117606

Estuaries Vol. 22, No. 3A, p. 624–635 September 1999

Refinement, Validation, and Application of a Benthic Condition Index for Northern Gulf of Mexico Estuaries

VIRGINIA D. ENGLE[1]
United States Geological Survey, Biological Resources Division
Gulf Breeze Project Office
1 Sabine Island Drive
Gulf Breeze, Florida 32561-5299

J. KEVIN SUMMERS
United States Environmental Protection Agency
National Health and Environmental Effects Research Laboratory, Gulf Ecology Division
1 Sabine Island Drive
Gulf Breeze, Florida 32561-5299

ABSTRACT: By applying discriminant analysis to benthic macroinvertebrate data, we have developed an indicator of benthic condition for northern Gulf of Mexico estuaries. The data used were collected by the United States Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP) in the Louisianian Province from 1991 to 1994. This benthic index represents a linear combination of the following weighted parameters: the proportion of expected species diversity, the mean abundance of tubificid oligochaetes, the percent of total abundance represented by capitellid polychaetes, the percent of total abundance represented by bivalve mollusks, and the percent of total abundance represented by amphipods. We successfully validated and retrospectively applied the benthic index to all of the benthic data collected by EMAP in the Louisianian Province. This benthic index was also calculated for independent data collected from Pensacola Bay, Florida, in order to demonstrate its flexibility and applicability to different estuarine systems within the same biogeographic region. The benthic index is a useful and valid indicator of estuarine condition that is intended to provide environmental managers with a simple tool for assessing the health of benthic macroinvertebrate communities.

Introduction

Information gained from monitoring benthic macroinvertebrate communities has been used widely to measure the status of and trends in the ecological condition of estuaries. Benthic macroinvertebrates are good indicators of estuarine condition because they are relatively sedentary within the sediment-water interface and deeper sediments (Dauer et al. 1987). Short-term disturbances such as hypoxia and long-term disturbances such as accumulation of sediment contaminants affect the population and community dynamics of benthic macroinvertebrates (Rosenberg 1977; Harper et al. 1981; Rygg 1986). Many of the effects of such disturbances include changes in benthic diversity, ratio of long-lived to short-lived species, biomass, abundance of opportunistic or pollution-tolerant organisms, and the trophic or functional structure of the community (Pearson and Rosenberg 1978; Santos and Simon 1980; Gaston 1985; Warwick 1986; Gaston and Nasci 1988; Gaston and Young 1992).

Researchers have successfully developed multimetric indices that combine the various effects of natural and anthropogenic disturbances on benthic communities. One such index, the benthic index of biotic integrity (B-IBI), was initially developed for freshwater systems (Clements et al. 1992; Lenat 1993; Kerans and Karr 1994; Chessman 1995; Lang and Reymond 1995), but variations have been successfully applied to estuaries (Engle et al. 1994; Ranasinghe et al. 1994; Weisberg et al. 1997). Multimetric benthic indices can help environmental managers who require a standardized means of tracking the ecological condition of estuaries. Environmental managers and policymakers also desire an easy, manageable method of identifying the extent of potentially degraded areas and a means of associating biotic responses with environmental exposures (Summers et al. 1995). For an indicator to be appropriate for the assessment of estuarine health, it should incorporate geographic variation and should recognize the in-

[1] Corresponding author; current address: United States Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Gulf Ecology Division, 1 Sabine Island Drive, Gulf Breeze, Florida 32561-5299; tele: 850/934-9354; fax: 850/934-4203; e-mail: engle.virginia@epa.gov.

herent multivariate nature of estuarine systems (Karr 1993; Wilson and Jeffrey 1994). While the statistical methods used to develop indicators may often be complex, it is the end product, an index of condition, that is of interest to resource managers. By applying a mathematical formula to multivariate benthic data, resource managers can calculate a scaled index that can be used to evaluate the benthic condition of estuaries in their region. Although biotic indices have been accused of oversimplifying or overgeneralizing biological processes, they play an important role in resource management (i.e., to provide criteria with which to characterize a resource as impaired or healthy) (Rakocinski et al. 1997). While ecological indicators such as biotic indices were developed to serve as tools for the preliminary assessment of ecological condition, they are not intended to replace a complete analysis of benthic community dynamics.

We originally developed a benthic index that characterized the environmental quality of estuaries in the northern Gulf of Mexico by summarizing the composition and diversity of benthic infaunal invertebrate communities using data from the 1991 demonstration project of the United States Environmental Protection Agency's (USEPA) Environmental Monitoring and Assessment Program (EMAP-E) in the Louisianian Province estuaries (Engle et al. 1994). That benthic index combined the Shannon-Wiener index (H', adjusted for salinity) and the percentages of total abundance represented by tubificid oligochaetes and bivalves. Although this index was intended to be preliminary, because data were not yet available for validation, it did successfully discriminate reference sites from sites degraded with respect to sediment contaminants, sediment toxicity, and hypoxia (Engle et al. 1994).

After the second year (1992) of EMAP-E sampling in the Louisianian Province, the benthic index was applied to an independent set of data for the purposes of validation and further testing. Benthic index values for this new dataset ranged from −11 to 11. The large negative values occurred at stations in the Mississippi River, indicating these sites had benthic conditions that were substantially more degraded than the original test sites used to develop the index. The Mississippi River sites had an average of 13 contaminants with high concentrations in 1992 and an average of four contaminants with high concentrations in 1991, while the 1991 degraded test sites ranged from 3 to 12 contaminants with high concentrations (Summers et al. 1992; Macauley et al. 1994). A new, revised benthic index was developed to include test sites from 1991 and 1992 that represented a broader set of environmental conditions. The revised benthic index was validated successfully using an independent subset of data from 1993 and 1994 and was applied retrospectively to all of the data collected from Gulf of Mexico estuaries during 1991–1994. This paper presents the process of developing the revised benthic index, its validation, and instructions for its application to independent data using Pensacola Bay, Florida, as an example.

Field Methods

A total of 341 stations located throughout the estuaries of the Gulf of Mexico was sampled as part of EMAP-E during July–September of 1991 and 1992 (Summers et al. 1991, 1992; Macauley et al. 1994). These estuaries were located geographically between Anclote Key, Florida, and the Rio Grande, Texas. Probability-based sites (239) were randomly located in large estuaries (≥ 250 km^2), small estuaries (< 250 km^2), and the Mississippi River (from New Orleans to its confluence with the Gulf of Mexico) (Summers et al. 1995). The remaining sites were specifically selected to answer sample design questions, perform quality control, and provide within-year and between-year replications.

At each site, three replicate benthic samples were collected using a Young-modified Van Veen grab that sampled a surface area of 440 cm^2. A small core (60 cc) was taken from each grab for sediment grain-size distribution, total organic carbon, and percent silt-clay. Fauna retained on a 0.5-mm screen were identified to the lowest practical taxonomic level and counted. Samples were also collected at each site for analysis of sediment chemistry and toxicity and water quality (Summers et al. 1991, 1992; Macauley et al. 1994).

Analysis Methods

The general approach for the development of a benthic index for EMAP-E was presented by Weisberg et al. (1993) and Engle et al. (1994). In developing a revised benthic index, new test sites were chosen that consisted of a subset of sites from 1991 and 1992 with known degraded or reference conditions. Designation as degraded or reference was based on guidelines for environmental indicators that would adversely affect benthos (e.g., sediment contaminant concentrations, degree of hypoxia, and survival of toxicity test organisms). Long et al.'s (1995) guidelines for sediment contaminants were used to differentiate reference and degraded sites. These guidelines were developed from a biological effects database (BEDS) that contained the concentrations of contaminants at which adverse biological effects occurred (i.e., altered benthic communities, sediment toxicity, and histopathological disorders in demersal fish). The guidelines, referred to as effects range-low (ER-L)

TABLE 1. Distribution of the degraded and reference test sites by salinity, sediment type, and state.

	Number of Degraded Sites	Number of Reference Sites
Salinity		
Tidal-freshwater (0‰)	8	2
Oligohaline (> 0–5‰)	2	3
Mesohaline (> 5–18‰)	4	5
Polyhaline (> 18–35‰)	7	11
Marine (> 35‰)	1	3
Sediment type		
Mud (> 80% Silt-Clay)	17	10
Mud-Sand (20–80% Silt-Clay)	5	10
Sand (<20% Silt-Clay)	0	4
State		
Florida	4	2
Alabama	3	3
Mississippi	2	2
Louisiana	12	10
Texas	1	7

and effects range-median (ER-M), delineate concentrations of contaminants at which adverse biological effects occur rarely (< ER-L), occasionally (> ER-L and < ER-M), or frequently (> ER-M).

Reference test sites met all of the following conditions: the minimum dissolved oxygen value over a 24-h period was greater than 3.0 mg l^{-1}, sediment contaminant concentrations never exceeded the ER-M concentration nor did more than three contaminants have concentrations exceeding ER-L values, and the control-corrected percent survival for *Ampelisca abdita* (10 d) and *Mysidopsis bahia* (96 h) in acute sediment bioassays was > 85%. Degraded test sites exhibited the impacts of low dissolved oxygen stress, contaminated sediment stress, or sediment toxicity and met at least one of the following conditions: the minimum dissolved oxygen measurement over a 24-h period was ≤ 2.0 mg l^{-1} (Summers and Engle 1993), sediment concentrations for at least one contaminant exceeded the ER-M value or concentrations of at least four contaminants exceeded the ER-L value, or control-corrected survival of either *A. abdita* or *M. bahia* was < 80%.

The test sites were chosen from the original sampled sites not only to represent extreme reference and degraded environmental conditions but also to cover the expected range of salinity, sediment types, and biogeographical divisions found in the estuaries of the northern Gulf of Mexico (Engle et al. 1994). The sites ranged in salinity regimes from tidal-freshwater (0‰) to marine (> 35‰), and most sites were located in muddy (> 80% silt-clay) sediment (Table 1). The location of the majority of sites in Louisiana is an artifact of the EMAP-E probability-based sample design (Louisiana has proportionately more estuarine area than the other four Gulf states).

TABLE 2. List of candidate benthic measures used to develop the benthic index.

Measures of species richness or diversity
Mean number of benthic species
Mean Shannon-Wiener diversity index (H')
Mean number of polychaete species
Measures of abundance
Mean abundance, total abundance, of all benthic organisms
Measures of taxonomic composition
Mean abundance, percent of total abundance, as amphipods
Mean abundance, percent of total abundance, as bivalves
Mean abundance, percent of total abundance, as capitellids
Mean abundance, percent of total abundance, as decapods
Mean abundance, percent of total abundance, as gastropods
Mean abundance, percent of total abundance, as molluscs
Mean abundance, percent of total abundance, as polychaetes
Mean abundance, percent of total abundance, as spionids
Mean abundance, percent of total abundance, as tubificids
Measures of trophic level composition
Mean abundance, percent of total abundance, as deposit feeders
Mean abundance, percent of total abundance, as omnivores and carnivores
Mean abundance, percent of total abundance, as suspension feeders

The benthic abundance and species-composition data collected by EMAP-E during 1991–1992 were used to calculate 29 candidate benthic measures (Table 2) that were used to develop the benthic index. These measures were chosen to represent a wide range of characteristics of benthic communities, including species richness and diversity, taxonomic composition, and trophic level abundance (Engle et al. 1994). Most of the measures were calculated as an arithmetic mean over three replicate samples. For example, the Shannon-Wiener Index ($H' = -\Sigma p_i \log_2 p_i$) was computed separately for each benthic grab sample, and the calculations were then averaged to result in a mean diversity index for a site. In addition to mean abundance measures, the proportion of the total abundance that was made up of different taxonomic groups was calculated. Proportional parameters were adjusted using an arc-sine transformation, and abundance parameters were adjusted using a $\log_{10}$(value + 1) transformation. Benthic species richness and diversity are often associated with natural habitat gradients such as salinity and sediment type (Flint and Younk 1983; Dauer et al. 1987; Montagna and Kalke 1992; Dauer 1993; Berger et al. 1995). Unless the natural variation is removed, however, these indicators may not reflect the responses of benthic communities to anthropogenic stress. Therefore, Pearson correlations were performed to test for significant relationships between all candidate indicators and habitat measures, including salinity, longitude, percent silt-clay, and total organic carbon content of sediments. Although many of the correlations were statistically significant at $p < 0.05$, only three had r^2 values exceeding 0.25 (mean Shannon-Wiener Diversity Index (H'), mean number of polychaete species, and mean number of benthic species).

These measures of diversity and species richness were adjusted to remove the effects of salinity. The adjustment was based on a modification of the method described by Engle et al. (1994), which estimates the expected value of the number of species at any given salinity and calculates the percent deviation of each observed value from the expected value. Salinity-adjusted candidate measures (proportion of expected diversity, proportion of expected mean number of species, and proportion of expected mean number of polychaete species) were computed by dividing the observed value by the expected value for each measure. Although correlations between the new measures and salinity were statistically significant ($p < 0.05$), the r^2 was very low (2.6% for the proportion of expected diversity; 9.6% for the proportion of expected number of polychaete species; 3.7% for the proportion of expected number of species), and the relationships between the new measures and salinity were deemed insignificant for practical purposes. The salinity-adjusted parameters were substituted for the original, unadjusted variables in the list of candidate measures.

A series of stepwise and canonical discriminant analyses was applied to the test dataset in order to select a subset of candidate biological indicators that best discriminated between the reference and degraded sites and to determine which linear combination of these variables showed the most substantial difference between degraded and reference sites (Engle et al. 1994). An estimate of classification efficiency determined how well the model classified the known test sites as degraded or reference. A benthic index was calculated by applying the coefficients derived from canonical discriminant analysis to the original variables and then normalizing the index values to a range of 0 to 10 for ease of interpretation.

The benthic index was calculated for all sites sampled in the Louisianian Province from 1991 to 1994. Validation was completed by evaluating the classification of an independent set of degraded and reference sites, by comparing the distribution of benthic index values from all years, and by comparing the classification of reference or degraded sites that were replicated within a year. The benthic index was also calculated for sites within Pensacola Bay, Florida, in order to demonstrate its flexibility and applicability to different estuarine systems within the same biogeographic region. The sampling program for Pensacola Bay employed sample designs and methods identical to those of EMAP-E in the Louisianian Province.

Results

Discriminant analysis is often a process of iteration. Stepwise discriminant analysis suggests parameters that best discriminate between groups of observations (STEPDISC procedure: SAS Institute 1990). Canonical discriminant analysis tests the classification efficiency of the model provided by stepwise discriminant analysis as well as providing estimated coefficients that can be used to calculate discriminant scores (CANDISC procedure: SAS Institute 1990). If the classification efficiency is not satisfactory, misclassified stations may be removed from the data or parameters that prove to be colinear may be omitted. The series of iterations of stepwise and then canonical discriminant analyses is performed until satisfactory results are achieved (Engle et al. 1994).

The final series of discriminant analyses was performed on a test dataset of 22 degraded sites and 24 reference sites, specifying an equal prior probability of being classified into either group. Stepwise discriminant analysis suggested seven candidate measures (Proportion of expected diversity, Tubificid abundance, Percent amphipods, Percent capitellids, Percent bivalves, Total abundance, Proportion of expected number of polychaete species) were required to discriminate between degraded and reference sites. Seventy-nine percent of the variance in the model was explained by these measures (Table 3). Pearson correlations were calculated for these measures to determine if any redundancy occurred among variables. The first five variables listed in Table 3 showed no significant correlations with each other ($p > 0.05$). The remaining two variables, however, correlated significantly with the Proportion of expected diversity ($p \leq 0.001$ or $p \leq 0.0001$). We concluded that Proportion of expected number of polychaete species and Total abundance were redundant and contributed little to the overall model. The first five measures (i.e., proportion of expected Shannon-Wiener Diversity Index (H'), mean abundance of tubificid oligochaetes, percent of total abundance represented by capitellid polychaetes, percent of total abundance represented by bivalves, and the percent of total abundance represented by amphipods), which accounted for 75% of the variation (Table 3), were included in the final canonical discriminant analysis.

Canonical discriminant analysis was performed using these five variables to confirm their partial contributions to the variance structure and to determine the classification efficiency and appropri-

TABLE 3. Results of stepwise discriminant analysis used to determine components that best discriminate between degraded and reference test sites.

Variable Entered	Partial R^2	F Statistic	Probability > F	Average Squared Canonical Correlation
Proportion of expected diversity	0.4452	35.303	0.0001	0.4452
Tubificid abundance	0.2915	17.696	0.0001	0.6069
Percent amphipods	0.1949	10.171	0.0027	0.6836
Percent capitellids	0.0906	4.086	0.0498	0.7122
Percent bivalves	0.1198	5.444	0.0248	0.7467
Total abundance	0.935	4.023	0.0519	0.7704
Proportion of expected number of polychaete species	0.1210	5.233	0.0278	0.7982

ate coefficients of the model (Table 4). One of our criteria for acceptance of a model was that the error rate for misclassification of sites by the discriminant criterion had to be less than 10%. Using this model, 9.1% of degraded sites and 4.2% of reference sites were misclassified, resulting in an overall successful classification rate of 93.4%. We accepted this model as our final model for calculating the benthic index.

The first step in computing the benthic index was to standardize the values of the five parameters in the model. All of the parameters were normalized to a mean = 0 and standard deviation = 1 using the mean and standard deviation of values from the EMAP-E probability-based sites in each year. The standardized parameters were combined to make a composite benthic index by applying the standardized coefficients produced by the canonical discriminant analysis (Table 4) in the following algorithm:

Discriminant Score

$$\begin{aligned} &= (1.5710 \times \text{Proportion of Expected Diversity}) \\ &\quad + (-1.0335 \times \text{Mean Abundance of Tubificids}) \\ &\quad + (-0.5607 \times \text{Percent Capitellids}) \\ &\quad + (-0.4470 \times \text{Percent Bivalves}) \\ &\quad + (0.5023 \times \text{Percent Amphipods}). \end{aligned}$$

TABLE 4. Results of canonical discriminant analysis used to determine coefficients for components of the benthic index.

Canonical discriminant analysis statistics:	
Squared canonical correlation	0.7467
Likelihood ratio	0.2533
Approximate F	23.5840
Probability > F	0.0001
Correct classification:	
Degraded sites	90.91%
Reference sites	95.83%
Total-sample standardized canonical coefficients:	
Proportion of expected diversity	1.5710
Tubificid abundance	−1.0335
Percent amphipods	0.5023
Percent capitellids	−0.5607
Percent bivalves	−0.4470

Discriminant scores were computed for the test dataset and then for all sample sites. The discriminant scores developed for the test dataset, which ranged from −3.21 to 4.29 were normalized to a range of 0 to 10 for ease in interpretation and presentation.

The index values calculated for the test dataset were then compared to the classification of sites as degraded or reference according to the discriminant analysis. The comparison indicated index values ≤ 3.0 represented degraded sites, index values > 5.0 represented reference sites, and index values between 3.0 and 5.0 represented sites with undetermined classification (these sites were misclassified by the original analysis). The distribution of the test sites (Fig. 1) shows the overlap of degraded or reference classification by the benthic index with the symbols indicating the original classification of test sites using a priori criteria. The relationships between the benthic index and each of its components for the test sites is shown in Fig. 2.

The benthic index was significantly correlated with salinity and percent silt-clay content of sediments, but the r for both of these correlations was < 0.39 (15% of variance explained). We determined these relationships were insignificant from an ecological perspective and that statistical signif-

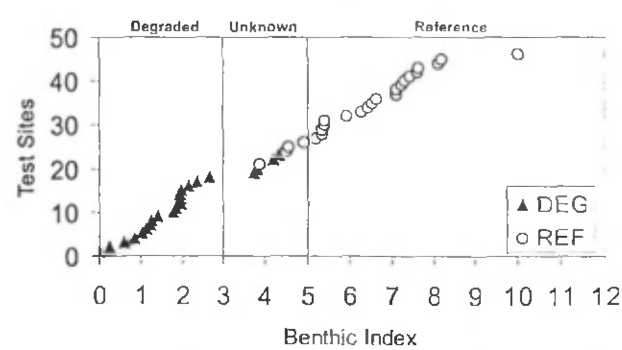

Fig. 1. Distribution of benthic index values for the test sites where ▲ = degraded sites and ○ = references sites determined by a priori criteria for dissolved oxygen, sediment chemistry, and sediment toxicity. Based on the benthic index, 3.0 is the cutoff value for degraded sites and 5.0 is the cutoff value for reference sites.

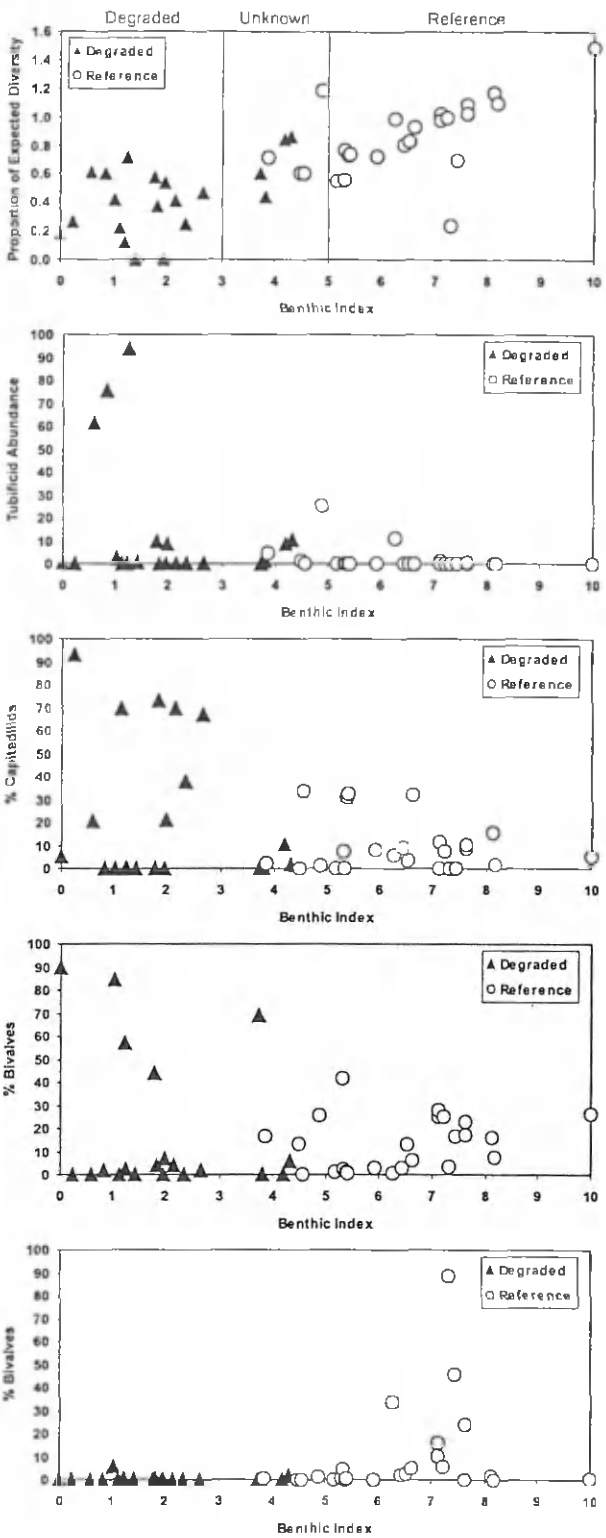

Fig. 2. Relationships between the benthic index and (a) proportion of expected diversity, (b) tubificid abundance, (c) percent capitellids, (d) percent bivalves, and (e) percent amphipods for the test sites where ▲ = degraded sites and ○ = references sites determined by a priori criteria for dissolved oxygen, sediment chemistry, and sediment toxicity.

icance was driven primarily by the large number of samples ($n = 338$). Anthropogenic impacts may be correlated with salinity and silt-clay as well, therefore, residual correlations between the benthic index and salinity or silt-clay may not indicate a lack of discriminatory power in the index. This is important because the benthic index was designed to be an indicator of environmental condition that is representative of the degree of sediment contamination and hypoxia experienced at a site, regardless of the inherent salinity and sediment characteristics.

Validation

The benthic index was computed for all sites sampled by EMAP-E in the Louisianian Province from 1991 to 1994. Because the index was developed from a subset of sites sampled in 1991 and 1992, validation of the benthic index was accomplished by using an independent set of data from two subsequent years, 1993 and 1994, as well as data from special study sites representing between-year and within-year replicates. Validation of the benthic index consisted of three steps: assessment of the correct classification by the index of an independent set of degraded and reference sites, comparison of the cumulative distribution function of the index among four years, and assessment of correct classification of replicate sites by the index.

Sites from 1993 and 1994 were classified as degraded or reference based on the criteria for dissolved oxygen, sediment chemistry, and sediment toxicity used to choose test sites in the development of the index. Of the 310 sites sampled in 1993 and 1994, only 195 could be classified as either degraded or reference and these were used in the first validation step. We randomly chose 50 subsets of the 195 sites where each subset consisted of 50 degraded and 50 reference sites. Correct classification occurred when the benthic index was either ≤ 3 at degraded sites or ≥ 5 at reference sites. Misclassification occurred when the benthic index was ≤ 3 at reference sites (false negative) or ≥ 5 at degraded sites (false positive). Using the 50 trials, we determined the percent of sites correctly classified as degraded and reference by the benthic index. The benthic index correctly classified 66–82% of degraded sites ($\bar{x} = 74\%$; SE = 0.5) and 70–84% of reference sites ($\bar{x} = 77\%$; SE = 0.4). The classification efficiency was adequate but not as high as we had hoped. This was partly due to the high degree of variability in the benthic communities in the Gulf of Mexico region. We may have sacrificed a level of precision in favor of a generalized index that is applicable across a wide geographic area with an inherently large spatial

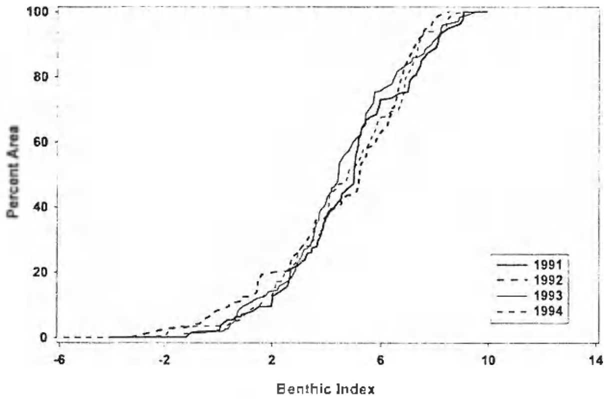

Fig. 3. Comparison of the cumulative distribution functions (CDF) of the benthic index calculated for the probabilistic sampling sites of the Louisianian Province for the years 1991–1994.

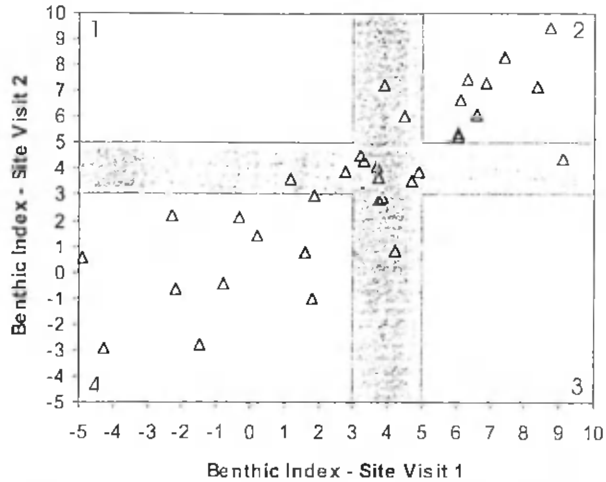

Fig. 4. Comparison of benthic index values from replicate visits within a sampling season to a site. Quadrant 2 indicates sites classified as reference for both visits; quadrant 4 indicates sites classified as degraded for both visits; quadrants 1 and 3 indicate sites that were classified differently for both visits. Sites that fall within the gray shaded area changed classification from degraded or reference to unknown (or vice versa) from visit 1 to visit 2.

variation. We also investigated the kappa coefficient (κ) to measure the degree of agreement (Stokes et al. 1995) between classification of a site by the benthic index versus classification by sediment chemistry, toxicity, and dissolved oxygen. The average kappa coefficient for the 50 trials was 0.509 ($\kappa \geq 0.4$ indicates moderate agreement) and the null hypothesis that there was no agreement (H_0: $\kappa = 0$) was rejected at the $\alpha = 0.05$ level of significance.

EMAP-E's probabilistic sample design facilitated the direct comparison of data among years within a province. Because the cycle for EMAP-E sampling of estuaries requires four years to complete (Summers et al. 1991), we made the assumption that there should be no statistical differences in the cumulative distribution of values of the benthic index (or any other indicator) among the four years. Any statistical difference in the distributions would prompt us to look for sampling error, changes in laboratory methods, measurement error, severe weather patterns, or pollution events that could have affected estuarine conditions in a given year, although a deviation from expectation could also reflect a real change in the benthic community and indicate that the index was performing correctly. A cumulative distribution function (CDF) was computed for the benthic index values weighted by the surface area (km^2) represented by the base stations for each year of sampling. The CDFs for the benthic index among the four years were not statistically different (Fig. 3) based on the Kolmogorov-Smirnov test and using the Bonferroni inequality adjustment to ensure that the overall experiment-wise error rate was at the desired five percent level of significance.

Within each year (excluding 1991), 13 estuaries were visited more than once in order to evaluate spatial and temporal replication. These sites were used to validate the consistency of classification by the benthic index. The same classification by the benthic index should be given to a single site on replicate visits within a sampling season. The distribution of benthic index values between the first and second visits to a site within the same sampling period was compared (Fig. 4). The shaded areas indicate the intersection of the cutoff values of 3.0 for degraded sites and 5.0 for reference sites. Ideally, all of the points should fall in quadrants 2 and 4 where sites were classified as degraded on both visits or as reference on both visits. Although several points fell within the shaded area, indicating that the site classification changed from degraded or reference to unknown (or vice versa) from the first visit to the second visit, no sites fell within quadrants 1 and 3, indicating that no sites were misclassified as degraded or reference. Correlation between the benthic index from the temporal replicates was 0.83. This validation of the benthic index was also determined to be successful.

Pensacola Bay, Florida

We applied our benthic index to independent data from Pensacola Bay, Florida, to illustrate how resource managers may use this benthic index to assess the biotic condition of an estuary in the Gulf of Mexico. The USEPA's Environmental Research Laboratory at Gulf Breeze, Florida (currently, the Gulf Ecology Division of the National Health and Environmental Effects Laboratory), conducted a survey in April 1992 of the water quality, sediment,

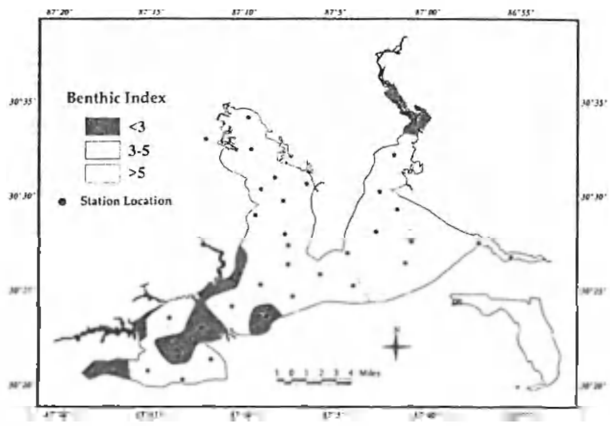

Fig. 5. Map of United States Environmental Protection Agency sampling stations for April 1992 in Pensacola Bay, Florida, with spatial distribution of conditions based on the benthic index (modified from Engle and Summers 1998).

and biological status of Pensacola Bay using a sample design and methodology similar to that of EMAP-E in the Louisianian Province. Using a probabilistic sample design, 40 sites were located throughout the main bay and the tributaries and bayous (Fig. 5). Benthic macroinvertebrate, sediment chemistry, and water quality samples were collected according to EMAP-E sampling protocols. The benthic index was calculated for each site by applying the formulas and methods detailed in the Appendix. The index was then compared to existing sediment chemistry information to identify any associations between degraded benthic communities and sediment contaminants.

Assessing the ecological status of Pensacola Bay is problematic because most of the bay exhibits signs of degraded sediment quality. By applying the same categorization guidelines for sediment chemistry and sediment toxicity (DO was inapplicable because sampling occurred in April) to the Pensacola Bay sites, we classified 36 (90%) of the sites as degraded. Severely contaminated sediments occur throughout the bay, with as many as 44 ($\bar{x}$ = 12) contaminants at a single site having concentrations exceeding ER-L values. Of interest to a resource manager who is concerned with biological quality is whether or not the benthic community is degraded throughout the bay or whether the benthic community reflects some gradient from severely degraded to moderately degraded condition. The benthic index identified 10 degraded sites, 16 moderately degraded sites, and 14 reference sites. Although sediment contaminants occurred at high concentrations at all sites, the sites with the highest contaminant concentrations also had the lowest benthic index values.

Benthic index values were significantly correlated ($p < 0.005$) with the number of contaminants having concentrations greater than ER-L values. Stepwise regression analysis indicated that 65% of the variation in the benthic index could be accounted for by lead, cadmium, tributyltin, DDT, the number of enriched metals, and the number of contaminants with concentrations greater than ER-L guidelines.

Discussion

Indices of biotic integrity (IBIs) have been used commonly to assess the ecological health of freshwater environments (Karr 1981; Kerans and Karr 1994). The goal of these indices has been to provide a simple, easily understood method of comparing the biological status of different streams or of tracking such status over time. Many freshwater IBIs were developed to identify streams that have been impacted by examining the responses of the benthic or fish communities to known anthropogenic influences. A biotic index that is applicable to estuarine environments would incorporate some of the same theories and methods as those used for freshwater biotic indices but would need to address the natural variability inherent in estuarine systems (e.g., salinity regimes) (Deegan et al. 1997; Weisberg et al. 1997).

Although the absolute values of an individual biotic index may be specific to a given ecoregion (Lenat 1993), often the theories and methods used in development are applicable across systems. A universal index that works in all systems or even in systems of the same ecological type, however, is unrealistic because communities are complex and geographically diverse (Lang et al. 1989). Historically, indices of biotic integrity have been separated into two types: fish indices based on ranks of community parameters such as number of species and abundance of trophic groups (Karr 1981; Bramblett and Fausch 1991; Oberdorff and Hughes 1992; Osborne et al. 1992; Lyons et al. 1995) and benthic invertebrate indices based on quantitative ranks assigned to species according to their known tolerance to anthropogenic stress (Lang et al. 1989; Clements et al. 1992; Lenat 1993; Chessman 1995; Lang and Reymond 1995). Use of freshwater biotic indices has been based on the assumption that some known anthropogenic stress (i.e., pollution) was the main factor affecting the abundance and community composition of benthic invertebrates (Modde and Drewes 1990). One of the difficulties in developing a biotic index for estuarine systems is that the tolerance values of many benthic invertebrates to heavy metals and other pollutants is unknown, whereas this information has been well documented for freshwater systems (Clements et al. 1992; Weisberg et al. 1997).

The freshwater indices and recent estuarine attempts to develop an IBI-type indicator show a basic difference from the approach described here. The indices developed by Weisberg et al. (1997) and Deegan et al. (1997) are based on the previous index work of Karr (1981). The original approach treated a set of parameters as a priori related to fish community structure and developed the strength of those relationships based on single stressor dose-response regressions. It is through dose-response relationships that Karr (1981) determined the breakpoints for his semi-quantitative multimetric IBI for freshwater fish communities. Weisberg et al. (1997) applied Karr's (1981) IBI approach to the benthic communities of Chesapeake Bay, creating a benthic IBI, or B-IBI, that used similar metrics, with the exception that they did not apply single stressor relationships to determine the threshold values for the metrics. Their use of reference (non-stressed) sites to assign scoring criteria to the metric values disregarded the strength of Karr's (1981) approach (i.e., the ability to interpret the components of a score with regard to a specific stressor). The reality that estuarine resources tend not to have single stressor attributes led our index development away from the Karr (1981) approach for the development of regional indices.

Weisberg et al.'s (1997) approach required significant subsetting of the data into seven separate benthic habitats. Few estuaries other than Chesapeake Bay and possibly Puget Sound (R. Llanso personal communication) have sufficient benthic information to permit such subsetting. Deegen et al.'s (1997) approach for an estuarine fish IBI also requires the development of separate indices for multiple habitats.

Like Karr's (1981) IBI, the Chesapeake Bay model must be redeveloped for each application. Weisberg et al.'s (1997) B-IBI for Chesapeake Bay showed a 90–95% fidelity rate. The strength of the model fit is appealing, but it must be realized that the model was built specifically for Chesapeake Bay, incorporating only the variability observed within that system. Any model construction, whether simulation, regression, or semi-quantitative multimetric index, can optimize on only two of the following model attributes: generality, reality, or precision. The Chesapeake Bay B-IBI optimizes on reality and precision. Attempts to apply the estuarine B-IBI to other systems geographically close to Chesapeake Bay (i.e., Delaware Bay, Long Island Sound, and the Hudson-Raritan estuary) were not successful, with fidelities of < 50% (S. Weisberg personal communication). This does not mean a B-IBI cannot be developed for these systems, but each application must be specifically developed for the system in question; thus, reparameterizing the model for each estuary. Development of a B-IBI for each estuary creates a strong, valid model for each system that is highly suspect with regard to comparisons across systems at scales greater than those captured by a single estuarine system.

While we adhere to the same principles in the initial selection of the sites used in index construction, we do not force the index results to demonstrate these principles through subjectively scaled index elements. We recommend scaling the factors appropriately based on the configuration of the data. The end result of our index development is that, although the components of the index were determined empirically, a case can be made for each to follow the paradigms of traditional benthic community analysis (e.g., diversity; see discussion in Engle et al. 1994). The contribution of each component to the index, however, is determined by the data. We believe this process adds a necessary level of independence to the determination of benthic condition.

We have attempted to develop an index that would be applicable across the wide variety of estuarine environments in the Gulf of Mexico. The proposed index optimizes on reality and generality and, as a result, its precision decreases somewhat due to broad variability incorporated into the model. While the IBI-type approaches have been shown to work well in specific streams or when constructed for a specific estuarine system (Weisberg et al. 1997), the successful application of an IBI to a large biogeographic region has yet to be accomplished. The statistical approach described in this paper proved to be applicable throughout the estuaries of the northern Gulf of Mexico and to work as well in Florida as it does in Texas. Our benthic index was independently validated by Rakocinski et al. (1997), who compared the results of canonical correspondence analysis (CCA) with our benthic index, using data from EMAP-E from 1991 and 1992. They found a good cross-validation between the two methods in that the CCA distinctly separated sites with low benthic index scores from those with high scores. They also found that the sites that were degraded as determined by low benthic index scores coincided with the area of the CCA that was interpreted to have high concentrations of sediment contaminants.

Another difficulty in applying biotic index methods to estuarine systems is that, in many cases, it is not only anthropogenic stress but also the natural variability of estuarine habitats that directly affects benthic macroinvertebrate community structure. Using conventional diversity indices, it is often difficult to distinguish whether a change in community structure is a result of a sudden change in salinity or is due to anthropogenic effects like sediment contamination. Given these constraints, we believe our benthic index for Gulf of Mexico estuarine systems is successful. The components of our benthic index are not highly correlated with salinity (i.e., the Shannon-Wiener index was adjusted for salinity and the other components were not correlated with salinity originally). Our benthic index delineates at a level of 74–77% benthic communities that have characteristics similar to those found in known degraded areas from benthic communities that are similar to those found in known reference areas. The difference in benthic community structure indicated by our benthic index is more likely associated with anthropogenic stress than with natural habitat variability. An exception occurs, however, when a low benthic index is associated primarily with hypoxia. The EMAP-E sampling design provides no method of distinguishing hypoxia that is anthropogenic in origin from natural hypoxic events.

A major criticism of the use of biotic indices stems from the belief that environmental managers favor an index over a detailed examination of the vast amount of information available from benthic monitoring programs (Elliott 1994). It is true that one goal of large-scale monitoring programs (e.g., EMAP) is to develop methods of condensing complex sets of data into manageable numerical indices without compromising too much of the information available in the details (Messer et al. 1991). The value of an index lies in its applicability across large geographical areas and its ability to provide regional assessments of ecological condition. The information derived from an index of environmental condition is useful to environmental managers and policymakers and decisionmakers who want to identify areas of potential degradation and track the status of environmental condition over time. A benthic index can be used to answer questions about the health of benthic communities in the estuaries of a large geographical region, the spatial or temporal variation of degraded areas of benthic communities, and comparisons of the status of benthic ecological conditions between the estuaries of different regions.

For any index to be useful to resource managers, it must be easy to develop and apply. The IBIs modeled after Karr (1981) are intuitive in that the models are forced to incorporate the conceptual framework of the developer. The EMAP-E benthic index employs the same generalized approach but assumes multi-stressor relationships and depends solely on the data to delineate which benthic parameters relate to the observed situation. The IBIs are perceived to be easier to employ. Clearly, they may be easier to develop than the benthic index proposed here, but the scoring on multiple habitats of four to seven parameters is more involved than inserting five parameters into an equation. The normalization is for ease of interpretation and does not need to be done (i.e., if unnormalized the cut-off between poor condition and marginal condition is 0.0). Unlike the IBIs in general, the index proposed herein is applicable over a wide range of environmental conditions and geography and provides comparable scores over these gradients.

Acknowledgments

We would like to thank the EMAP-E sampling teams at Texas A&M University, Avanti Corporation, and Gulf Coast Research Laboratory for collecting all biological, physical, and chemical data used in our analyses. Special thanks go to R. Heard and his team at Gulf Coast Research Laboratory and A. McAllister at the University of Mississippi for the taxonomic identification and enumeration of the benthic samples. We also thank M. Bergen at Southern California Coastal Water Research Project for her insightful review of this manuscript. Contribution No. 988, United States Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Gulf Ecology Division, 1 Sabine Island Drive, Gulf Breeze, Florida 32561-5299.

Literature Cited

Berger, V. Y., A. D. Naumov, and A. I. Babkov. 1995. The relationship of abundance and diversity of marine benthos to environmental salinity. *Russian Journal of Marine Biology* 21:41–46.

Bramblett, R. G. and K. D. Fausch. 1991. Variable fish communities and the index of biotic integrity in a western Great Plains river. *Transactions of the American Fisheries Society* 120: 752–769.

Chessman, B. C. 1995. Rapid assessment of rivers using macroinvertebrates: A procedure based on habitat-specific sampling, family level identification, and a biotic index. *Australian Journal of Ecology* 20:122–129.

Clements, W. H., D. S. Cherry, and J. H. Van Hassel. 1992. Assessment of the impact of heavy metals on benthic communities at the Clinch River (Virginia): Evaluation of an index of community sensitivity. *Canadian Journal of Fisheries and Aquatic Sciences* 49:1686–1694.

Dauer, D. M. 1993. Biological criteria, environmental health and estuarine macrobenthic community structure. *Marine Pollution Bulletin* 26:249–257.

Dauer, D. M., R. M. Ewing, and A. J. Rodi. 1987. Macrobenthic distribution within the sediment along an estuarine salinity gradient. *Internationale Revue der Gesamten Hydrobiologie* 72:529–538.

Deegan, L. A., J. T. Finn, S. G. Ayvazian, C. A. Ryder-Kieffer, and J. Buonaccorsi. 1997. Development and validation of an estuarine biotic integrity index. *Estuaries* 20:601–617.

Elliott, M. 1994. The analysis of macrobenthic community data. *Marine Pollution Bulletin* 28:62–64.

Engle, V. D. and J. K. Summers. 1998. Determining the causes of benthic condition. *Environmental Monitoring and Assessment* 51:381–397.

Engle, V. D., J. K. Summers, and G. R. Gaston. 1994. A benthic index of environmental condition of Gulf of Mexico estuaries. *Estuaries* 17:372–384.

Flint, R. W. and J. A. Younk. 1983. Estuarine benthos: Long-term community structure variations, Corpus Christi Bay, Texas. *Estuaries* 6:126–141.

GASTON, G. R. 1985. Effects of hypoxia on macrobenthos of the inner shelf off Cameron, Louisiana. *Estuarine, Coastal and Shelf Science* 20:603–613.

GASTON, G. R. AND J. C. NASCI. 1988. Trophic structure of macrobenthic communities in the Calcasieu Estuary, Louisiana. *Estuaries* 11:201–211.

GASTON, G. R. AND J. C. YOUNG. 1992. Effects of contaminants on macrobenthic communities in the Upper Calcasieu Estuary, Louisiana. *Bulletin of Environmental Contamination and Toxicology* 49:922–928.

HARPER, JR., D. E., L. D. MCKINNEY, R. R. SALZER, AND R. J. CASE. 1981. The occurrence of hypoxic bottom water off the upper Texas coast and its effects on the benthic biota. *Contributions in Marine Science* 24:53–79.

KARR, J. R. 1981. Assessment of biotic integrity using fish communities. *Fisheries* 6:21–27.

KARR, J. R. 1993. Defining and assessing ecological integrity beyond water quality. *Environmental Toxicology and Chemistry* 12:1521–1531.

KERANS, B. L. AND J. R. KARR. 1994. A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. *Ecological Applications* 4:768–785.

LANG, C., G. L'EPLATTENIER, AND O. REYMOND. 1989. Water quality in rivers of western Switzerland: Application of an adaptable index based on benthic invertebrates. *Aquatic Sciences* 51:224–234.

LANG, C. AND O. REYMOND. 1995. An improved index of environmental quality for Swiss rivers based on benthic invertebrates. *Aquatic Sciences* 57:172–180.

LENAT, D. 1993. A biotic index for the southeastern United States: Derivation and list of tolerance values, with criteria for assigning water-quality ratings. *Journal of the North American Benthological Society* 12:279–290.

LONG, E. R., D. D. MACDONALD, S. L. SMITH, AND F. D. CALDER. 1995. Incidence of adverse biological effects within ranges of chemical concentrations in marine and estuarine sediments. *Environmental Management* 19:81–97.

LYONS, J., S. NAVARRO-PEREZ, P. A. COCHRAN, E. SANTANA C., AND M. GUZMAN-ARROYO. 1995. Index of biotic integrity based on fish assemblages for the conservation of streams and rivers in west-central Mexico. *Conservation Biology* 9:569–584.

MACAULEY, J. M., J. K. SUMMERS, P. T. HEITMULLER, V. D. ENGLE, G. T. BROOKS, M. BABIKOW, AND A. M. ADAMS. 1994. Annual Statistical Summary: EMAP-Estuaries Louisianian Province—1992. EPA/620/R-94/002. United States Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Gulf Breeze, Florida.

MESSER, J. J., R. A. LINTHURST, AND W. S. OVERTON. 1991. An EPA program for monitoring ecological status and trends. *Environmental Monitoring and Assessment* 17:67–78.

MODDE, T. AND H. G. DREWES. 1990. Comparison of biotic index values for invertebrate collections from natural and artificial substrates. *Freshwater Biology* 23:171–180.

MONTAGNA, P. A. AND R. D. KALKE. 1992. The effect of freshwater inflow on meiofaunal and macrofaunal populations in the Guadalupe and Nueces estuaries, Texas. *Estuaries* 15:307–326.

OBERDORFF, T. AND R. M. HUGHES. 1992. Modification of an index of biotic integrity based on fish assemblages to characterize rivers of the Seine Basin, France. *Hydrobiologia* 228:117–130.

OSBORNE, L. L., S. L. KOHLER, P. B. BAYLEY, D. M. DAY, W. A. BERTRAND, M. J. WILEY, AND R. SAUER. 1992. Influence of stream location in a drainage network on the index of biotic integrity. *Transactions of the American Fisheries Society* 121:635–643.

PEARSON, T. H. AND R. ROSENBERG. 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanography and Marine Biology Annual Review* 16:229–311.

RAKOCINSKI, C. F., S. S. BROWN, G. R. GASTON, R. W. HEARD, W. W. WALKER, AND J. K. SUMMERS. 1997. Macrobenthic responses to natural and contaminant-related gradients in northern Gulf of Mexico estuaries. *Ecological Applications* 7:1278–1298.

RANASINGHE, J. A., S. B. WEISBERG, J. B. FRITHSEN, D. M. DAUER, L. C. SCHAFFNER, AND R. J. DIAZ. 1994. Chesapeake Bay benthic community restoration goals. Report CBP/TRS 107/94. United States Environmental Protection Agency, Chesapeake Bay Program, Annapolis, Maryland.

ROSENBERG, R. 1977. Benthic macrofaunal dynamics, production, and dispersion in an oxygen-deficient estuary of West Sweden. *Journal of Experimental Marine Biology and Ecology* 26:107–133.

RYGG, B. 1986. Heavy-metal pollution and log-normal distribution of individuals among species in benthic communities. *Marine Pollution Bulletin* 17:31–36.

SANTOS, S. L. AND J. L. SIMON. 1980. Response of soft-bottom benthos to annual catastrophic disturbance in a south Florida estuary. *Marine Ecology Progress Series* 3:347–355.

SAS INSTITUTE, INC. 1990. SAS/STAT® User's Guide, Version 6, 4th edition, Volumes 1–2. SAS Institute, Inc., Cary, North Carolina.

STOKES, M. E., C. S. DAVIS, AND G. G. KOCK. 1995. Categorical Data Analysis Using the SAS System. SAS Institute Inc., Cary, North Carolina.

SUMMERS, J. K. AND V. D. ENGLE. 1993. Evaluation of sampling strategies to characterize dissolved oxygen conditions in Gulf of Mexico estuaries. *Environmental Monitoring and Assessment* 24:219–229.

SUMMERS, J. K., J. M. MACAULEY, AND P. T. HEITMULLER. 1991. Environmental Monitoring and Assessment Program. Implementation Plan for Monitoring the Estuarine Waters of the Louisianian Province—1991 Demonstration. EPA/600/5-91/228. United States Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Gulf Breeze, Florida.

SUMMERS, J. K., J. M. MACAULEY, P. T. HEITMULLER, V. D. ENGLE, A. M. ADAMS, AND G. T. BROOKS. 1992. Annual Statistical Summary: EMAP-Estuaries Louisianian Province—1991. EPA/600/R-93/001. United States Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Gulf Breeze, Florida.

SUMMERS, J. K., J. F. PAUL, AND A. ROBERTSON. 1995. Monitoring the ecological condition of estuaries in the United States. *Toxicological and Environmental Chemistry* 49:93–108.

WARWICK, R. M. 1986. A new method for detecting pollution effects on marine macrobenthic communities. *Marine Biology* 92:557–562.

WEISBERG, S. L., J. B. FRITHSEN, A. F. HOLLAND, J. F. PAUL, K. J. SCOTT, J. K. SUMMERS, H. T. WILSON, R. VALENTE, D. HEIMBUCH, J. GERRITSEN, S. C. SCHIMMEL, AND R. W. LATIMER. 1993. EMAP-Estuaries Virginian Province 1990 Demonstration Project Report. EPA/620/R-93/006. United States Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Narragansett, Rhode Island.

WEISBERG, S. B., J. A. RANASINGHE, D. M. DAUER, L. C. SCHAFFNER, R. J. DIAZ, AND J. B. FRITHSEN. 1997. An estuarine benthic index of biotic integrity (B-IBI) for Chesapeake Bay. *Estuaries* 20:149–158.

WILSON, J. G. AND D. W. JEFFREY. 1994. Benthic biological pollution indices in estuaries, p. 311–327. *In* K. J. M. Kramer (ed.), Biomonitoring of Coastal Waters and Estuaries. CRC Press, Boca Raton, Florida.

SOURCES OF UNPUBLISHED MATERIALS

Llanso, R. Personal communication. Washington State Department of Ecology, P. O. Box 47600, Olympia, Washington 98504-7600.

Weisberg, S. Personal communication. Southern California Coastal Water Research Project, 7171 Fenwick Lane, Westminster, California 92683-5218.

Appendix

Calculation of the benthic index for data from Pensacola Bay, Florida.

Step 1. Calculate arithmetic mean and standard deviation of the benthic variables using only the 40 base sites:

	Mean ($\bar{x}$)	Std. Dev. (s)
Proportion of expected Shannon-Wiener Diversity	0.6445	0.2959
Mean Abundance of Tubificids (log transformed)	0.0906	0.2052
Percent Capitellids (arc-sine transformed)	0.1752	0.2036
Percent Bivalves (arc-sine transformed)	0.0324	0.0608
Percent Amphipods (arc-sine transformed)	0.0334	0.0705

Step 2. Standardize variables to mean = 0 and standard deviation = 1 by applying the following formula to all of the data:

$$x_i' = \frac{S^*(x_i - \bar{x}) + M}{s_x} \tag{A2}$$

where

x_i' = new standardized value

S = desired standard deviation (1)

M = desired mean (0)

x_i = observation's original value

$\bar{x}$ = variable's mean

s_x = variable's standard deviation

Step 3. Calculate the discriminant score of the benthic index as follows using the standardized variables from Step 2:

Discriminant Score

= (1.5710 × Proportion of expected diversity)

+ (−1.0335 × Mean Abundance of tubificids)

+ (−0.5607 × Percent capitellids)

+ (−0.4470 × Percent bivalves)

+ (0.5023 × Percent amphipods).

Step 4. Calculate the benthic index by normalizing the discriminant scores from Step 3 using the following formula, which includes the minimum (−3.21) and range (7.50) of discriminant scores from the original test data used to develop the benthic index:

Benthic Index

= ((Discriminant Score − (−3.21))/7.50) × 10

Received for consideration, November 25, 1996
Accepted for publication, November 2, 1998