J. Great Lakes Res. 22(3):602-623
Internat. Assoc. Great Lakes Res., 1996

Calculation and Evaluation of Sediment Effect Concentrations for the Amphipod *Hyalella azteca* and the Midge *Chironomus riparius*

Christopher G. Ingersoll,[1] Pamela S. Haverland,[1] Eric L. Brunson,[1] Timothy J. Canfield,[1] F. James Dwyer,[1] Christopher E. Henke,[1] Nile E. Kemble,[1] David R. Mount,[2] and Richard G. Fox[3]

[1]*Midwest Science Center*
National Biological Service
4200 New Haven Road
Columbia, Missouri 65201

[2]*Mid-continent Ecological Division-Duluth*
U.S. Environmental Protection Agency
6201 Congdon Boulevard
Duluth, Minnesota 55804

[3]*Great Lakes National Program Office*
U.S. Environmental Protection Agency
77 W. Jackson
Chicago, Illinois 60604

ABSTRACT. Procedures are described for calculating and evaluating sediment effect concentrations (SECs) using laboratory data on the toxicity of contaminants associated with field-collected sediment to the amphipod Hyalella azteca *and the midge* Chironomus riparius. *SECs are defined as the concentrations of individual contaminants in sediment below which toxicity is rarely observed and above which toxicity is frequently observed. The objective of the present study was to develop SECs to classify toxicity data for Great Lake sediment samples tested with* Hyalella azteca *and* Chironomus riparius. *This SEC database included samples from additional sites across the United States in order to make the database as robust as possible. Three types of SECs were calculated from these data: (1) Effect Range Low (ERL) and Effect Range Median (ERM), (2) Threshold Effect Level (TEL) and Probable Effect Level (PEL), and (3) No Effect Concentration (NEC). We were able to calculate SECs primarily for total metals, simultaneously extracted metals, polychlorinated biphenyls (PCBs), and polycyclic aromatic hydrocarbons (PAHs). The ranges of concentrations in sediment were too narrow in our database to adequately evaluate SECs for butyltins, methyl mercury, polychlorinated dioxins and furans, or chlorinated pesticides. About 60 to 80% of the sediment samples in the database are correctly classified as toxic or not toxic depending on type of SEC evaluated. ERMs and ERLs are generally as reliable as paired PELs and TELs at classifying both toxic and non-toxic samples in our database. Reliability of the SECs in terms of correctly classifying sediment samples is similar between ERMs and NECs; however, ERMs minimize Type II error (false negatives) relative to ERLs and minimize Type I error (false positives) relative to NECs. Correct classification of samples can be improved by using only the most reliable individual SECs for chemicals (i.e., those with a higher percentage of correct classification). SECs calculated using sediment concentrations normalized to total organic carbon (TOC) concentrations did not improve the reliability compared to SECs calculated using dry-weight concentrations. The range of TOC concentrations in our database was relatively narrow compared to the ranges of contaminant concentrations. Therefore, normalizing dry-weight concentrations to a relatively narrow range of TOC concentrations had little influence on relative concentrations of contaminants among samples. When SECs are used to conduct a preliminary screening to predict the potential for toxicity in the absence of actual toxicity testing, a low number of SEC exceedances should be used to minimize the potential for false negatives; however, the risk of accepting higher false positives is increased.*

INDEX WORDS: Toxicity, sediment, Great Lakes, thresholds, amphipods, midges, chironomids, Hyalella.

INTRODUCTION

Over the past decade, a variety studies have reported toxicity associated with field-collected sediments (USEPA 1994, ASTM 1995, Burton *et al.* 1996). However, it is often difficult to determine relationships between levels of contamination and toxicity in these and other studies because the sediments typically contain a variety of both organic and inorganic contaminants. Sediment Effect Concentrations (SECs) have been used to determine concentrations of individual contaminants in sediment below which toxicity is rarely observed and above which toxicity is frequently observed (Long *et al.* 1995, MacDonald *et al.* 1996). However, only a limited number of SECs for freshwater sediments have been published (Persaud *et al.* 1992, Batts and Cubbage 1995). The objective of the present study was to develop SECs to classify toxicity data for Great Lakes sediment samples tested with *Hyalella azteca* and *Chironomus riparius*. The SEC database included samples from additional sites across the United States in order to make the database as robust as possible.

Ideally, SECs could be used to: (1) interpret historical sediment chemistry data, (2) identify chemicals or areas of concern, (3) identify the need for more detailed studies before an action is taken, (4) identify a potential problem before discharging a chemical, (5) establish a link between a contaminant source and sediment quality, (6) trigger regulatory action, or (7) establish target remediation objectives. The strength of SECs generated using data from studies with individual chemicals spiked into sediment or with an equilibrium partitioning (EQP) approach is that cause and effect relationships can be established (Di Toro *et al.* 1991, USEPA 1992). While, all seven of the uses listed above for SECs could be satisfied with either of these approaches, both the spiked-sediment and EQP approaches were developed primarily for evaluating the effects of individual chemicals. However, contaminated sediments typically contain complex mixtures of chemicals which could act independently, additively, synergistically, or antagonistically. Therefore, the application of SECs developed using these two approaches is often uncertain in field-collected sediments (Swartz and DiToro 1997).

One of the main strengths of SECs generated using data from tests conducted with field-collected samples is that the potential effects of mixtures of chemicals are explicitly addressed (Long and Morgan 1991, USEPA 1992, MacDonald 1994, Long *et al.* 1995, MacDonald *et al.* 1996). However, there are a number of limitations associated with the co-occurrence-based approaches that have been used to generate SECs including cause and effect is difficult to establish and use of these SECs may be restricted to the geographical area where the sediments were collected. Hence, the last four uses for SECs listed above are difficult to accommodate using co-occurrence-based approaches such as a weight-of-evidence approach (Long and Morgan 1991). For example, these SECs should not be used independently to establish trigger levels for clean up of sediment. One of the major strengths of SECs developed with data on field-collected sediments is in their use for predicting the potential for toxicity in field-collected sediment samples. A primary use of SECs developed with field-collected sediments should be to provide guidance for determining sites which may require further investigation (Long and Morgan 1991, MacDonald 1994). Moreover, the ability of any sediment toxicity test or SEC to predict benthic community effects should be considered before any approach is used to routinely evaluate sediment quality (Canfield *et al.* 1994, 1996a, 1996b).

As part of the Assessment and Remediation of Contaminated Sediment (ARCS) program (Ross *et al.* 1992, Fox and Tuchman 1996), whole-sediment toxicity tests were conducted with the amphipod *Hyalella azteca* (14- and 28-d tests) and the midges *Chironomus riparius* (14-d test) and *Chironomus tentans* (10-d test) with sediments collected from three Great Lakes Areas of Concern: Indiana Harbor, IN; Buffalo River, NY; and Saginaw River, MI (USEPA 1993, Burton *et al.* 1996). Only a limited number of samples were successfully tested with *C. tentans*; therefore, we did not use these data to calculate SECs. Sediment chemistry, benthic community analysis (Canfield *et al.* 1996a), elutriate toxicity (Hall *et al.* 1996), mutagenicity (Papoulias and Buckler 1996, Papoulias *et al.* 1996), and toxicity ranking (Canfield *et al.* 1996a, Wildhaber and Schmitt 1996) of sediment samples were also evaluated as part of the ARCS program (USEPA 1993). In addition to the ARCS data, we evaluated toxicity and chemistry data generated with sediments collected from the following sites: (1) Waukegan Harbor, IL (Ingersoll and Nelson 1990); (2) the upper Mississippi River, MN (USEPA 1996); (3) the upper Clark Fork River, MT (Kemble *et al.* 1994); (4) the Trinity River, TX (USEPA 1996); (5) Mobile Bay, AL (USEPA 1996), and (6) Galveston Bay, TX (Roach *et al.* 1993).

MAR 16 1999
9703

We used three different approaches to calculate various SECs including: (1) Effect Ranges Low and Median (ERL and ERM; Long and Morgan 1991), (2) Threshold Effect and Probable Effect Levels (TEL and PEL; MacDonald 1994, Smith *et al.* 1996), and (3) No Effect Concentrations (NEC; Kemble *et al.* 1994). These approaches for calculating SECs were evaluated relative to their: (1) reliability in terms of correctly classifying the toxicity of sediment samples within the data set, (2) predictive ability for correctly classifying the toxicity of sediment samples from independent data sets, and (3) comparability within the data set or to other published SECs such as Apparent Effect Threshold (AET), ERM, PEL, or EQP values. Procedures are also described in companion papers for confirming the response of test organisms in the laboratory with the response of benthic communities in the field (e.g., Sediment Quality Triad; Canfield *et al.* 1996a) and for confirming the cause of sediment toxicity (e.g., sediment spiking and Toxicity Identification Evaluations, TIE; Ankley and Thomas 1992, Ankley *et al.* 1996). Smith *et al.* (1996) also used our data and additional data from North America to calculate and evaluate TELs and PELs for freshwater sediments.

This paper evaluates SECs calculated using dry-weight concentrations for the entire database because these SECs were generally more reliable than SECs calculated using sediment concentrations normalized to total organic carbon (TOC) concentrations for non-ionic organics or SECs calculated using pore-water metals concentrations. We have included one example in the paper comparing the reliability ERMs calculated using dry-weight concentrations vs sediment concentrations normalized to TOC concentrations for polycyclic aromatic hydrocarbons (PAHs) and polychlorinated biphenyls (PCBs). The reader is encouraged to consult USEPA (1996) for these and additional applications of the database including: (1) Great Lakes vs the entire database, (2) dry-weight vs TOC-normalized concentrations for PAHs and total PCBs, (3) whole-sediment vs pore-water metal concentrations, and (4) total-metal vs simultaneously extracted metal (SEM) concentrations. A copy of the database is available in USEPA (1996) or from the authors if the reader is interested in pursuing additional applications.

METHODS

The following steps were taken to develop the database: (1) chemistry and toxicity data were generated, (2) samples were classified as "toxic" or "not toxic" based on statistical analyses of the toxicity tests, (3) toxic samples were classified as "effects" or "no concordance" based on chemistry, (4) minimum data requirements were established, and (5) SECs were calculated. SECs were calculated for the following tests: (1) 14-d *C. riparius* [CR14], (2) 14-d *H. azteca* [HA14], and (3) 28-d *H. azteca* [HA28]. Data from the following tests were combined: (1) 10-d with 14-d *H. azteca*, (2) 29- to 32-d with 28-d *H. azteca*, and (3) 13-d with 14-d *C. riparius* (Table 1). Chemical concentrations used to calculate SECs were normalized to: (1) dry weight, (2) total organic carbon (for non-ionic organics; Di Toro *et al.* 1991), or (3) AVS (for divalent metals; Di Toro *et al.* 1990). Calculations and graphics were performed using SAS version 6.08 (SAS 1992).

TABLE 1. Percentage of whole-sediment samples identified as toxic in Chironomus riparius *14-d (CR14),* Hyalella azteca *14-d (HA14), or* H. azteca *28-d (HA28) tests.*

	Toxic[1]	Toxic-S[2]	Toxic-G[3]	Toxic-M[4]
All samples				
CR14	26 (42)	24 (42)	9 (34)	ND
HA14	41 (32)	25 (32)	20 (25)	24 (21)
HA28	39 (62)	26 (62)	34 (44)	11 (36)
Great Lakes				
CR14	37 (27)	33 (27)	11 (19)	ND
HA14	48 (27)	30 (27)	20 (25)	24 (21)
HA28	48 (27)	41 (27)	24 (25)	5 (21)
Upper Mississippi River				
HA28	0 (5)	0 (5)	ND	ND
Clark Fork River				
CR14	7 (15)	7 (15)	7 (15)	ND
HA28	53 (15)	13 (15)	53 (15)	20 (15)
Trinity River				
HA14	0 (5)	0 (5)	ND	ND
HA28	0 (5)	0 (5)	ND	ND
Mobile Bay				
HA28	0 (5)	0 (5)	ND	ND
Galveston Bay				
HA28	60 (5)	60 (4)	25 (4)	ND

[1]Toxic: Significant reduction in survival, growth, or maturation relative to the control ($p < 0.05$, N in parentheses).
[2]Toxic-S: Significant reduction in survival.
[3]Toxic-G: Significant reduction in growth.
[4]Toxic-M: Significant reduction in maturation (*H. azteca* only).
[5]ND: Not determined.

Toxicity Testing

Toxicity tests with the amphipod *Hyalella azteca* were conducted for 10 to 32 d following procedures outlined in Ingersoll and Nelson (1990), USEPA (1994), and ASTM (1995). Tests were generally started within 3 weeks of sediment collection. The control sediment was a fine silt- and clay-particle size soil obtained from an agricultural area. Twenty amphipods were exposed in 200 mL of sediment with 800 mL of overlying water in 1-L beakers. Four replicate beakers were tested at 20°C on a 16L:8D photoperiod at a light intensity of about 50 to 100 foot candles. Overlying water was renewed daily and the amphipods were fed a suspension of Purina® Rabbit Chow three times a week. Endpoints measured at the end of the amphipod tests were survival, growth (as length), or sexual maturation. Toxicity tests with the midge *Chironomus riparius* were conducted for 13 to 14 d using similar procedures to those used in the tests with amphipods except midges were < 48-h old at the start of the tests and midges were fed a mixture of algae, Cerophyl®, and Hartz® Dog Treats daily. Endpoints measured at the end of the midge tests were survival and growth (as length). A sample was designated as "toxic" if there was a statistically significant reduction in survival, growth, or maturation relative to the response in the control sediment (USEPA 1996).

Physical characterizations of sediments included organic carbon content, water content, and particle size. Chemical characterizations of sediments included total metals (Ag, As, Cd, Cr, Cu, Fe, Hg, Mn, Ni, Pb, Se, Zn), organometals (butyltins and methyl mercury), acid volatile sulfide (AVS) and simultaneously extracted metals (SEM), chlorinated pesticides, total polychlorinated biphenyls (PCBs), polychlorinated dioxins and furans, or polycyclic aromatic hydrocarbons (PAHs). Metal concentrations in pore water were also measured in selected samples (USEPA 1996).

Sediment toxicity data were obtained for the sites listed below (Table 1). Suspected toxicants were mixtures of metals and organic compounds except for the samples from the upper Clark Fork River where suspected toxicants were primarily As, Cd, Cu, Pb, and Zn.

Great Lakes ARCS

Sediments were collected from three Great Lakes Areas of Concern: Indiana Harbor (5 samples tested in August 1989), Buffalo River (6 samples tested in October 1989), and Saginaw River (4 samples tested during Survey 1 in December 1989 and 8 samples tested during Survey 3 in June 1990). Toxicity tests were conducted with *H. azteca* (14- and 28-d tests) and *C. riparius* (14-d test). All of the Indiana Harbor samples were extremely toxic to amphipods in the 14-d test. Therefore, a 28-d test with amphipods was not conducted. We assumed for the calculation of SECs that Indiana Harbor samples toxic to amphipods in a 14-d test would also be toxic in a 28-d test.

Waukegan Harbor

Toxicity tests were conducted with 4 sediment samples from Waukegan Harbor, IL in November 1987 (*H. azteca* 10- and 29-d tests, *C. riparius* 13-d test).

Upper Mississippi River

Toxicity tests were conducted with 5 sediment samples from the upper Mississippi River near Minneapolis, MN in September 1987 (*H. azteca* 32-d test).

Upper Clark Fork River

Sediment samples were collected from Milltown Reservoir (8 samples tested in July 1991) and the Clark Fork River, MT (7 samples tested in September 1991). Toxicity tests were conducted with *H. azteca* (28-d test) and *C. riparius* (14-d test).

Trinity River

Toxicity tests were conducted with 5 sediment samples from the Trinity River, near Dallas, TX in June 1988 (*H. azteca* 10- and 32-d tests).

Mobile Bay

Toxicity tests were conducted with 6 sediment samples from Mobile Bay, AL in March 1988 (*H. azteca* 28-d test). The test was conducted under static conditions with 10‰ salinity in the overlying water.

Galveston Bay

Toxicity tests were conducted with 5 sediment samples from Galveston Bay, TX in July 1990 (*H*

azteca 28-d test). The test was conducted under static conditions with 10‰ salinity in the overlying water.

Classification of Effects and Minimum Data Requirements for Use of an SEC

To increase the likelihood that associations between sediment chemistry and toxicity would be observed, the data were screened to determine if at least a 10-fold difference in concentration for at least one chemical among the samples was met from each site (Long and Morgan 1991, MacDonald 1994). The chemicals measured in each sample were classified in terms of their association with the observed toxicity. Each of the chemicals in the toxic samples were classified as an "effect" or "no concordance" depending on whether the ratio of the concentration in the sample to the mean concentration in the non-toxic samples was > 1 or ≤ 1. Concentrations of chemicals in non-toxic samples were designated as "no effects." Samples designated with the no concordance descriptor were also included with the no-effect samples for calculation of SECs. Long and Morgan (1991), MacDonald (1994), Long *et al.* (1995), MacDonald *et al.* (1996), and Smith *et al.* (1996) used a similar designation; however, they considered a chemical to be associated with a toxic effect only if the mean concentration in toxic samples at a site was at least two fold greater than the mean concentration in non-toxic samples at a site. We chose to use a ratio of > 1 instead of > 2 to classify a sample as an "effect" in order to minimize Type II error (toxic sample classified as not toxic). We used an SEC for a chemical only if: (1) five or more of the samples were toxic for the chemical and (2) the number of toxic samples with concentrations above the SEC was greater than the number of toxic samples with concentrations below the SEC.

Calculation of Effect Range Low (ERL) and Effect Range Median (ERM)

We calculated ERLs and ERMs using procedures described by Long and Morgan (1991) and Long *et al.* (1995). Our ERLs and ERMs are calculated for individual toxicity tests (e.g., the *H. azteca* 28-d test) in order use consistent endpoints for determining a toxic response. In contrast, Long and Morgan (1991) merged data from about 75 sources. These sources included marine and freshwater field surveys, spiked-sediment tests, and EQP. Effect ranges were calculated by Long *et al.* (1995) for 9 metals, 13 individual PAHs, 3 groups of PAHs, and 3 synthetic organic contaminants. Strengths of the Long *et al.* (1995) approach include: (1) ranges instead of absolutes (e.g., AET) are calculated, (2) a preponderance of evidence from diverse sources is used to generate the ranges (e.g., weight of evidence), and (3) probability of observing effects can be estimated. Limitations to the Long *et al.* (1995) approach include: (1) the quality of the data was variable, (2) different types of data were merged (e.g., acute lethality and benthic community structure were combined to calculate effect ranges), (3) concentrations were calculated on a dry-weight basis (data on sediment organic carbon and AVS concentrations were not available for all data sets), and (4) no-effect data are not used in the calculation of ERLs or ERMs.

Long *et al.* (1995) calculated ERLs and ERMs using the following procedure. Concentrations observed or predicted by different methods to be associated with effects were sorted in ascending order, and the lower 10 percentile (ERL) and 50 percentile (ERM) effect concentrations were calculated. An ERL was defined by Long and Morgan (1991) and Long *et al.* (1995) as the concentration of a chemical in sediment below which adverse effects were rarely observed or predicted among sensitive species. An ERM was defined as the concentration of a chemical in sediment above which effects are frequently or always observed or predicted among most species. Use of percentiles minimized the influence of single data points (e.g., potential outliers associated with AETs) on SECs. No-effect data were used to evaluate the reliability of ERLs and ERMs calculated using only effect data (Long *et al.* 1995). We chose to calculate ERLs using the 15 percentile rather than using the 10 percentile of effects to reduce the potential for Type II error (false negatives; MacDonald *et al.* (1996; see below)).

Calculation of Threshold Effect Level (TEL) and Probable Effect Level (PEL)

We calculated TELs and PELs using procedures described by MacDonald (1994) and MacDonald *et al.* (1996). Our TELs and PELs are calculated for individual toxicity tests in order use consistent endpoints for determining a toxic response. MacDonald (1994) and MacDonald *et al.* (1996) calculated TELs and PELs by expanding two to three fold the database originally developed by Long and Morgan (1991) and by excluding freshwater data. Effect ranges were calculated by MacDonald *et al.* (1996) for 9 metals, 7 pesticides, 13 individual PAHs, 3 groups of PAHs, total PCBs, and one phthalate ester. A similar procedure was used to calculate freshwater TELs and PELs for 8 metals, 6 individual PAHs, total PCBs, and 8 pesticides (Smith *et al.* 1996). Strengths and limitations to this approach are similar to the ERL/ERM approach. However, calculation of TELs and PELs take both effect and no-effect data into consideration.

MacDonald *et al.* (1996) and Smith *et al.* (1996) calculated TELs and PELs using the following procedure. Concentrations observed or predicted by different methods to be associated with effects were sorted and the lower 15 percentile (ERL) and 50 percentile (ERM) concentrations of the effects data set were calculated. In addition, the 50 percentile (No Effect Range Median; NERM) and 85 percentile (No Effect Range High; NERH) concentrations of the no-effects data set were calculated. The TEL was calculated as the geometric mean of the ERL and NERM, whereas the PEL was calculated the geometric mean of the ERM and NERH. The geometric mean was used rather than the arithmetic mean because the two data sets are typically not normally distributed. An analogous procedure has been used to calculate Maximum Acceptable Toxicant Concentrations (MATCs) from the geometric mean of the no-observable- and low-observable-effect concentrations (LOEC and NOEC; MacDonald 1994). For each of the values (ERL, ERM, NERM, and NERH), a series of percentiles was evaluated to optimize correct classification of toxicity using the TELs and PELs (MacDonald 1994). The approaches described by Long *et al.* (1995) and MacDonald *et al.* (1996) have been used by NOAA (Long and Morgan 1991), Environment Canada (CCME 1995), and the state of Florida (MacDonald 1994) to derive sediment quality guidelines. Additional organizations that are considering the use of these approaches to derive sediment quality guidelines include the state of California (Lorenzato *et al.* 1991), the International Council for Exploring the Sea, and the National Rivers Authority in the United Kingdom (R. Fleming, WRc, Marlow, Bucks, United Kingdom, personal communication).

Calculation of No Effect Concentration (NEC)

We also calculated No Effect Concentrations (NECs) which are analogous to Apparent Effect Thresholds (AETs). An AET is defined as the sediment concentration of a given chemical above which statistically significant effects (e.g., sediment toxicity) are always observed (Barrick *et al.* 1988). If any chemical exceeds its AET for a particular response, an adverse effect is expected for that response. If all concentrations of chemicals are below their AET for a particular response, then no adverse effect is expected. The AET approach has been applied to contaminated sediment in marine environments (e.g., in the Puget Sound, Barrick *et al.* 1988, and in California, Becker *et al.* 1989); however, the AET approach has rarely been used to evaluate freshwater sediments (Kemble *et al.* 1994).

A NEC is calculated as the maximum concentration of a chemical in a sediment that did not significantly adversely affect the particular response (e.g., survival, growth, or maturation) compared to the control. We chose to use the term NEC instead of AET because: (1) we calculated NECs for whole-sediment or pore-water concentrations, while AETs are typically calculated for just whole-sediment concentrations; (2) a minimum of 25 to 50 samples is recommended for calculating an AET and we used < 25 samples to calculate some of our NECs; and (3) we calculated effects relative to a control sediment, whereas AETs are typically calculated relative to reference sediments.

Evaluations of SECs

SECs were evaluated relative to their potential to: (1) correctly classify toxic samples as toxic (toxic sample that exceeds an SEC [hit]); (2) correctly classify non-toxic samples as not toxic (non-toxic sample that does not exceed an SEC [no hit]); (3) incorrectly classify non-toxic samples as toxic (Type I error; false positive; non-toxic sample that exceeds an SEC [hit]); and (4) incorrectly classify toxic samples as not toxic (Type II error; false negative; toxic sample that does not exceed an SEC [no hit]). The SECs were evaluated relative to their: (1) reliability in terms of correctly classifying the toxicity of sediment samples within the data set, (2) predictive ability for correctly classifying the toxicity of sediment samples from independent data sets, and (3) comparability within the data set or to other published SECs.

RESULTS AND DISCUSSION

Toxicity of Sediment Samples

The percentage of sediment samples identified as toxic are listed in Table 1. Survival or growth of *C. riparius* in 14-d tests (CR14) were significantly re-

duced in 26% of the 42 samples tested. In the CR14 tests when both survival and growth were measured, survival (24%) was reduced more frequently than growth (9%). Sediments from the Clark Fork River (7%) were less toxic in the CR14 test than sediments from the Great Lakes (37%). Survival, growth, or maturation of *H. azteca* in 14-d tests (HA14) were reduced in 41% of the 32 samples tested. In the HA14 test, survival (25%), growth (20%), and maturation (24%) were reduced in a similar percentage of samples. None of the Trinity River samples were toxic; however, 48% of the Great Lakes samples were toxic in the HA14 test. Survival, growth, or maturation of *H. azteca* in the 28-d tests (HA28) were reduced in 39% of the 62 samples tested. In the HA28 test, survival (26%) and growth (34%) were reduced in a higher percentage of samples compared to maturation (11%). None of the upper Mississippi River, Trinity River, or Mobile Bay samples were toxic; however, 48 to 60% of the Great Lakes, Clark Fork River, or Galveston Bay samples were toxic in the HA28 test.

In summary, both survival and growth endpoints provided unique information for assessing sediment toxicity and should be measured in either HA14 or HA28 tests. The HA28 test seldom identified toxic samples that were not identified as toxic in the HA14 test (USEPA 1996). However, the majority of the samples used to make these comparisons were highly contaminated. We have not compared responses in HA14 vs HA28 tests using moderately contaminated samples. Additional exposures conducted with moderately contaminated sediment may exhibit a higher percentage of sublethal effects in the HA28 test compared to HA14 test (Kemble *et al.* 1994). Using CR14 test, only one additional sample was identified as toxic compared to responses in the HA14 or HA28 tests (USEPA 1996). A primary consideration in selecting an organism for toxicity testing should be its ability to identify toxic samples (Burton *et al.* 1996). In future evaluations, it may be a more efficient use of resources to test additional samples with *H. azteca* alone rather than testing fewer sediments using both *H. azteca* and *C. riparius*.

Calculation and Evaluation of SECs

Due to space limitations in the journal, only the HA28 SECs calculated on a dry-weight basis are listed in Table 2. USEPA (1996) lists SECs for HA14 and CR14 tests. We were able to calculate SECs primarily for total metals, simultaneously extracted metals (SEM metals), total PCBs (for the HA28 test), and PAHs (Table 2; USEPA 1996). The ranges of concentrations in the samples were too narrow or there were too few measured concentrations in our database to adequately evaluate SECs for butyltins, methyl mercury, polychlorinated dioxins and furans, PCBs (for the CR14 or HA14 tests), or chlorinated pesticides. Either less than five samples were toxic for these chemicals or the number of toxic samples with concentrations above the SEC was less than the number of toxic samples with concentrations below the SEC.

For a particular chemical, ERMs were typically higher than paired PELs and ERLs were typically higher than paired TELs (Table 2; USEPA 1996). This resulted from a lower distribution of concentrations in non-toxic samples (e.g., low NERM or NERH). Although the concentrations of these paired SECs differed, the percentage of samples correctly classified by paired SECs was similar for all three toxicity tests (USEPA 1996). These analyses indicated ERMs and ERLs were generally as reliable as paired PELs and TELs at classifying both toxic and non-toxic samples in our database. Therefore, the remainder of the paper describes results of evaluations using ERLs and ERMs instead of TELs or PELs.

The SECs for a particular chemical vary considerably in ability to make correct or incorrect classifications of toxic or non-toxic samples (Table 2; USEPA 1996). Therefore, the reliability of different types of SECs was evaluated by plotting observed and expected toxicity of samples based on the minimum number of exceedances of these individual SECs. For example, Figure 1a is a plot of the percentage of samples correctly classified as a function of the number of individual ERMs exceeded. In the CR14 test, if an exceedance of an ERM for only one chemical is used to classify a sample as a hit, about 65% of the samples were correctly classified as toxic or not toxic and about 30% of the non-toxic samples were classified as hits (Type I error; false positive). However, the Type II error was only 5% (toxic samples classified as no hit; false negative). As the criterion for a hit is increased to 2 or more ERM exceedances per sample in the CR14 test, the percentage of samples correctly classified increased to about 80%, Type II error increased to almost 20%, and Type I error decreased to < 5%. In the HA14 test, the highest correct classification of about 75 to 85% occurs in the range of about 2 to 4 ERM exceedances (Fig. 1a). In this range, Type I and Type II errors are equal (i.e., cross over of the

TABLE 2. Sediment effect concentrations (SECs) calculated using dry-weight and total concentrations (for metals) in the Hyalella azteca *28-d tests for the entire database.*

SEC	CONC	N	TOX	EFFECT	HIT	CORRECT	TOXHIT	NOTNOT	NOTHIT	TOXNOT
Aluminum (μg/g)										
ERL	14000*	25	11	5	23	44	40	4	52	4
ERM	58000	25	11	5	10	64	24	40	16	20
TEL	26000*	25	11	5	21	44	36	8	48	8
PEL	60000*	25	11	5	7	52	12	40	16	32
NEC	73000*	25	11	5	1	52	0	52	4	44
Arsenic (μg/g)										
ERL	13	52	24	16	27	63	31	33	21	15
ERM	50	52	24	16	11	63	15	48	6	31
TEL	11	52	24	16	32	62	35	27	27	12
PEL	48	52	24	16	11	63	15	48	6	31
NEC	100	52	24	16	2	54	2	52	2	44
Cadmium (μg/g)										
ERL	0.70	62	24	18	35	66	31	35	26	8
ERM	3.9	62	24	18	12	71	15	56	5	24
TEL	0.58	62	24	18	35	66	31	35	26	8
PEL	3.2	62	24	18	14	74	18	56	5	21
NEC	8.0	62	24	18	8	71	11	60	2	27
Chromium (total; μg/g)										
ERL	39	62	24	13	31	60	24	35	26	15
ERM	270	62	24	13	7	73	11	61	0	27
TEL	36	62	24	13	31	60	24	35	26	15
PEL	120	62	24	13	7	73	11	61	0	27
NEC	95	62	24	13	9	73	13	60	2	26
Copper (μg/g)										
ERL	41	52	24	21	26	73	35	38	13	12
ERM	190	52	24	21	14	69	21	48	6	25
TEL	28	52	24	21	29	71	37	35	19	10
PEL	100	52	24	21	17	71	23	46	8	21
NEC	580	52	24	21	3	56	4	52	2	42
Iron (%)										
ERL	20	37	22	7	14	57	27	30	11	32
ERM	28	37	22	7	5	49	11	38	3	49
TEL	19	37	22	7	14	57	27	30	11	32
PEL	25	37	22	7	6	46	11	35	5	49
NEC	29	37	22	7	3	43	5	38	3	54
Manganese (μg/g)										
ERL	730	44	22	11	18	68	30	39	11	20
ERM	1700	44	22	11	8	59	14	45	5	36
TEL	630	44	22	11	26	59	34	25	25	16
PEL	1200	44	22	11	9	61	16	45	5	34
NEC	4500*	44	22	11	1	48	0	48	2	50
Nickel (μg/g)										
ERL	24	62	24	14	20	68	19	48	13	19
ERM	45	62	24	14	7	73	11	61	0	27
TEL	20	62	24	14	24	65	21	44	18	18
PEL	33	62	24	14	14	74	18	56	5	21
NEC	43	62	24	14	9	73	13	60	2	26
Lead (μg/g)										
ERL	35	62	24	19	26	74	27	47	15	11
ERM	99	62	24	19	12	74	16	58	3	23
TEL	37	62	24	19	32	71	31	40	21	8
PEL	82	62	24	19	18	77	21	55	6	16
NEC	130	62	24	19	8	71	11	60	2	27

Continued

TABLE 2. Continued.

SEC	CONC	N	TOX	EFFECT	HIT	CORRECT	TOXHIT	NOTNOT	NOTHIT	TOXNOT
Zinc (mg/g)										
ERL	110	62	24	20	37	63	31	32	29	8
ERM	550	62	24	20	16	74	19	55	6	19
TEL	98	62	24	20	39	63	32	31	31	6
PEL	540	62	24	20	16	74	19	55	6	19
NEC	1300	62	24	20	7	69	10	60	2	29
Naphthalene (ng/g)										
ERL	13	62	24	18	46	53	34	21	40	5
ERM	98	62	24	18	18	61	15	47	15	24
TEL	15	62	24	10	43	56	32	24	37	6
PEL	140	62	24	18	17	63	15	48	13	24
NEC	1400	62	24	18	6	68	8	60	2	31
Fluorene (ng/g)										
ERL	10	62	24	15	58	42	37	5	56	2
ERM	140	62	24	15	16	61	13	48	13	26
TEL	10	62	24	15	58	42	37	5	56	2
PEL	150	62	24	15	13	60	11	48	13	27
NEC	3000	62	24	15	3	63	3	60	2	33
Phenanthrene (ng/g)										
ERL	27	62	24	22	36	65	31	34	27	8
ERM	330	62	24	22	19	66	18	48	13	21
TEL	19	62	24	22	38	68	34	34	27	5
PEL	410	62	24	22	17	63	15	48	13	24
NEC	20000	62	24	22	2	61	2	60	2	37
Anthracene (ng/g)										
ERL	10	62	24	17	56	45	37	8	53	2
ERM	140	62	24	17	18	61	15	47	15	24
TEL	10	62	24	17	56	45	37	8	53	2
PEL	170	62	24	17	15	63	13	50	11	26
NEC	2000	62	24	17	4	65	5	60	2	34
Fluoranthene (ng/g)										
ERL	33	62	24	20	37	56	27	29	32	11
ERM	180	62	24	20	24	55	16	39	23	23
TEL	31	62	24	20	37	56	27	29	32	11
PEL	320	62	24	20	21	56	15	42	19	24
NEC	10000	62	24	20	3	63	3	60	2	35
Pyrene (ng/g)										
ERL	40	62	24	22	39	60	31	29	32	8
ERM	350	62	24	22	22	61	18	44	18	21
TEL	44	62	24	22	36	55	26	29	32	13
PEL	490	62	24	22	21	61	18	45	16	21
NEC	9000	62	24	22	4	65	5	60	2	34
Benzo(a)anthracene (ng/g)										
ERL	19	62	24	19	38	55	27	27	34	11
ERM	200	62	24	19	16	68	16	52	10	23
TEL	16	62	24	19	39	56	29	27	34	10
PEL	280	62	24	19	16	68	16	52	10	23
NEC	3000	62	24	19	6	68	8	60	2	31
Chrysene (ng/g)										
ERL	30	62	24	17	36	52	24	27	34	15
ERM	500	62	24	17	11	73	15	58	3	24
TEL	27	62	24	17	38	52	26	26	35	13
PEL	410	62	24	17	16	68	16	52	10	23
NEC	3000	62	24	17	6	68	8	60	2	31
Benzo(a)pyrene (ng/g)										
ERL	84	62	24	14	25	60	19	40	21	19
ERM	470	62	24	14	8	71	11	60	2	27
TEL	32	62	24	14	31	53	21	32	29	18
PEL	370	62	24	14	12	71	15	56	5	24
NEC	1000	62	24	14	6	68	8	60	2	31

Continued

TABLE 2. Concluded.

SEC	CONC	N	TOX	EFFECT	HIT	CORRECT	TOXHIT	NOTNOT	NOTHIT	TOXNOT
Indeno(1,2,3-c,d)pyrene (ng/g)										
ERL	30	57	21	15	32	56	25	32	32	12
ERM	250	57	21	15	14	67	14	53	11	23
TEL	17	57	21	15	33	54	25	30	33	12
PEL	240	57	21	15	14	67	14	53	11	23
NEC	770	57	21	15	6	70	9	61	2	28
Benzo(g,h,i)perylene (ng/g)										
ERL	13	62	24	18	40	52	27	24	37	11
ERM	280	62	24	18	11	73	15	58	3	24
TEL	16	62	24	18	38	52	26	26	35	13
PEL	250	62	24	18	14	71	16	55	6	23
NEC	1200	62	24	18	6	68	8	60	2	31
Benzo(b,k)fluoranthene (ng/g)										
ERL	37	43	13	6	19	53	14	40	30	16
ERM	71*	43	13	6	14	51	7	44	26	23
TEL	27	43	13	6	19	53	14	40	30	16
PEL	180*	43	13	6	10	56	5	51	19	26
NEC	4000*	43	13	6	1	67	0	67	2	30
Dibenz(a,h)anthracene (ng/g)										
ERL	10	43	13	6	40	37	30	7	63	.
ERM	15*	43	13	6	12	56	7	49	21	23
TEL	10	43	13	6	40	37	30	7	63	.
PEL	28*	43	13	6	8	56	2	53	16	28
NEC	870*	43	13	6	1	67	0	67	2	30
Polycyclic Aromatic Hydrocarbons (PAH-total, ng/g)										
ERL	240	62	24	22	41	56	31	26	35	8
ERM	2200	62	24	22	24	58	18	40	21	21
TEL	260	62	24	22	38	55	27	27	34	11
PEL	3400	62	24	22	20	61	16	45	16	23
NEC	62000	62	24	22	5	66	6	60	2	32
PAH-low molecular weight (ng/g)										
ERL	80	62	24	21	37	63	31	32	29	8
ERM	650	62	24	21	21	63	18	45	16	21
TEL	76	62	24	21	37	63	31	32	29	8
PEL	1200	62	24	21	16	61	13	48	13	26
NEC	29000	62	24	21	4	65	5	60	2	34
PAH-high molecular weight (ng/g)										
ERL	170	62	24	22	42	55	31	24	37	8
ERM	1700	62	24	22	24	58	18	40	21	21
TEL	190	62	24	22	38	52	26	26	35	13
PEL	2300	62	24	22	23	60	18	42	19	21
NEC	33000	62	24	22	4	68	8	60	2	31
Polycyclic Aromatic Hydrocarbons (PCB-total, ng/g) Polychlorinated biphenyls										
ERL	50	29	10	5	9	69	17	52	14	17
ERM	730	29	10	5	3	76	10	66	0	24
TEL	32	29	10	5	9	69	17	52	14	17
PEL	240	29	10	5	3	76	10	66	0	24
NEC	190	29	10	5	4	72	10	62	3	24

*: Unreliable SEC (e.g., less than five of the samples were designated as toxic for the chemical or the number of toxic samples with concentrations below the SEC was greater than the number of toxic samples with concentrations above the SEC).

N: Total number of samples used to calculate each SEC.

TOX: Number of toxic samples.

EFFECT: Number of toxic samples where the concentration of a chemical was greater than the mean concentration of the chemical in the non-toxic samples at a site.

HIT: Number of samples with concentrations greater than the SEC.

TOTAL

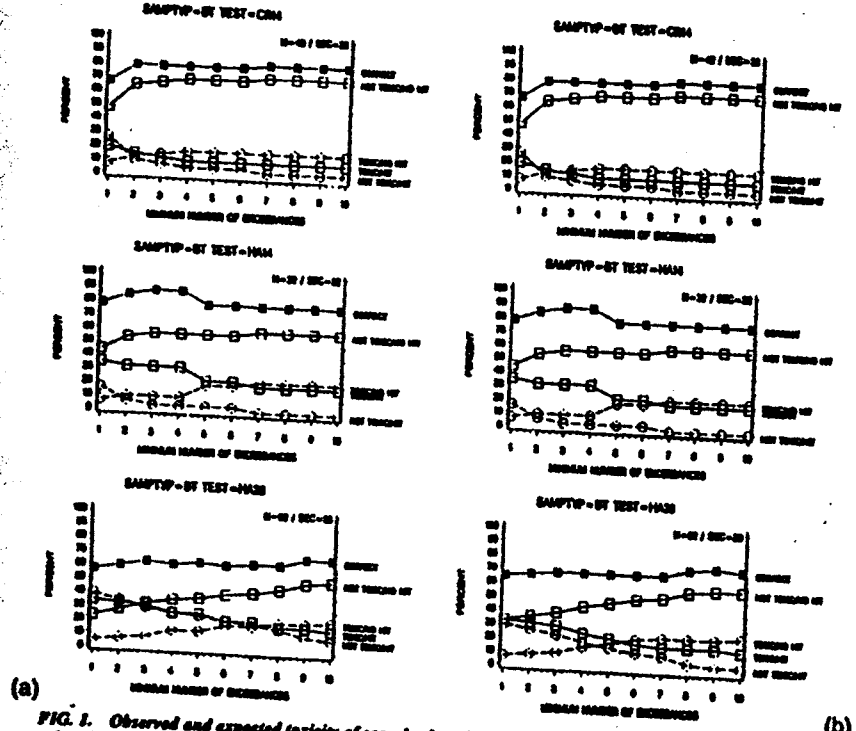
CORRECT: Percentage samples correctly classified as toxic.

TOXHIT: Percentage of toxic samples correctly classified as toxic (toxic sample and hit).

NOTNOT: Percentage of non-toxic samples correctly classified as not toxic (non-toxic sample and no hit).

NOTHIT: Percentage of non-toxic samples incorrectly classified as toxic (Type I error; non-toxic sample and hit [false positive])

TOXNOT: Percentage of toxic samples incorrectly classified as not toxic (Type II error; toxic sample and no hit [false negative]).

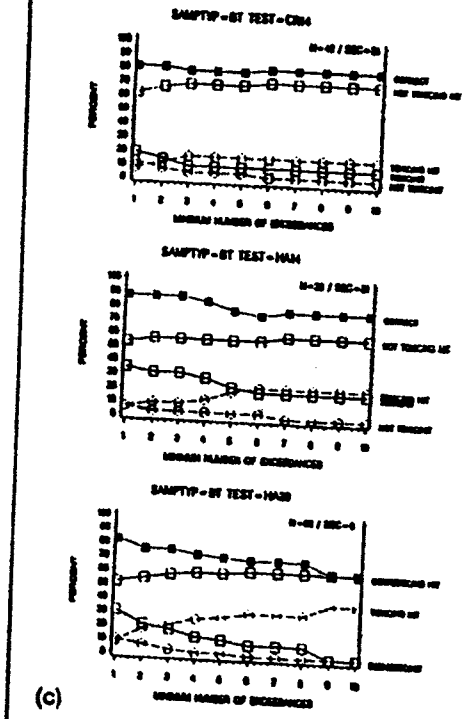


(a) (b)

FIG. 1. Observed and expected toxicity of samples based on the minimum number of ERM exceedances using dry-weight concentrations. N = number of samples and SEC = number of SECs used. Figure 1a uses all individual ERMs listed in USEPA (1996) or Table 2 regardless of percentage correct classification by these individual ERMs. Figure 1b uses only those chemicals for which individual ERMs correctly classify ≥ 60% of the samples. Figure 1c uses only those chemicals for which individual ERMs correctly classify ≥ 70% of the samples.

not toxic/hit line with the toxic/no hit line). In contrast, the highest correct classification of about 70% for the HA28 test occurs across 3 to 10 exceedances with Type I equal to Type II error at about 7 exceedances.

Figure 1b plots correct classification of samples as a function of ERM exceedances using only those chemicals for which individual ERMs correctly classify ≥ 60% of the samples (Fig. 1a used all the individual ERMs reported in USEPA (1996) regardless of percentage correct classification). The number of ERMs used in the calculations remained the



(c)

same for the CR14 and HA14 tests (Fig. 1a vs 1b); however, selecting a criterion of 60% reduced the number of ERMs used in the calculations from 25 to 20 in the HA28 test (Table 2; USEPA 1996). As a result, correct classification increased by about 5 to 10% across 1 to 10 exceedances in the HA28 test (Fig. 1a vs 1b). This increased correct classification resulted from a reduction in Type I error (not toxic/hit).

Selecting a criterion of 70% correct classification for individual ERMs reduced the number of ERMs used in the calculation from 22 to 21 in the CR14 and HA14 tests, and from 25 to 9 in the HA28 test (Fig. 1a vs 1c; Table 2; USEPA 1996). At 1 ERM exceedance, correct classifications increased in all three tests by about 10 to 20% and both Type I and Type II error were only about 10% (Fig. 1a vs 1c). However, increasing the minimum number of exceedances decreased correct classifications of samples in the HA28 test (Fig. 1c). This drop in correct classification results from increased Type II error (toxic/no hit) when fewer ERMs are used in the calculation (Fig. 1a vs 1c).

In summary, correct classification of samples can be improved by using ERMs with a higher percentage of correct classification. For example, using a 70% criterion for selection of ERMs, only 1 ERM had to be exceeded in any of the three tests to achieve about 80 to 90% correct classification of samples with only about 10% Type I and Type II errors. By lowering the criterion to 60% for selection of ERMs, exceeding 2 to 5 ERMs still resulted in about 70 to 80% correct classification of samples.

Figure 2a plots correct classification of samples as a function of ERL exceedances (regardless of percentage correct classification by individual ERLs). Type II error (false negatives) remains relatively low (< 10%) across the range of 1 to 10 ERL exceedances. The highest correct classification of about 60 to 70% occurs at > 5 to 6 ERL exceedances. However, Type I error (false positives) was always higher (> 20 to > 40%) compared to Type II error resulting in lower percentage correct classification with ERLs compared to ERMs (Fig. 1a vs 2a).

Selecting a criterion of 60% did not substantially improve classification by ERLs in the CR14 or HA14 tests, but correct classification increased about 10 to 30% in the HA28 test (Fig. 2a vs 2b; Table 2). Using a 70% criterion for selection of ERLs, 70 to 90% of the samples were correctly classified with 8 to 10 exceedances in the CR14 and HA14 tests and with 1 to 2 exceedances in the HA28 test (Fig. 2a vs 2c). However, in the HA28 test only non-toxic samples are correctly classified at 3 or more ERL exceedances because of the high Type II error (toxic/no hit) resulting from using just 2 ERLs (Fig. 2c). In summary, correct classification of samples can be improved by using multiple ERLs with a high percentage of correct classification. However, samples which exceeded multiple ERLs were typically samples which also exceeded ERMs. Hence, exceeding a few ERMs or multiple ERLs resulted in similar correct classification of samples.

For NECs, about 70 to 90% of samples are cor-

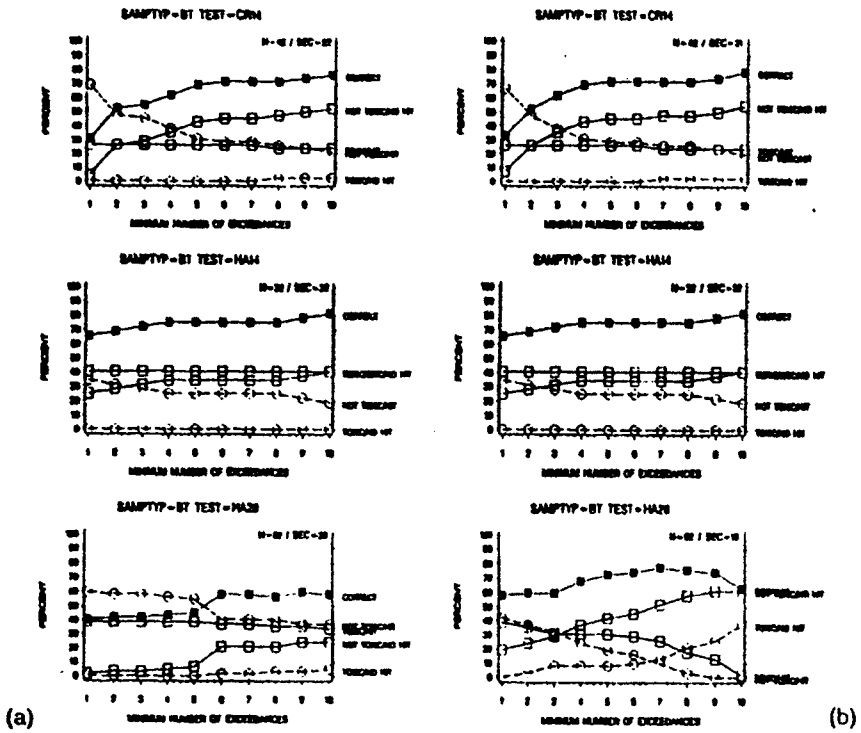


FIG 2. Observed and expected toxicity of samples based on the minimum number of ERL exceedances using dry-weight concentrations. See legend of Figure 1 for a description of Figure 2.

rectly classified at 1 to 10 exceedances regardless if all the NECs are used (Fig. 3) or if the 60% criterion or the 70% criterion are used (USEPA 1996; Table 2). Type I and Type II errors are generally equal at about 1 to 3 exceedances. However, Type II error (false negatives) often starts at 5 to 10% with only 1 exceedance. Increasing the minimum number of exceedances decreased correct classification of samples. This drop in correct classification results from increased Type II error if multiple exceedances of NECs are required to classify a sample as toxic.

Figure 4 directly compares correct classification, Type I error, and Type II error as a function of the minimum number of ERL, ERM, or NEC exceedances for the HA28 test listed in Table 2. ERLs only classify about 40 to 60% of the samples correctly. The higher Type I error associated with ERLs compared to either ERMs or NECs results in this lower correct classification by ERLs. ERMs

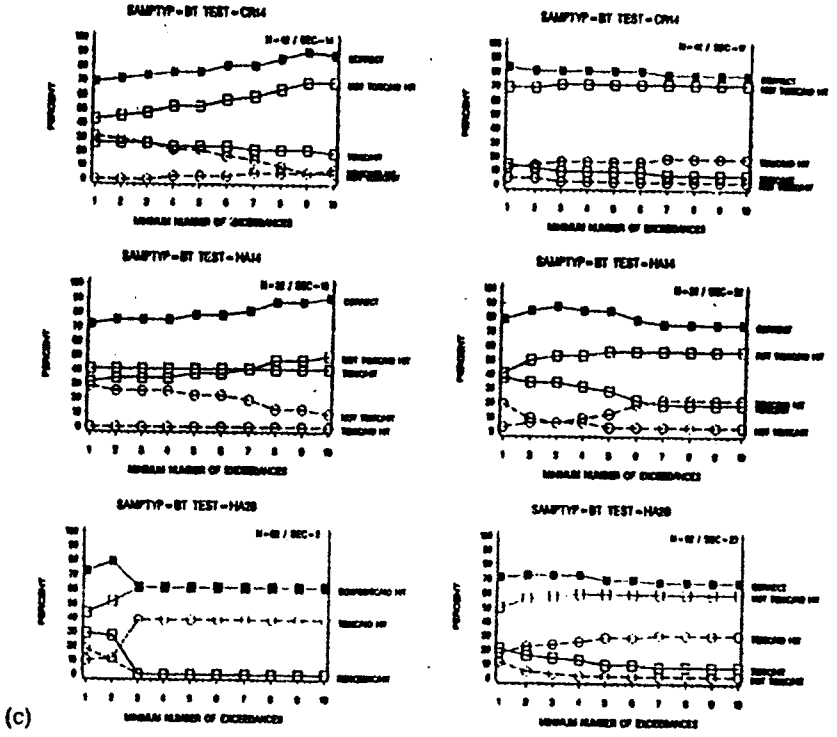


FIG 3. Observed and expected toxicity of samples based on the minimum number of NEC exceedances using dry-weight concentrations and using all individual NECs regardless of the percent correct classification. See legend of Figure 1 for a description of Figure 3.

and NECs correctly classify a similar percentage of samples; however, Type II error is consistently higher with NECs compared to either ERMs or ERLs. In summary, these analyses indicate the reliability of correct classifications is similar between ERMs and NECs; however, ERMs minimize Type I error relative to ERLs and minimize Type II error relative to NECs. The high Type I error typically associated with ERLs is the primary reason Long *et al.* (1995) and MacDonald *et al.* (1996) recommend ERMs and PELs, but not ERLs or TELs should be used to predict toxicity of samples. However, ERLs and TELs can be used to efficiently identify concentrations below which toxicity is rarely observed.

Table 2 and USEPA (1996) list SECs for PAHs and total PCBs calculated using dry-weight concentrations. USEPA (1996) also lists SECs for PAHs and total PCBs calculated using sediment concen-

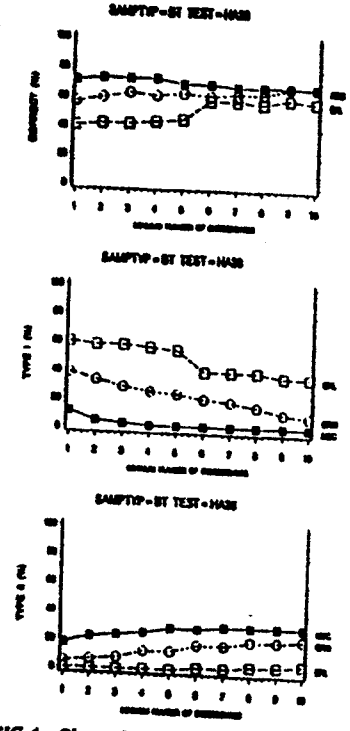


FIG 4. Observed and expected toxicity of HA28 samples based on the minimum number of NEC, ERM, and ERL exceedances using dry-weight concentrations regardless of percentage correct classification by these individual ERMs. See legend of Figure 1 for a description of Figure 4.

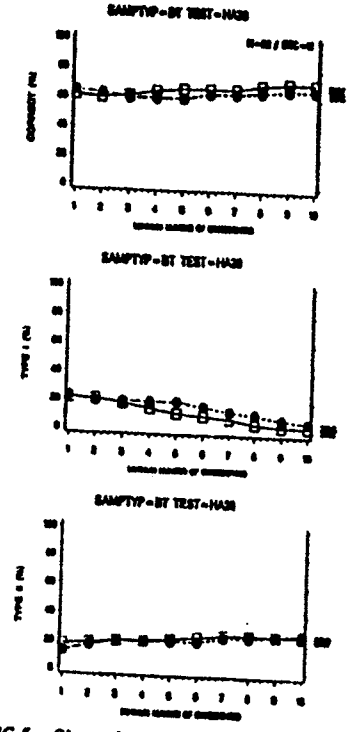


FIG 5. Observed and expected toxicity of HA28 samples based on the minimum number of ERM exceedances using only individual ERMs for PAHs or PCBs calculated using: (1) dry-weight concentrations or (2) sediment concentrations normalized to total organic carbon (TOC) concentrations which correctly classify ≥ 60% of the samples. See legend of Figure 1 for a description of Figure 5.

trations normalized to total organic carbon (TOC) concentrations. None of the individual ERMs calculated using sediment concentrations normalized to TOC concentrations correctly classified ≥ 70% of the samples (USEPA 1996). Therefore, Figure 5 plots correct classification of samples as a function of ERM exceedances using only individual ERMs which correctly classify ≥ 60% of the samples in the HA28 test (n = 11 SECs). Correct classification of samples ranged between 60 to 70% and Type I

and Type II errors were similar based on exceedances of ERMs using either dry-weight concentrations or sediment concentrations normalized to TOC concentrations. Correct classifications were also similar using PELs and NECs calculated using dry-weight concentrations or calculated sediment concentrations normalized to TOC concentrations.

One would expect SECs calculated using sediment concentrations normalized to TOC concentrations to be more reliable than SECs calculated using dry-weight concentrations since TOC reportedly controls the bioavailability of non-ionic organic contaminants such as PAHs and PCBs in sediment (Di Toro *et al.* 1991). The range of TOC concentrations in our database was relatively narrow compared to the ranges of contaminant concentrations. The mean concentration of TOC was 2.7% with a 95% confidence interval of only ± 0.65% (n = 62). In contrast, the concentration ranges of contaminants normalized to dry weight typically varied by several orders of magnitude. Therefore, normalizing dry-weight concentrations to a relatively narrow range of TOC concentrations had little influence on relative concentrations of contaminants among samples. Similar findings were reported by Barrick *et al.* (1988) for AETs and Long *et al.* (1995) for ERMs calculated using sediment concentrations normalized to TOC concentrations. It is surprising that there is not at least a trend of increased reliability with SECs calculated using sediment concentrations normalized to TOC concentrations. The lower reliability of SECs calculated using sediment concentrations normalized to TOC concentrations may indicate PAHs and PCBs were not causing the toxicity, but were only associated with the toxic chemicals. Use of sediment toxicity identification evaluations (TIE) or studies using spiking of sediment are needed to establish these cause and effect relationships (Ankley and Thomas 1992, Lamberson and Swartz 1992, Ankley *et al.* 1996).

Predictive Ability of SECs

The predictive ability of SECs in this paper was evaluated by first calculating SECs using just the Great Lakes (GL) portion of the database (n = 27 samples). We were able to calculate GL SECs primarily for total metals, simultaneously extracted metals (SEM metals), and PAHs (USEPA 1996). These GL SECs were then used to predict responses in independent HA28 and CR14 tests with Clark Fork River (CFR) sediments (n = 15 samples). The CFR sediments contained elevated concentrations of As, Cd, Cu, Pb, and Zn. Concentrations of PAHs, PCBs, and chlorinated pesticides were not elevated in these samples (Kemble *et al.* 1994). In the CFR tests, 7% of the sediments were toxic in the CR14 test and 53% of the samples were toxic in the HA28 test (Table 1).

Figure 6 plots correct classification of CFR samples as a function of the number of exceedances of individual GL ERMs which correctly classified ≥ 70% of the GL samples. For the CR14 test, about 80 to 90% of the CFR samples were correctly classified at 1 to 2 exceedances of GL ERMs. The majority of the samples were not toxic and did not exceed GL ERMs for the CR14 test. Type II error (toxic/no hit) was always < 10% and Type I error (not toxic/hit) was 20% at 1 exceedance dropping to < 10% with > 2 GL ERM exceedances. For the

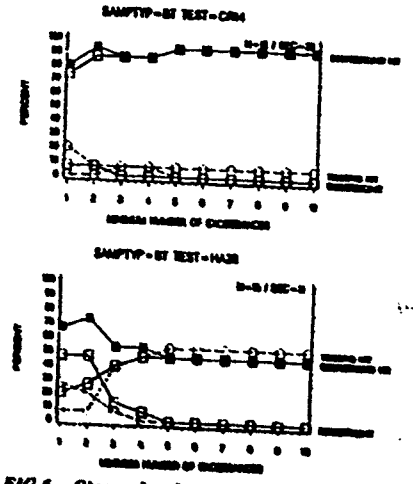


FIG 6. Observed and predicted toxicity of Clark Fork River samples based on the minimum number of Great Lakes ERM exceedances using dry-weight concentrations and using only those chemicals for which individual SECs correctly classify ≥ 70% of the Great Lakes samples. See legend of Figure 1 for a description of Figure 6.

HA28 test, about 70% of the CFR samples were correctly classified, Type II error was <10%, and Type I error was 20 to 30% at 1 to 2 GL ERM exceedances. Above 2 GL ERM exceedances in the HA28 test, Type II error increases more than the decrease in Type I error, resulting in a substantial drop in correct classification of samples. Evaluations using GL PELs and GL NECs resulted in similar predictive ability compared to GL ERMs for the CR14 and HA28 tests with CFR sediments.

The CFR sediments primarily contained high concentrations of Cu and Zn resulting in exceedances of GL ERMs for these two metals. Requiring more than 2 exceedances of GL ERMs resulted in a high Type II error (toxic samples misclassified as not toxic). Hence, classification based on multiple exceedances of SECs in a preliminary screening of sediments which contain a limited number of contaminants may result in high Type II error. For example, Type II error was < 10% and Type I error was 10 to 30% at 1 to 2 GL ERM exceedances in both the CR14 and HA28 tests with CFR sediments, but Type II error was high with multiple exceedances in the HA28 test (Fig. 6). Therefore, a low number of SEC exceedances should be used to conduct a preliminary screening to predict the potential for toxicity in the absence of actual toxicity testing. This would minimize the potential for false negatives (i.e., Type II error) at the risk of accepting higher false positives (i.e., Type I error).

We have included this one example of how the predictive ability of SECs can be evaluated using an independent data set. We are currently in the process of using our SECs calculated from the entire database to predict the response of *Hyalella azteca* and *Chironomus riparius* in a variety of independent data sets generated by other laboratories (i.e., McGee *et al.* 1994, Pastorok *et al.* 1994, Schlekat *et al.* 1994, Batts and Cubbage 1995, Day *et al.* 1995, Hoke *et al.* 1995, J. Field, NOAA, Seattle, WA, and M.D. Sprenger, USEPA, Edison, NJ, unpublished data).

Comparability to Published SECs

Example comparisons are plotted of our SECs relative to other published SECs for benzo[a]pyrene (BaP; Fig. 7) and copper (Fig. 8). Our SECs are typically lower than the AET (Figs. 7 and 8) and EQP values (Fig. 7) and are relatively similar to paired marine ERMs, ERLs, PELs, or TELs and freshwater PELs or TELs (Figs. 7 and 8; Smith *et al.* 1996, USEPA 1996). The SECs based on EQP and AET approaches are typically near the maximum concentration for the particular chemical in our database. This is not surprising since EQP values represent concentrations of single chemicals predicted to be toxic whereas the other SECs listed in Figures 7 and 8 represent concentrations of a chemical associated with toxicity in mixtures of chemicals in field-collected sediments (Hoke *et al.* 1995).

Smith *et al.* (1996) reported 14 of their 23 TELs and 15 of their 23 PELs were within a factor of 3 for at least two other published SECs. These results indicate SECs developed using a variety of approaches and data sets are often comparable. The SECs calculated by Smith *et al.* (1996) were also comparable to our SECs listed in Table 2 for HA28 tests. However, reliability of the SECs in Smith *et al.* (1996) was generally lower than the reliability of our SECs. The database used by Smith *et al.* (1996) to calculate SECs included our data and a variety of additional data sources from North America. This lower reliability of SECs reported by Smith *et al.* (1996) resulted from including data for additional species from studies reporting no effects without matching effect data (i.e., intolerant species or short exposure duration) or by including data from benthic community surveys (i.e., difficult to compare sediment chemistry to distributions of benthos). Additional comparisons are ongoing to further evaluate comparability and predictive ability of published SECs to our SECs using additional independent data sets (i.e., McGee *et al.* 1994, Pastorok *et al.* 1994, Schlekat *et al.* 1994, Batts and Cubbage 1995, Day *et al.* 1995, Hoke *et al.* 1995; J. Field, NOAA, Seattle, WA and M.D. Sprenger, USEPA, Edison, NJ, unpublished data).

Ultimately, the best measure of comparability among SECs is not to compare similarity in absolute concentrations, but to compare how different types of SECs correctly (or incorrectly) predict toxicity in independent samples. For example, Figure 9 plots predictions of toxicity in our HA28 tests as a function of exceedances of freshwater PELs (PELF; Smith *et al.* 1996), *Hyalella azteca* AETs (AET5; Batts and Cubbage 1995; assuming 2% TOC), EQP (USEPA 1988, Hoke *et al.* 1995; assuming 2% TOC) and SLCs (SLC1; lowest effect level for Screening Level Concentrations; Persoud *et al.* 1992). At 1 to 6 exceedances of these published PELs, AETs, and EQP values, toxicity is correctly predicted in about 60 to 80% of the samples whereas SLCs only correctly predict toxicity in

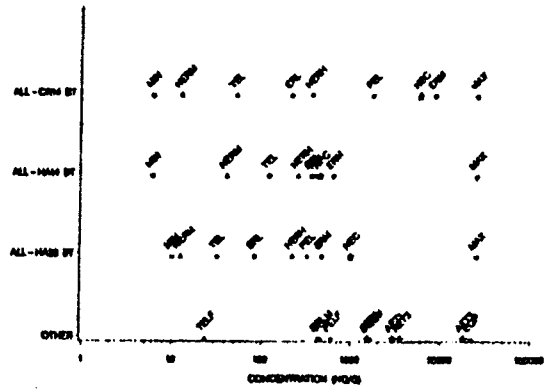


FIG 7. Comparability of our SECs for the entire database to published SECs for benzo[a]pyrene (BaP) based on dry-weight concentrations. Marine ERL and ERM (ERLM and ERMM; Long et al. 1995); marine TEL and PEL (TELM and PELM; MacDonald, et al. 1995); freshwater TEL and PEL (TELF and PELF; Smith et al. 1996); (4) marine AETs (AET1 for amphipods, AET2 for oysters, AET3 for Microtox, AET4 for benthos; Barrick et al. 1988); freshwater AETs (AET5 for Hyalella azteca and AET6 for Microtox; Batts and Cubbage 1995 (assumed 2% total organic carbon (TOC)); (6) Screening Level Concentrations (SLC1 for lowest effect level, SLC2 for severe effect level (assumed 2% TOC); Persaud et al. 1992); and (7) EQP (USEPA 1988, Hoke et al. 1995; assumed 2% TOC).

about 40 to 60% of the samples. The higher Type I error (false positives) associated with SLCs compared to the other values results in this lower correct prediction by SLCs. The PEL, AET, and EQP values correctly predict toxicity in a similar percentage of samples; however, Type II error (false negatives) is consistently higher with AET and EQP values compared to either PELs or SLCs. In summary, these analyses indicate predictive ability is similar between published PEL, AET, and EQP values; however, these PELs minimize Type I error relative to SLCs and minimize Type II error relative to AET and EQP values. In addition, the predictive ability of these published SECs is comparable to the reliability of our SECs listed in Table 2.

Throughout this paper we have evaluated the reliability using the frequency of exceeding individual SECs. Canfield *et al.* (1996a, 1996b) and Kemble *et al.* (1996) evaluated the reliability of our ERMs using a toxic quotient approach. A toxic quotient was calculated for each sample by first dividing the concentration of individual chemicals by their respective ERM and then summing each of these individual values. Figure 10 plots the relationship between the frequency of ERM exceedances and the sum of the ERM toxic quotient for HA28 samples using all ERMs regardless of the percent correct classification. The frequency of observed toxicity in samples increases at either a sum ERM toxic quotient of about 10 to 20 or at a frequency of ERM exceedances of about 3 to 7. A similar relationship is evident if only individual ERMs are used that correctly classify ≥ 60% or ≥ 70% of the samples; however, a lower number of ERM exceedances or lower sum ERM toxic quotients are needed to consistently estimate observed toxicity.

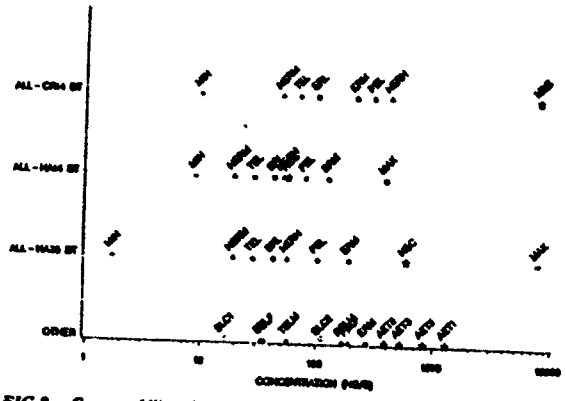


FIG 8. Comparability of our SECs to published SECs for copper based on dry-weight concentrations. See legend of Figure 7 for an description of the abbreviations in Figure 8.

In summary, either the sum toxic quotient approach or the frequency of SEC exceedances are equally reliable at classifying samples as either toxic or not toxic in our database (Canfield *et al.* 1996a, 1996b; Kemble *et al.* 1996).

CONCLUSIONS

ERMs and ERLs are generally as reliable as paired PELs and TELs at classifying samples as either toxic or not toxic in our database. Reliability of the SECs in terms of correctly classifying sediment samples is similar between ERMs and NECs; however, ERMs minimize Type I error (false positives) relative to ERLs and minimize Type II error (false negatives) relative to NECs. ERMs and NECs rather than ERLs should be used to predict toxicity of samples due to the lower Type I error associated with them. However, ERLs can be used to efficiently identify concentrations below which toxicity is rarely observed. Correct classification of samples can be improved by using only the most reliable individual ERMs or NECs for chemicals (i.e., those with a higher percentage of correct classification). When SECs are used to conduct a preliminary screening to predict the potential for toxicity in the absence of actual toxicity testing, a low number of SEC exceedances should be used to minimize the potential for false negatives (i.e., Type II error); however, the risk of accepting higher false positives (i.e., Type I error) is increased.

SECs generated using data from field-collected samples should not be used independently to establish trigger levels for clean up of sediments. The strength of SECs developed using data from tests with field-collected sediments is in their use in predicting the potential for toxicity in independent field-collected sediment samples. A primary use of SECs developed with field-collected sediments should be to provide guidance for determining sites which may require further investigation. The ability of any SEC or sediment toxicity test to predict benthic community effects should be considered before either of these approaches are used to routinely evaluate sediment quality (Canfield *et al.* 1994, 1996a, 1996b).

Our SECs were calculated from toxicity tests with field-collected samples. If a chemical concentration exceeds an SEC generated using data from these tests with field-collected samples, it does not necessarily mean the chemical *caused* the observed effect. Rather, the SEC is the concentration of a chemical that is *associated* with the effect. Field-collected sediments typically contain complex mixtures of contaminants. Additional information is needed to identify the specific contaminants that were actually responsible for the toxicity. Confirmation of sediment toxicity due to individual or groups of contaminants or the interactive effects of sediment toxicants can be evaluated by using TIE procedures (Ankley and Thomas 1992, Ankley *et al.* 1996) or by conducting toxicity tests with chemicals spiked into sediments (Lamberson and Swartz 1992). Once the probable cause(s) of sediment toxicity has been identified, better decisions can be made regarding remediation options.

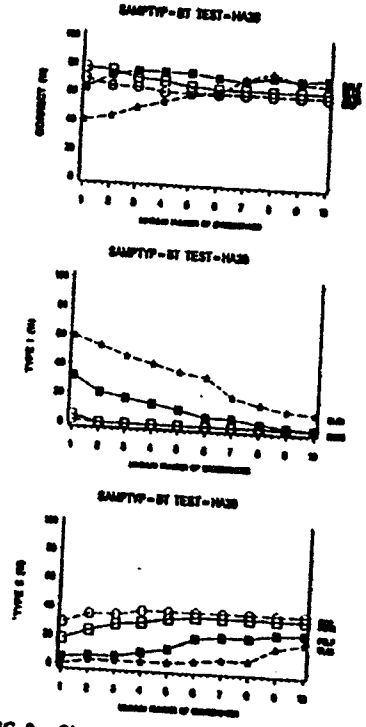


FIG. 9. Observed and predicted toxicity of HA28 samples based on dry-weight concentrations and the minimum number of exceedances of published freshwater PELs (PELF; Smith et al. 1996), Hyalella azteca AET: (AETS; Batts and Cubbage 1995; assumed 2% TOC), EQP (USEPA 1988, Hoke et al. 1995; assumed 2% TOC) and SLCs (SLC1; lowest effect level for Screening Level Concentrations; Persaud et al. 1993). See legend of Figure 1 for a description of Figure 9.

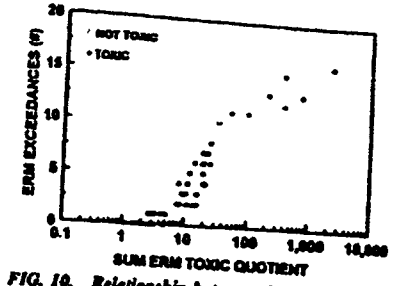


FIG. 10. Relationship between the frequency of ERM exceedances and the sum of the ERM toxic quotient for toxic and non-toxic HA28 samples using all ERMs regardless of the percent correct classification using dry-weight concentrations. Adapted from Canfield et al. (1996a,b) and Kemble et al. (1996).

ACKNOWLEDGMENTS

Although the sediment effect concentrations (SECs) can be used as guidance for evaluating contaminated sediment, there is no intent expressed or implied that these SECs represent USEPA or National Biological Service (NBS) criteria. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. We would like to thank the following individuals for insight and guidance in developing procedures to evaluate SECs: G.T. Ankley, USEPA, Duluth, MN; W.G. Brumbaugh, NBS, Columbia, MO; F.D. Calder and G. Sloane, Florida Department of Environmental Protection, Tallahassee, FL; R. Fleming, WRc, Marlow, Bucks, United Kingdom; P.F. Landrum, NOAA, Ann Arbor, MI, E.R. Long and J. Field, NOAA, Seattle, WA; D.D. MacDonald, MacDonald Environmental Sciences, Ladysmith, BC, M. Briley, USEPA, Washington, DC, and S. Smith,

Environment Canada, Hull, Quebec. The information in this report has been funded in part by the USEPA under Contract Number DW14933874 with the Midwest Science Center, Columbia, MO.

REFERENCES

American Society for Testing and Materials (ASTM). 1995. Standard test methods for measuring the toxicity of sediment-associated contaminants with freshwater invertebrates. E1706-95b. In *ASTM Annual Book of Standards*, Vol. 11.05, pp. 1204-1285. Philadelphia, PA.

Ankley, G.T., and Thomas, N. 1992. Interstitial water toxicity identification evaluation approach. In *Sediment classification methods compendium*, pp. 5-1 to 5-14. EPA 823-R-92-006, Washington, DC.

_____, Schubauer-Berigan, M.K., and Dierkes, J.R. 1996. Application of toxicity identification evaluation techniques to pore water from Buffalo River sediments. *J. Great Lakes Res.* 22:534-544.

Barrick, R., Becker, S., Brown, L., Beller, H., and Pastorok, R. 1988. *Sediment quality values refinement: 1988 Update and Evaluation of Puget Sound AET, Vol. I.* PTI Contract C717-01, PTI Environmental Services, Bellevue, WA.

Batts, D., and Cubbage, J. 1995. *Summary guidelines for contaminated freshwater sediments.* Washington Department of Ecology, Publication no. 95-308, Olympia, WA.

Becker, D.S., Barrick, R.C., and Read, L.B. 1989. *Evaluation of the AET approach for assessing contamination in sediments in California.* PTI Contract C739-01, PTI Environmental Services, Bellevue, WA.

Burton, G.A., Ingersoll, C.G., Burnett, L., Henry, H., Hinman, M., Klaine, S., Landrum, P., and Ross, P. 1996. A comparison of sediment toxicity test methods at three Great Lakes Areas of Concern. *J. Great Lakes Res.* 22:495-511

Canadian Council of Ministers of the Environment. 1995. *Protocol for the derivation of Canadian sediment quality guidelines for the protection of aquatic life.* Prepared by the CCME Task Group on water quality guidelines, Ottawa, ON.

Canfield, T.J., Kemble, N.E., Brumbaugh, W.G., Dwyer, F.J., Ingersoll, C.G., and Fairchild, J.F. 1994. Use of benthic invertebrate community structure and the sediment quality triad to evaluate metal-contaminated sediment in the upper Clark Fork River, Montana. *Environ. Toxicol. Chem.* 13:1999-2012.

_____, Dwyer, F.J., Fairchild, J.F., Ingersoll, C.G., Kemble, N.E., Mount, D.R., La Point, T.W., Burton, G.A., and Swift, M.C. 1996a. Assessing contamination of Great Lakes sediment using benthic invertebrates and the sediment quality triad approach. *J. Great Lakes Res.* 22:565-583.

_____, E.L. Brunson, F.J. Dwyer, C.G. Ingersoll, and N.E. Kemble. 1996b. Assessing upper Mississippi river sediments using benthic invertebrates and the sediment quality triad. Manuscript in review.

Day, K.E., Dutka, B.J., Kwan, K.K., Batista, N., Reynoldson, T.B., and Metcalfe-Smith, J.L. 1995. Correlations between solid-phase microbial screening assays, whole-sediment toxicity tests with macroinvertebrates, and *in situ* benthic community structure. *J. Great Lakes Res.* 21:192-206.

Di Toro, D.M., Mahony, J.H., Hansen, D.J., Scott, K.J., Hicks, M.B., Mayr, S.M., and Redmond, M. 1990. Toxicity of cadmium in sediments: The role of acid volatile sulfides. *Environ. Toxicol. Chem.* 9:1487-1502.

_____, Zarba, C.S., Hansen, D.J., Berry, W.J., Swartz, R.C., Cowan, C.E., Pavlou, S.P., Allen, H.E., Thomas, N.A., and Paquin, P.R. 1991. Technical basis for establishing sediment quality criteria for nonionic organic chemicals using equilibrium partitioning. *Environ. Toxicol. Chem.* 10:1541-1583.

Fox, R.G., and Tuchman, M. 1996. The assessment and remediation of contaminated sediments (ARCS) program. *J. Great Lakes Res.* 33:493-494.

Hall, N.E., Fairchild, J.F., La Point, T.W., Heine, P.R., Ruessler, D.S., and Ingersoll, C.G. 1996. Problems and recommendations in using algal toxicity testing to evaluate contaminated sediments. *J. Great Lakes Res.* 22:545-556.

Hoke, R.A., Ankley, G.T., and Peters, J.F. 1995. Use of a freshwater sediment quality database in an evaluation of sediment quality criteria based on equilibrium partitioning and screening-level concentrations. *Environ. Toxicol. Chem.* 14:451-459.

Ingersoll, C.G., and Nelson, M.K. 1990. Testing sediment toxicity with *Hyalella azteca* (Amphipoda) and *Chironomus riparius* (Diptera). In *Aquatic Toxicology and Risk Assessment: Thirteenth Volume*, ASTM STP 1096, eds. W.G. Landis and W.H. van der Schalie, pp. 93-109. American Society for Testing and Materials, Philadelphia, PA.

Kemble, N.E., Brumbaugh, W.G., Brunson, E.L., Dwyer, F.J., Ingersoll, C.G., Monda, D.P., and Woodward, D.F. 1994. Toxicity of metal-contaminated sediments from the upper Clark Fork River, Montana to aquatic invertebrates in laboratory exposures. *Environ. Toxicol. Chem.* 13:1985-1997.

_____, E.L. Brunson, T.J. Canfield, F.J. Dwyer, and C.G. Ingersoll. 1996. Laboratory toxicity test with *Hyalella azteca* exposed to whole sediments from the Upper Mississippi River. Manuscript in review.

Lamberson, J.O. and Swartz, R.C. 1992. Spiked-sediment toxicity test approach. In *Sediment classification methods compendium*, pp. 4-1 to 4-10. EPA 823-R92-006, Washington, DC.

Long, E.R., and Morgan, L.G. 1991. *The potential for biological effects of sediment-sorbed contaminants tested in the National Status and Trends Program.* NOAA Technical Memorandum NOS OMA 52, Seattle, WA.

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

Lorenzato, S.G., Gunther, A.J., and O'Conner, J.M. 1991. *Summary of a workshop concerning sediment quality assessment and development of sediment quality objectives.* California State Water Resources Control Board, Sacramento, CA.

MacDonald, D.D. 1994. *Approach to the assessment of sediment quality of Florida coastal waters. Volume 1-Development and evaluation of sediment quality assessment guidelines.* Report prepared for the Florida Department of Environmental Protection, Tallahassee, FL. November 1994.

_____, Carr, R.S., Calder, F.D., Long, E.R., and Ingersoll, C.G. 1996. Development and evaluation of sediment quality guidelines for Florida coastal waters. *Ecotoxicology*: In press.

McGee, B.L., Schlekat, C.E., Boward, D.M., and Wade, T.L. 1994. Sediment contamination and biological effects in a Chesapeake Bay marina. *Ecotoxicology* 4: 39-59.

Papoulias, D.M., and Buckler, D.R. 1996. Mutagenicity of Great Lakes sediments. *J. Great Lakes Res.* 22: 591-601.

_____, Buckler, D.R., and Tillitt, D.E. 1996. Optimization of the Ames/Salmonella mutagenicity assay for use with extracts of aquatic sediments. *J. Great Lakes Res.* 22:584-590.

Pastorok, R.A., Peek, D.C., Sampson, J.R., and Jacobson, M.A. 1994. Ecological risk assessment for river sediments contaminated by creosote. *Environ. Toxicol. Chem.* 13:1929-1941.

Persuad, D., Jaagumagi, R., and Hayton, A. 1992. *Guidelines for the protection and management of aquatic sediment quality in Ontario.* ISBN 0-7729-9248-7. Ontario Ministry of the Environment, Water Resources Branch, Toronto, ON.

Roach, R.W., Carr, R.S., Howard, C.L., and Cain, D.W. 1993. *Assessment of produced water impacts in Galveston Bay System.* U.S. Fish and Wildlife Report, Clear Lake Ecological Services Office, Houston, TX.

Ross, P.E., Burton, G.A. Jr., Creceliut, E.A., Filkins, J.C., Glesy, J.P. Jr., Ingersoll, C.G., Landrum, P.F., Mac, M.J., Murphy, T.J., Rathbun, J.E., Smith, V.E., Tatem, H., and Taylor, R.W. 1992. Assessment of sediment contamination at Great Lakes areas of concern: The ARCS program. *J. Aquat. Ecosystem Health* 1:193-200.

Schlekat, C.E., McGee, B.L., Boward, D.M., Reinharz, E., Velinsky, D.J., and Wade, T.L. 1994. Tidal river sediments in the Washington, D.C. area. III. Biological effects associated with sediment contamination. *Estuaries* 17:334-344.

Smith, S.L., MacDonald, D.D., Keenleyside, K.A., Ingersoll, C.G., and Field, J. 1996. A preliminary evaluation of sediment quality assessment values for freshwater ecosystems. *J. Great Lakes Res.* 22:624-638.

Statistical Analysis Systems. 1992. *SAS Users Guide.* Cary, SC: SAS Institute, Inc.

Swartz, R.C., and DiToro, D.M. 1997. Sediments as complex mixtures: An overview of methods to assess ecotoxicological significance. In *Ecological Risk Assessments of Contaminated Sediments*, Chapter 16, G.R. Biddinger, T. Dillon, and C.G. Ingersoll (eds.). SETAC Press, Pensacola, FL: in press.

U.S. Environmental Protection Agency. 1988. *Interim sediment criteria values for nonpolar hydrophobic organic contaminants.* EPA, SDC #17, Criteria and Standards Division, Washington, DC.

_____. 1992. *Sediment classification methods compendium.* EPA 823-R92-006, Washington, DC.

_____. 1993. *Assessment and remediation of contaminated sediments (ARCS) program. Biological and chemical assessment of contaminated Great Lakes sediment.* EPA 905-R93-006, Chicago, IL.

_____. 1994. *Methods for measuring the toxicity and bioaccumulation of sediment-associated contaminants with freshwater invertebrates.* EPA 600-R24-024, Duluth, MN

_____. 1996. *Calculation and evaluation of sediment effect concentrations for the amphipod Hyalella azteca and the midge Chironomus riparius.* EPA 905-R96-008, Chicago, IL.

Wildhaber, M.L., and Schmitt, C.J. 1996. Hazard ranking of contaminated sediments based on chemical analysis, laboratory toxicity tests, and benthic community composition: Prioritizing sites for remedial action. *J. Great Lakes Res.* 22:639-652.

Submitted: 15 April 1996
Accepted: 13 July 1996