteristics of Category 4 conditions in **Table 4**. However, in sediments classified as Category 2, toxicity or the lack thereof is more uncertain.

Table 4. Percent incidence of highly toxic samples and average percent amphipod survival in marine sediment samples classified according to numerical sediment quality guidelines.

	Percent highly toxic* samples		Average, control-adjusted amphipod survival	
Chemical characteristics relative to sediment guidelines	National** database (n=1068)	Combined summary (n=1513)	National** database (n=1068)	Combined summary (n=1513)
Category 1:				
• mean ERM quotients <0.1	11	9	93	93
• mean PEL quotients <0.1	10	8	93	93
• no ERLs exceeded	11	9	92	92
• no TELs exceeded	9	8	92	92
Category 2:				
• mean ERM quotients 0.11 - 0.5	30	21	81	86
• mean PEL quotients 0.11 - 1.5	25	21	84	86
• 1-5 ERMs exceeded	32	32	79	79
• 1-5 PELs exceeded	24	18	83	88
Category 3:				
• mean ERM quotients 0.51-1.5	46	49	74	70
• mean PEL quotients 1.51 - 2.3	50	49	66	68
• 6-10 ERMs exceeded	52	57	63	59
• 6-20 PELs exceeded	47	48	71	70
Category 4:				
• mean ERM quotients >1.5	75	76	43	41
• mean PEL quotients >2.3	77	73	47	46
• >10 ERMs exceeded	85	80	41	41
• >20 PELs exceeded	88	83	38	37

* mean survival significantly different from controls and <80% of controls

** data from Long et al., 1998

The ERLs and mean ERM quotients for saltwater were more efficient at correctly predicting non-toxicity (100% and 93% correct, respectively) than SEM:AVS ratios (80% correct) based on analyses of data compiled to field-validate the SEM:AVS criteria (Long et al., 1998b). Also, the ERMs and mean ERM quotients were slightly more predictive of toxic conditions (33% and 42% correct, respectively) than the SEM:AVS ratios (26% correct). These data suggest that the predictive abilities of SQGs based on bulk trace metals data are not improved with SEM-to-AVS normalizations (Long et al., 1998b).

Limitations

The SQGs should be used with caution and common sense. There are no SQGs available for many substances that can be highly toxic in sediments. The abilities of the SQGs to correctly predict toxicity of co-varying substances for which there are no SQGs are unknown. The SQGs were derived in units of dry weight sediments; therefore, they do not account for the potential effects of geochemical factors in sediments that may influence contaminant bioavailability. The SQGs were not intended for use in predicting effects in wildlife or humans through bioaccumulation pathways. The SQGs were neither calculated nor intended as toxicological thresholds; therefore, there is no certainty that they will always correctly predict either non-toxicity or toxicity. The SQGs were derived with data from soft sedimentary deposits; they should not be applied to assessments of upland soils, gravel, coarse sand, tar, slag, or metal ore.

The SQGs are best applied when accompanied by measures of effects such as laboratory toxicity tests and/or benthic community analyses and/or bioaccumulation tests, which lead to the preparation of a weight of evidence. Furthermore, they are best applied in a comprehensive assessment framework involving the establishment of clear study objectives, *a priori* methods for data analyses, and well-understood decision points regarding the uses of the data.

References

Long, E. R., and D. D. MacDonald. 1998. Recommended uses of empirically derived, sediment quality guidelines for marine and estuarine ecosystems. *Human and Ecological Risk Assessment* 4(5): 1019-1039.

Long, E.R., and L. G. Morgan. 1990. The potential for biological effects of sediment-sorbed contaminants tested in the National Status and Trends Program. NOAA Technical Memorandum NOS OMA 52. National Oceanic and Atmospheric Administration. Seattle, Washington.

Long, E. R., and C. J. Wilson. 1997. On the identification of toxic hot spots using measures of the sediment quality triad. *Marine Pollution Bulletin* 34 (6): 373-374.

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(1): 81-97.

Long, E. R., L. J. Field, and D. D. MacDonald. 1998a. Predicting toxicity in marine sediments with numerical sediment quality guidelines. *Environmental Toxicology and Chemistry* 17(4)

Long, E. R., D. D. MacDonald, J. Cubbage, C. G. Ingersoll. 1998b. Predicting the toxicity of sediment-associated trace metals with simultaneously-extracted tracemetal:acid-volatile sulfide concentrations and dry weight-normalized concentrations: A critical comparison. *Environmental Toxicology and Chemistry* 17(5)

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