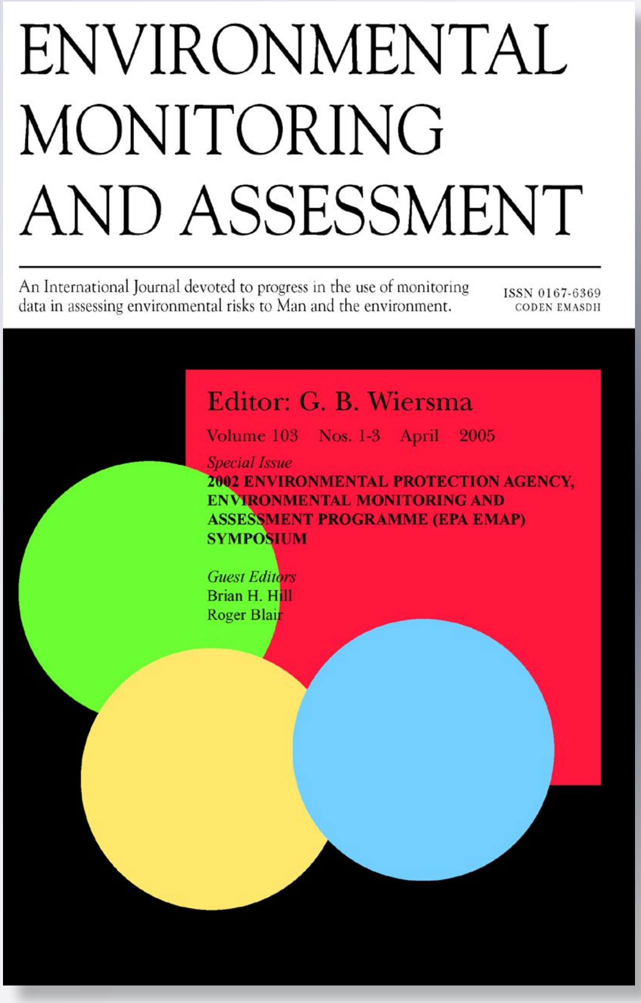
An approach for determining bioassessment performance and comparability

Jerry Diamond, James ~~R.~~ Stribling, Lisa Huff & Jaime Gilliam

Environmental Monitoring and Assessment
An International Journal Devoted to Progress in the Use of Monitoring Data in Assessing Environmental Risks to Man and the Environment

ISSN 0167-6369
Volume 184
Number 4

Environ Monit Assess (2012)
184:2247-2260
DOI 10.1007/s10661-011-2114-7



ENVIRONMENTAL MONITORING AND ASSESSMENT

An International Journal devoted to progress in the use of monitoring data in assessing environmental risks to Man and the environment.

ISSN 0167-6369
CODEN EMASDH

Editor: G. B. Wiersma

Volume 103 Nos. 1-3 April 2005

Special Issue
2002 ENVIRONMENTAL PROTECTION AGENCY, ENVIRONMENTAL MONITORING AND ASSESSMENT PROGRAMME (EPA EMAP) SYMPOSIUM

Guest Editors
Brian H. Hill
Roger Blair

Springer

Environ Monit Assess (2012) 184:2247–2260
DOI 10.1007/s10661-011-2114-7

An approach for determining bioassessment performance and comparability

Jerry Diamond · James ~~R.~~ Stribling · Lisa Huff · Jaime Gilliam

Received: 4 November 2010 / Accepted: 26 April 2011 / Published online: 25 May 2011

Abstract Many organizations in the USA collect aquatic bioassessment data using different sampling and analysis methods, most of which have unknown performance in terms of data quality produced. Thus, the comparability of bioassessments produced by different organizations is often unknown, ultimately affecting our ability to make comprehensive assessments on large spatial scales. We evaluated a pilot approach for determining bioassessment performance using macroinvertebrate data obtained from several states in the Southeastern USA. Performance measures evaluated included precision, sensitivity, and responsiveness to a human disturbance gradient, defined in terms of a land disturbance index value for each site, combined with a value for specific conductance, and instream habitat quality. A key finding of this study is the need to harmonize ecoregional reference conditions among states so as to yield more comparable and consistent bioassessment results. Our approach was also capable of identifying potential areas for refinement such as reevaluation of less precise, sensitive, or responsive metrics that may result in suboptimal index performance. Higher performing bioassessments can yield information beyond "impaired" versus "unimpaired" condition. Acknowledging the limitations of this pilot study, we would recommend that performance evaluations use at least 50 sites, 10 of which are ecoregional reference sites. Efforts should be made to obtain data from the entire human disturbance gradient in an ecoregion to improve statistical confidence in performance measures. Having too few sites will result in an under-representation of certain parts of the disturbance gradient (e.g., too few poor quality sites), which may bias sensitivity and responsiveness estimates.

Keywords Bioassessment · Macroinvertebrates · Method performance · Data quality · Streams

J. Diamond (✉) · J. ~~R.~~ Stribling · J. Gilliam
Tetra Tech, Inc., Owings Mills, MD, USA
e-mail: jerry.diamond@tetratech.com

L. Huff
Alabama Department of Environmental Management, Montgomery, AL, USA

Introduction

A large number of public and private organizations in the U.S. collect aquatic biological data using a variety of sampling and analysis methods (ITFM 1995a; Carter and Resh 2001). In the majority of cases, these data are collected for the purpose of assessing condition of a waterbody to help ensure that the goals of the US Clean Water Act are met. While the information collected by an individual organization is usually beneficial to its own program, providing comprehensive assessment or condition information on a regional, state, or national level has been problematic due to unknown data quality and unknown comparability of data or assessments produced by different programs and methods (ITFM 1995a; Diamond et al. 1996; NWQMC 2001; Mackey 2002; GAO 2004). Reliability and confidence in long-term, broad scale datasets is directly related to maintenance of data quality and the ability of the scientific community to summarize and communicate that information (Costanza et al. 1992; Edwards 2004). This problem extends to other uses of bioassessment information (such as total maximum daily load studies, habitat restoration, and assessment of best management practices, in which data from multiple organizations are often needed to answer management questions.

The issues of unknown data quality and undetermined bioassessment comparability ultimately affect both data generators and data users. Having bioassessment data that meet minimum quality criteria, and having comparable bioassessment methods, data generators (e.g., state, tribal, and national bioassessment programs) could more effectively pool resources, potentially increasing the spatial or temporal coverage of information or improving characterization of reference conditions within a shared ecoregion (ITFM 1995b; Diamond et al. 1996; Barbour et al. 1999; NWQMC 2001). Data users (e.g., US EPA's Office of Water) would also benefit from having more and/or better information with which to answer some of our most fundamental questions such as: what is the condition of the nation's surface waters and are protection and restoration programs working effectively? Further, if the quality of bioassessment data is known, it may be possible to use those data for purposes other than their original intent, thus prolonging their usefulness and value (GAO 2004).

While several research studies have examined bioassessment comparability, particularly for stream macroinvertebrates (Astin 2006; Jessup and Gerritsen 2006; Houston et al. 2002), most of these did so via comparisons of assessment ratings without knowledge of the underlying data quality or performance of each assessment protocol (but see Herbst and Silldorff 2006). The term *performance* as used in this paper refers to the resulting data quality, assuming the method is performed correctly and is used with types of samples that the method is supposedly capable of analyzing. In water-quality monitoring, method performance is typically characterized in terms of various quantitative data quality characteristics such as precision (how similar are duplicate measurements), sensitivity (how small a change in condition can the method reliably detect), and bias (systematic difference from the true value due to the method). The values specified by a program or data user for these various data quality characteristics are often termed *performance criteria* (NWQMC 2001). While method performance has been characterized for many chemical and microbiological analytical methods, performance of bioassessment methods has generally not been characterized thus far.

The need for better guidance on determining bioassessment performance and comparability has been voiced in several scientific workshops and conferences, as well as state and regional biologist associations. For example, neighboring states sharing an ecoregion often have a need to share reference sites (i.e., least impaired or relatively "natural" site) in order to make more robust assessments for that ecoregion (Houston et al. 2002). At issue is whether the reference site data and assessments are comparable between the two states such that they can be used interchangeably (or together) in each state. To address this need, a pilot study was conducted by several southeastern US states including Kentucky, Georgia, North and South Carolina, Tennessee, and Alabama using benthic macroinvertebrate assessments. This paper presents the approach developed for this pilot and discusses several recommendations resulting from this study. While the results pertain to benthic macroinvertebrate assemblages, the approach was designed to accommodate any aquatic assemblage. State identities in terms of results are kept anonymous to avoid unintended comparisons or judgments regarding their programs.

Methods

Sampling design

All of the quantitative performance characteristics were calculated based on the sampling design shown in Fig. 1. Benthic macroinvertebrate sampling was performed at each site using normal state sampling protocols. For each ecoregion and each state examined, the desired goal was a minimum of 20 sites, with 10 of the sites along a human disturbance gradient, and 10 sites that are considered by the state to be representative of reference conditions for the ecoregion.

The number of sites sampled by each state is summarized in Table 1. For some ecoregions and states, the desired number of reference and/or test sites was unavailable. Most of the states did not have data from 10 reference sites in the ecoregion being evaluated due to either a lack of reference-quality sites (because of extensive human use impacts or the small size of the ecoregion in their

Fig. 1 Schematic of sampling design on which stream macroinvertebrate data are based for performance evaluations

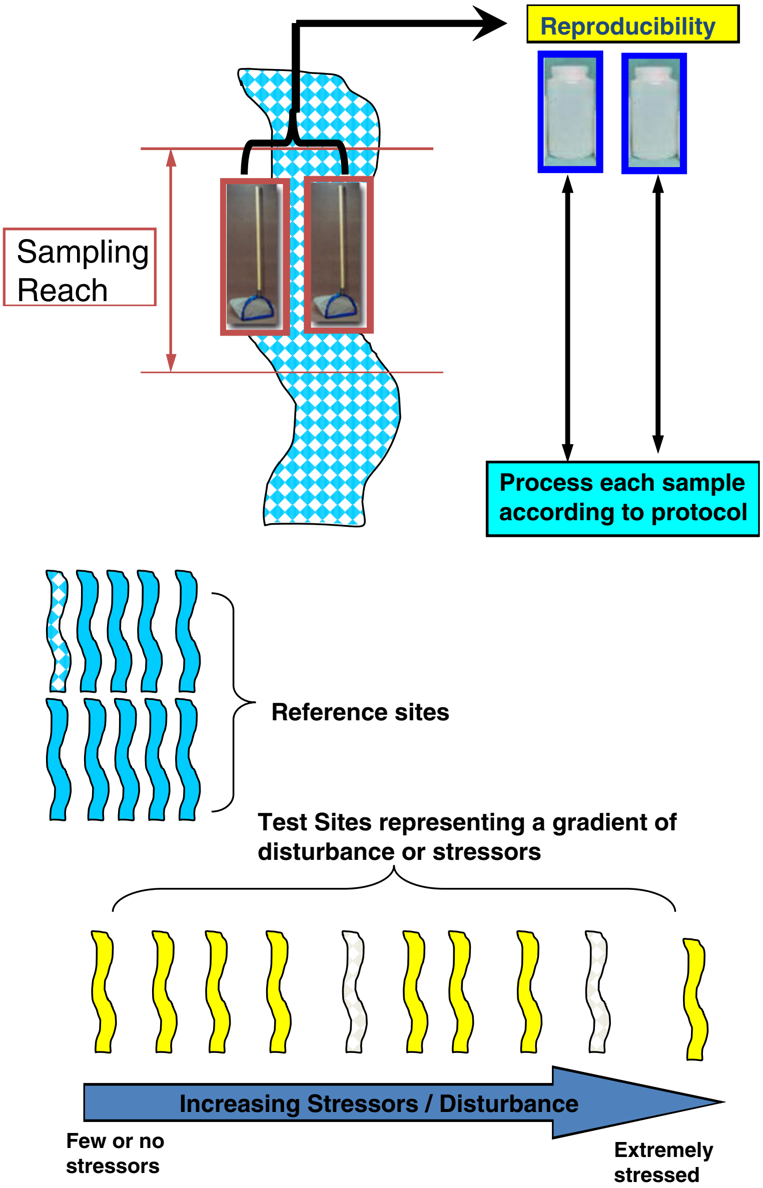

Reproducibility
Sampling Reach
Process each sample according to protocol
Reference sites
Test Sites representing a gradient of disturbance or stressors
Increasing Stressors / Disturbance
Few or no stressors
Extremely stressed

Table 1 Summary of sampling conducted by each state in the pilot study

State	No. reference sites	No. test sites	No. replicated sites	No. side-by-side sites	Ecoregion sampled
A	7	8	4	16	45
B	3	9	4	10	45
C	9	32	7	None	75
D	3	8	3	6	71
E	5	21	6	12	71
F	8	30	6	None	45

own state) or because only a subset of their reference site data were available at the time of this study. Also, half the states did not have 10 "test" sites spanning a human disturbance gradient as desired.

In some cases, states supplied biological data that were not used because all of the necessary ancillary data were not available for a given site (e.g., certain desired stressor data were missing) or because biological data were collected in different years or index periods, which would introduce an additional source of uncertainty in our analyses.

The sampling design used in this study was a compromise between cost/effort and desired statistical rigor. Including more reference and test sites in these analyses is likely to improve performance estimates. Each state used its normal sampling and sample processing methods, which differ in some respects among states. Differences in overall results due to sampling or processing methods were not evaluated in this study due to resource constraints. For each ecoregion and each state examined, two replicate samples were collected from at least three of the sites (preferably one reference and two along the disturbance gradient) in the field. All samples were handled, processed, sorted, and analyzed according to the program's normal methods.

In addition to the sampling design described above, several states also conducted joint sampling of a few sites within an ecoregion. Table 1 summarizes the number of sites jointly sampled by different states in each ecoregion. These joint samplings were used to examine similarity of assessments between states and, in conjunction with the bioassessment performance information calculated for each state, were used to assess the degree of comparability of state assessment protocols.

Characterizing human disturbance gradient

A key component in the study design and in the analyses of bioassessment performance is characterization of the human disturbance gradient (HDG) and how sites are scored along this gradient (Karr and Chu 1999; Fore 2004; Yuan and Norton 2003). The HDG used in this study was based in part on a modification of the Landscape Development Index (LDI; Brown and Vivas 2005), which has been used successfully in Florida (Fore 2004) and in other regions to characterize human disturbance status (Tetra Tech, Inc 2006). The LDI is calculated as the percentage area of particular types of land use multiplied by the coefficient of energy use associated with that land use, summed over all land use types in the catchment (Brown and Vivas 2005):

$$\mathrm{LDI} = \sum (\mathrm{LDI}i \times \%\mathrm{LU}i)$$

where,

LDI*i* A coefficient defining the nonrenewable energy flow for land use *i*, and

%LU*i* The percentage of land area in the catchment with land use *i*.

In this study, the human disturbance gradient using the LDI approach incorporated 14 different land use categories, including open water, low, medium, and high intensity residential, row crop, pasture, and deciduous, evergreen, and mixed forest. LDIs were calculated for each site using the most recent MRLC land cover data and ArcGIS (ESRI, Inc., Redlands, CA). Land use information within a 2 km radius of a given site was used to compute an LDI value for each biological sampling site.

The HDG used in this study, similar to that used in Florida (Fore 2004), was based on a composite of the LDI, instream physical habitat quality scores, and conductivity as an indicator of water quality. Several studies have demonstrated a relationship between conductivity and general water quality in freshwater systems within a given ecoregion or geology/soils region (e.g., Wiley et al. 2003). Higher conductivity is often associated with urban runoff, wastewater effluents, and irrigation return water (Brown and Vivas 2005). Instream habitat quality also profoundly influences benthic macroinvertebrate assemblage structure and function and indirectly represents physical habitat stressors that may be present such as sedimentation, scour, and stream bank instability (Rosenberg and Resh 1993; Barbour et al. 1999). Instream habitat quality score and conductivity data for each site were normalized to the mean reference values obtained from the same ecoregion by the state. Sites having a higher intensity of urban or agricultural land use in the surrounding riparian corridor or that had habitat or conductivity that was worse than reference conditions in the same ecoregion, were scored higher (i.e., more stressed) on the HDG. An HDG score was assigned to each indicator using threshold levels summarized in Table 2. These threshold levels were derived from literature values as well as based on state databases for conductivity and physical habitat quality in a given ecoregion. The values shown for Level 1 in Table 2 represent reference site conditions reported by the respective state. An overall HDG score was calculated by averaging the scores of the individual components (conductivity, habitat quality, and LDI). Analyses comparing the Florida approach (Fore 2004) and this study indicated high similarity in how sites were scored on the human disturbance gradient ($r > 0.8$, $p < 0.05$).

Given the importance of using an accurate representation of disturbance condition based on non-biological information, bioassessment performance will depend on two sources of error: error due to the bioassessment protocol itself and error due to classification of sites on the HDG. Therefore, attempts were made to have more than one test site in a given HDG class for each state and ecoregion.

Performance characteristics

A performance-based protocol requires that data quality is documented to the extent that one could make an informed decision on the appropriateness of resulting data or assessments for a specified management question or objective. Performance characteristics examined in this study included precision, sensitivity, and responsiveness. Bias is not truly known for bioassessments and was, therefore, not examined in this pilot study. Table 3 summarizes the definitions for each of these characteristics, as used in this study. All analyses were based on the final assessment value (e.g., IBI score) as well as component metric values that a given state typically uses (e.g., EPT). Separate performance calculations were made for each state and each Level III ecoregion sampled (Table 1).

Calculation of performance measures

All performance measures were calculated using data from a single index period and year of sampling for each state and ecoregion. Precision of a bioassessment protocol was calculated for each pair of replicate values within a given ecoregion using relative percent difference:

$$\left[\text{absolute value }(\text{measure}_1 - \text{measure}_2)/\text{mean}\right] \times 100)$$

where measure_1 and measure_2 are the index values (or metric values) obtained using the replicate samples from a given site. RPD values for each replicate pair were then averaged to yield overall precision for each state protocol and ecoregion examined.

Sensitivity was evaluated using ANOVA and Tukey's HSD test ($p < 0.05$) with HDG class as the treatment categories and either the index or metric values as the dependent variable. Index and metric values met requirements of normality and variance homogeneity. Sensitivity of each state bioassessment protocol was then compared in terms of the lowest disturbance level that could be detected in comparison with the reference values.

Table 2 Approach for calculating a human disturbance gradient (HDG) score for each biological sampling site

Ecoregion	State(s)	Conductivity (μmhos/cm)					Habitat quality (max = 200)[a]				
		Level 1	Level 2	Level 3	Level 4	Level 5	Level 1	Level 2	Level 3	Level 4	Level 5
45	A[b]	<225.3		225.3–396.7		>396.7	>148.6		148.6–121.4		<121.4
45	B	<28	28–49.9	49.9–71.8	71.8–93.7	>93.7	>116.2	116.2–110.4	110.4–104.6	104.6–98.8	<98.8
45	C	<30.43	30.43–44.07	44.07–100	100–200	>200	>189.5	189.5–168	168–146.5	146.5–125	<125
71	D	<172	172–284	284–396	396–508	>508	>153.8	153.8–141.6	141.6–129.4	129.4–117.2	<117.2
71	E	<400.6	400.6–592.2	592.2–738.8	738.8–975.4	>975.4	>141.6	141.6–127.2	127.2–112.8	112.8–98.4	<98.4
75	F	<90.8	90.8–154.6	154.6–218.4	218.4–282.2	>282.2	>159	159–138	138–117	117–96	<96

Ecoregion	State(s)	LDI				
		Level 1	Level 2	Level 3	Level 4	Level 5
45	A[b]	<1.8		1.8–2.29		>2.29
45	B	<1.55	1.55–1.93	1.93–2.3	2.3–2.68	>2.68
45	C	<1.26	1.26–1.52	1.52–1.8	1.8–3	>3
71	D	<2.16	2.16–2.6	2.6–3.03	3.03–3.47	>3.47
71	E	<1.56	1.56–2.06	2.06–2.57	2.57–3.07	>3.07
75	F	<1.72	1.72–2.42	2.42–3.12	3.12–3.82	>3.82

[a] A lower score indicates more disturbed conditions

[b] Only three levels of the disturbance gradient could be determined for this state dataset because there were too few sites and the range of test sites was too narrow to reliably identify more levels

Table 3 Performance characteristics that were documented for bioassessment methods in this study

Performance characteristic	Definition
Precision	Variability in biological index or indicator measure using duplicate field samples or samples from multiple reference sites (field sampling and lab error, as well as random variability; reproducibility)
Sensitivity	Level of human disturbance at which the biological index or indicator is statistically different from reference condition or from another site, if a reference site is unavailable.
Responsiveness	Degree to which a biological index or indicator measure decreases monotonically with increasing disturbance level.

Responsiveness was calculated using linear regression of the index or metric values against the HDG score to calculate responsiveness as determined by the significance and sign of the slope and associated R^2 value. A significant slope in the appropriate direction (i.e., index value indicates poorer condition as disturbance level increases) and an R^2 value above 0.5 would indicate moderate–high responsiveness of the index.

Side-by-side data

Four states participated in joint sampling exercises to investigate method comparability. Two neighboring states sharing an ecoregion sampled several (6–10) sites spanning a range of potential disturbance levels. Sites were ranked using the within-state index scores given by each of the states, which represents biological condition classes using their respective system. Rankings were then compared between two states using Kendall's Tau nonparametric correlation (Statistica Version 8.0, Statsoft, Inc., Tulsa, OK).

Results

Reference site human disturbance scores

Reference condition often spanned a fairly wide range of habitat quality, conductivity, and LDI values in this study, even within a single ecoregion (Table 4) indicating a range of potentially stressed sites being used to define reference condition for certain ecoregions and states. For example, Ecoregion 45 had habitat quality scores ranging between 97 and 200 at reference sites among the three states that sampled that ecoregion (Table 4). Qualitatively, these habitat scores ranged from fair to excellent according to the state's narrative assessments. This ecoregion also had a particularly wide range of HDG component values (Tables 2 and 4) for the three states suggesting potential difficulties in identifying minimally impaired reference conditions there. There also may have been substantial differences in reference site quality between states in terms of human disturbance levels as demonstrated by Ecoregion 71 (Tables 2 and 4). The apparent difference in range of HDG values between two states in this ecoregion (based on the limited data available for this study) appeared to be primarily related to a difference in the LDI range, with one state having some reference sites with higher disturbance

Table 4 Observed range of the three component variables used in the human disturbance gradient for state Reference sites used in this study

Ecoregion	State	Habitat score range (out of 200)	Conductivity (µmhos/cm)	LDI score (Max=5)	HDG score
75	A	155–180	41–242	1.02–1.18	1–2
71	B	132–166	60–463	1.73–3.63	1–4
	C	133–156	384–435	2.10–2.70	2
45	D	104–176	54–133	1.31–2.42	1–4
	E	97–113	38–111	1.17–1.67	2–4
	F	157–200	17–42	1.00–3.19	1–3

The lower the LDI score, the fewer human stressor sources present and the more the stream is assumed to be high quality values based on the LDI and land use information (Table 4).

The range of conductivity values associated with reference conditions was also different as evidenced by the difference in Level 1 thresholds for the two states in Ecoregion 71 (Table 2). A similar result was observed for conductivity and habitat quality thresholds for Level 1 in Ecoregion 45 (Table 2). For example, the Level 1 threshold for conductivity in state "A" was nearly 10 times the threshold observed for the same ecoregion in states "B" and "C" (Table 2). These results should be treated with caution due to the relatively few sites available for analysis in this pilot study. However, they suggest that the HDG scale might not be uniform within a Level III ecoregion class.

Assessment index performance

Overall, performance results indicated that most states had similar assessment index performance characteristics (Table 5). All of the state assess-

Table 5 Summary of preliminary performance characteristics measured for state component metric and assessment index score in each ecoregion sampled in southeast USA

State	Metric	Precision (mean RPD)	Sensitivity (HDG level different from reference)	Responsiveness (R^2 values)
A	TR	13.28	2	0.3802
	EPTR	18.02	3	0.4672
	NCBI	7.08	3	0.3543
	%Domtaxa	17.02	–	0.0145
	%EPT	18.60	3	0.0873
	Index	16.11	3	0.2864
B	%Oligo	0.00	–	0.4879
	%Tanyp	4.74	–	0.0045
	%Filt	89.17	–	0.0553
	TolTax	72.89	–	0.1114
	Index	11.22	–	0.052
C	TR	10.34	–	0.2842
	EPTR	16.27	3	0.3176
	HBI	1.17	3	0.0495
	%EPT	0.00	4	0.2752
	%C+O	2.51	–	0.0026
	%Cling	11.24	–	0.13
	Index	4.42	–	0.271
D	TR	24.56	2	0.2846
	Intol	22.09	2	0.1998
	Intol Ab	31.12	2	0.2711
	NCBI	9.91	2	0.049
	Index	6.87	2	0.0596
E	Abundance	9.18	–	0.154
	TR	7.41	3	0.0154
	EPTR	10.05	2	0.0367
	NCBI	5.33	2	0.0002
	%EPT	8.12	2	0.0355
	Index	5.64	2	0.0103
F	Abundance	11.45	2	0.0122
	TR	24.68	–	0.0001
	EPTR	24.69	2	0.1026
	% EPT	31.11	2	0.0972
	% C+O	66.65	–	0.0311
	NCBI	15.85	3	0.1682
	% Domtaxa	29.07	–	0.2939
	%Cling	12.38	–	0.0015
	Index	11.60	2	0.0374

Performance measures are based on a suboptimal number of sites in most cases and should therefore be treated as very preliminary. A dash for sensitivity indicates that no HDG level was significantly different from the best reference condition

TR Taxa richness; *EPTR* richness of EPT taxa; *NCBI* North Carolina Biotic Index; % *Dom taxa* percentage of dominant taxa; %*EPT* percent of abundance that is EPT individuals; %*Oligo* % oligochaeta; %*Tanyp* %tanypodinae; %*Filt* % filterers; *TolTax* number of tolerant taxa; %*Cling* % clinger taxa; *HBI* Hilsenhoff Biotic Index (modified); %*C+O* % individuals identified as chironomids or oligochaetes; *Intol* % intolerant taxa; *IntolAb* abundance of intolerant individuals

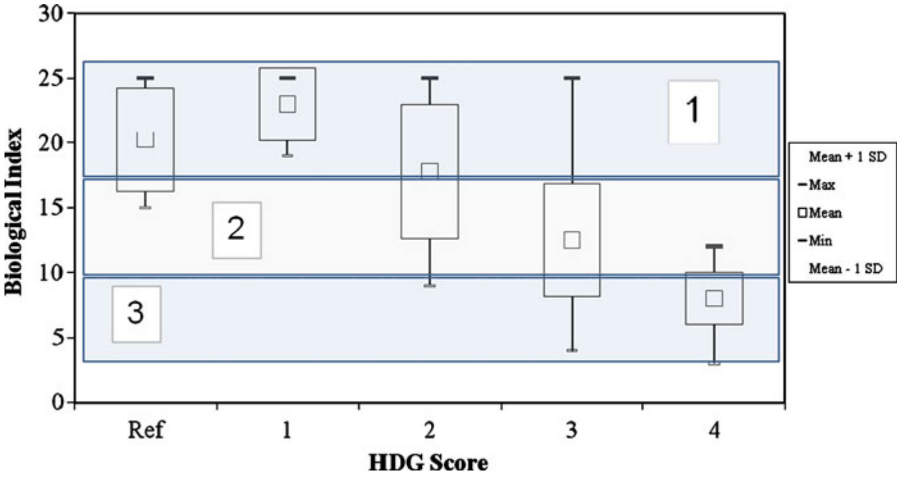


Fig. 2 Example of an index that is fairly responsive, distinguishing three levels of disturbance as denoted by *numbered areas* of the graph

ment indices had relative percent difference (RPD) values <20% (0.20) and most of the indices had RPD values <7% (0.07), indicating fairly high precision of replicate measures based on the assessment index. In terms of sensitivity, four of the six state assessment indices could distinguish HDG class 2 from reference, where the best reference condition recorded by the state was equivalent to HDG level 1. Greater sensitivity is demonstrated if a state index can distinguish a lower HDG class (e.g., level 2 rather than level 3) from reference. Thus, states D, E, and F appear to demonstrate greater sensitivity than states A, B, or C (Table 5). For most states, an HDG class of 4 or 5 would constitute an impaired site. For two state indices (states B and C), we were unable to detect a significant difference between HDG level 1 and any of the other HDG levels (Table 5). Index responsiveness, as determined using our approach, was generally low ($R^2 < 30\%$ or 0.30) for most states (Table 5). While this result may be due at least in part to the small sample size available for most states in this study, we did not see a relationship between number of sites and responsiveness R^2 values. For example, state “C,” which had the fewest total number of sites (11), had one of the highest responsiveness R^2 values (Table 5). This suggests that having more sites does not necessarily guarantee higher index responsiveness. The most responsive state index could distinguish three different condition classes equating to good, fair, poor (e.g., Fig. 2) as evidenced by the fact that the three categories were significantly different from each other. For most states, two condition classes (good/fair and poor) were distinguishable statistically.

Metric performance

For the most part, no one metric displayed the best performance across states and ecoregions (Table 5). For most states, metric precision RPD values were <20% and in many cases <10%. Metrics that appeared to be most precise overall were taxa richness (TR), North Carolina Biotic Index (NCBI), and % clingers. Precision of certain metrics varied between states (e.g., TR RPD values ranged between 7% and 25%). Given similar taxonomic levels identified among these states, differences in precision among states for a given taxonomic metric may be due to a combination of differences in field sampling and sample processing replicability.

Most metrics used by each state were fairly sensitive using the simplistic HDG scale in this study (Table 5). Functional group metrics (e.g., % clingers) or percentage metrics (e.g., % Dominance) appeared less sensitive overall while NCBI appeared to be relatively sensitive as indicated by a significantly higher (i.e., biologically poorer) value with increasing HDG level (Fig. 3). General taxa richness metrics were also often less sensitive than other metrics using the HDG in this study and the data available from each state.

Of the metrics examined, EPT taxa and other taxa richness metrics showed the most respon-

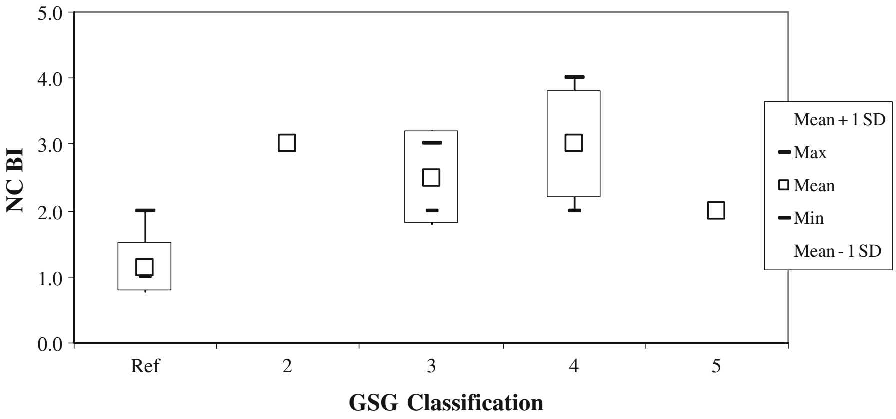


Fig. 3 Example of a metric displaying high sensitivity. The reference condition has a lower NCBI score than other HDG classes. (Note: a lower NCBI score equates to better biological condition)

sive relationships with increasing HDG score (Table 5); however, all R^2 values were fairly low (most <30% or 0.30). This result is not surprising given the relatively few sites available for each state and the simplistic HDG used. However, certain metrics appeared especially unresponsive in more than one state including %EPT and HBI (Table 5). It was not uncommon for an individual metric to appear more responsive than the final assessment index for a given state (e.g., TR, EPTR, and NCBI for state "A" vs its Index, Table 5).

Joint bioassessment data

Joint state bioassessment data demonstrated that states generally classified sites similarly (Table 6). At nearly all sites, similar percentage ranges with regards to the average reference value were observed between states for the same sites and sites generally had similar narrative ratings between states. Concordance for pairs of state assessments was generally >90% indicating similar rankings of sites. If a two-class ranking is used (impaired vs. not impaired), concordance was >95% for all sites and States. Out of 22 sites that had paired state assessments, two sites had bioassessments that differed by more than one assessment classification; e.g., one state assessed a site as slightly impaired and the other state assessed the same site as excellent. The few differences observed may have been attributable to a difference in reference conditions between the states.

Table 6 Summary of Kendall Tau concordance analysis of bioassessment indices for joint site bioassessments

Ecoregion	States	# Sites	Kendall's coefficient of rank correlation, τ
71	A & B	6	0.931
45	C & D	10	0.954
66	C & A	6	0.966

All pairwise assessments were significantly correlated ($p < 0.05$)

Discussion

All forms of monitoring are subject to issues of data quality and comparability of methods and data. The approach presented in this paper is intended to help a program document bioassessment performance, which can be useful not only to the program itself but also to other programs in providing information as to whether biological assessments can be considered comparable and therefore combined into a larger spatial scale assessment. Acknowledging the limitations of this study (relatively few sites from each state and ecoregion and a fairly simplistic calculation of human disturbance level), most state bioassessment indices analyzed in this study could distinguish two or perhaps three condition classes, including reference condition. We would anticipate that performance levels improve if more site data were available and the disturbance gradient was addressed more thoroughly. Based on our preliminary results, we would recommend at least 50 sites, with 10 sites being reference sites as determined by the state for a given ecoregion in order to obtain data from the entire human disturbance gradient in an ecoregion so as to improve statistical confidence in performance measures. Having too few sites for some states in this study resulted in an under-representation of certain parts of the disturbance gradient (e.g., few poor quality sites), undoubtedly biasing sensitivity and responsiveness estimates in those cases.

One obvious need resulting from this exercise is harmonization of reference condition expectations. Many studies have indicated that differences in reference condition criteria among organizations may be the single most important factor affecting both bioassessment performance and comparability (Stoddard et al. 2006). By conducting the analyses recommended in this document, organizations are likely to be in a more knowledgeable position to address issues such as disparate reference condition criteria.

Another result suggested by this study is that for some states and ecoregions, certain metrics may be less precise, sensitive, or responsive than others, potentially resulting in suboptimal assessment index performance. It was not unusual for example to find that certain metrics outperformed the overall assessment index due to the incorporation of apparently poorer performing metrics. While the results of this study should be treated cautiously due to small sample sizes, the approach presented in this paper could be helpful to a state program in refining their assessment index and the component metrics used.

If a bioassessment program needs to distinguish only impaired from unimpaired sites, then most state indices examined in this study appear to satisfy that objective. If, however, a program needs finer resolution of condition status (e.g., in order to help prioritize management needs; protect truly excellent sites from gradual degradation over time; assess results of instituting best management practices or restoration activities), then results of this preliminary analysis suggest that most of the bioassessment protocols examined may not have sufficient sensitivity or responsiveness to meet such an objective. The need to distinguish more than two levels of biological condition in bioassessment protocols has become increasingly apparent, particularly in terms of identifying and maintaining exceptional or high quality ecological conditions when they occur. For many state programs, a relatively high quality biological condition can degrade to a somewhat poor biological condition before the change in condition is recognized and management action is taken (Davies and Jackson 2006). A bioassessment index that is more sensitive is less likely to encounter this issue.

It is important to note that index performance is typically characterized once so that expected data quality can be determined. Routine quality-control analyses are conducted each time the protocol is used to document that performance expectations are met. As a general rule-of-thumb, characterization of data quality is performed at a rate of 10% of sites/samples using repeated field sampling, sample reprocessing, and for calculation of indicator variables (Stribling et al. 2008a). Performance documentation should be repeated if any method or gear changes or the protocol is being extended to different types of habitats, ecosystems, or ecoregions. Each of these factors is known to potentially affect results of a bioassessment protocol (Barbour et al. 1999), and differences in some of these methods among states likely contributed to the overall results observed. Routine QC analyses are used to document that personnel and equipment actually meet performance expectations for the protocol.

If a program already has the minimum recommended dataset with which to analyze performance, as described in this paper, there is no need for additional sampling; performance analyses can be conducted on existing data. If a program does not have the minimum data requirements, some additional sampling is required. In carrying out the procedures described in this study, it is important that the performance characterization process is unbiased and a true representation of the protocol as routinely practiced.

By conducting a thorough evaluation of bioassessment performance as recommended in this study, an organization or program could determine which part(s) of the protocol may limit current performance. In so doing, this evaluation could help identify refinements that are likely to produce the greatest improvement in performance with the least additional expenditure of resources.

While several organizations have already embarked on the process of evaluating performance of either parts of their protocols (e.g., taxonomic data quality) or their entire protocol, much yet could be done. For example, it may be desirable (perhaps necessary) for a program such as the Clean Water Act 305(b) ambient monitoring program and assessment to use data only from protocols that have documented some minimum level of performance (i.e., precision, sensitivity, etc.) to ensure a minimum level of data quality nationally. Currently, most programs attempt to ensure a consistent performance level through standardized sampling and assessment protocols (e.g., USGS NAWQA program (Cuffney et al. 1993); USEPA Mid-Atlantic Assessment (Herlihy et al. 2000)). While these programs ensure consistency within their program, they are inconsistent with each other, and they are costly, resulting in reduced sampling frequency (and trend analysis) and coarse spatial coverage. Furthermore, bioassessment performance as outlined in this paper has not been documented in these programs, nor is it known whether the performance being achieved by each satisfies program objectives. At least in the case of wadeable stream benthic macroinvertebrate assessments (and perhaps the case could be as readily made for large river macroinvertebrate and fish assemblages (Flotemersch et al. 2006)), most state resource agencies and many other programs across the country have assessment information spanning many years, which could potentially be combined to develop regional and national assessments if bioassessment performance is known to meet some useful minimum standard.

Encouraging minimum performance standards for bioassessments would fully use existing monitoring resources, yielding a less costly and more informative national assessment program. States and tribes, as part of "Credible Data" requirements, may also find that developing bioassessment performance standards ensures that data used in evaluating status and trends and listing impaired waters are of appropriate quality. Finally, states and tribes that share common basins or borders may have much to gain in documenting and comparing bioassessment performance. If a certain level of performance can be consistently achieved among organizations, there may be more opportunities for data sharing and combining data to derive larger scale assessments. The pilot study examined in this paper evaluated and compared performance for this very reason.

In addition to evaluating and analyzing performance measures, it is just as critical to document observed performance characteristics in a useful way. Participating states in this pilot developed a summary form as a standardized framework with which each state could document their bioassessment performance. Ultimately, it may be feasible to document bioassessment performance in relatively simple terms such that protocols can be readily compared for a given objective. These measures along with other method information could be summarized, such as in the online methods database National Environmental Methods Index (www.nemi.gov), which currently houses performance measures and other information for over 1200 chemical and microbiological methods. Using such a system, organizations could learn from each other as to what factors or method modifications are likely to yield the most benefit in terms of addressing different monitoring objectives while balancing overall bioassessment quality, efficiency, and cost.

While having similar bioassessment performance is a critical aspect of comparability, it is not sufficient. On an assessment level, many other factors will determine comparability. These factors include sample collection method, sampling design, taxonomic levels used, and reference condition criteria used. Several studies have provided approaches and case studies examining performance of individual aspects of a bioassessment protocol such as sampling (Cao et al. 2005; Blocksom and Flotemersch 2005; Peterson and Zumberge 2006; Stribling et al. 2008b), sorting efficiency for macroinvertebrates, taxonomic data quality (Hawkins and Norris 2000; Stribling et al. 2003, 2008a), and scoring systems (Blocksom 2003; Astin 2006; Reynoldson et al. 1997). Although still preliminary, many researchers have reported that if performance is similar and sampling methods are relatively similar, assessments are likely to be comparable for most objectives (Herbst and Silldorff 2006; Jessup and Gerritsen 2006). More studies examining this issue could help define "boundaries" within which method differences or reference condition criteria differences are of relatively little consequence and assessments can be considered comparable.

Acknowledgements We appreciate the contributions made to this study by Debbie Arwine, Tennessee DEC, Michelle Brossett, Georgia EPD, William Crouch, North Carolina DENR, James Glover, South Carolina DHEC, Jim Harrison, EPA Region 4, and Mark Vogel, Kentucky DOW. David Bressler provided GIS and LDI analyses. This work was funded in part by Task Order 123 within USEPA contract #68-C-02-108.

References

Astin, L. (2006). Data synthesis and bioindicator development for nontidal streams in the interstate Potomac River Basin, USA. *Ecological Indicators, 6*, 664–685.

Barbour, M. T., Gerritsen, J., Snyder, B. D., & Stribling, J. B. (1999). *Rapid bioassessment protocols for use in streams and Wadeable rivers: Periphyton, Benthic macroinvertebrates and fish* (2nd ed.). Washington, D.C.: US Environmental Protection Agency, Office of Water, EPA 481-B-99-002.

Blocksom, K. (2003). A performance comparison of metric scoring methods for a multimetric index for mid-Atlantic highland streams. *Environmental Management, 31*, 670–682.

Blocksom, K., & Flotemersch, J. (2005). Comparison of macroinvertebrate sampling methods for nonwadeable streams. *Environmental Monitoring and Assessment, 102*, 243–262.

Brown, M. T., & Vivas, M. B. (2005). Landscape development intensity index. *Environmental Monitoring and Assessment, 101*, 289–309

Cao, Y., Hawkins, C. P., & Storey, A. D. (2005). A method for measuring the comparability of different sampling methods used in biological surveys: Implications for data integration and synthesis. *Freshwater Biology, 50*, 1105–1115.

Carter, J. L., & Resh, V. H. (2001). After site selection and before data analysis: Sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. *Journal of the North American Benthological Society, 20*, 658–676.

Costanza, R., Funtowicz, S. O., & Ravetz, J. R. (1992). Assessing and communicating data quality in policy-relevant research. *Environmental Management, 16*(1), 121–131.

Cuffney, T., Gurtz, M., & Meador, M. (1993). *Methods for collecting benthic invertebrate samples as part of the National Water Quality Assessment Program*. Reston: U.S. Geological Survey, Open-File Report 93-406.

Davies, S. P., & Jackson, S. K. (2006). The biological condition gradient: A descriptive model for interpreting change in aquatic ecosystems. *Ecological Applications, 16*, 1251–1266.

Diamond, J. M., Barbour, M. T., & Stribling, J. B. (1996). Characterizing and comparing bioassessment methods and their results: A perspective. *Journal of the North American Benthological Society, 15*, 713–727.

Edwards, P. N. (2004). A vast-machine: Standards as social technology. *Science, 304*(7), 827–828.

Flotemersch, J. F., Stribling, J. B., & Paul, M. J. (2006). *Concepts and approaches for the bioassessment of non-wadeable streams and rivers. EPA/600/R-06/127*. Cincinnati: U.S. Environmental Protection Agency.

Fore, L. S. (2004). *Development and testing of biomonitoring tools for macroinvertebrates in Florida streams*. Tallahassee: Florida Department of Environmental Protection.

GAO (2004). Watershed management: Better coordination of data collection efforts. General Accounting Office, GAO-04-382. http://www.gao.gov/new.items/d04382.pdf.

Hawkins, C. P., & Norris, R. H. (2000). Effects of taxonomic resolution and use of subsets of the fauna on the performance of RIVPACS-type models. In: J. F. Wight, D. W. Sutcliffe, & M. T. Furse (Eds.), *Assessing the biological quality of fresh waters: RIVPACS and other techniques* (pp. 217–228). Ambleside: Freshwater Biological Association.

Herbst, D., & Silldorff, E. (2006). Comparison of the performance of different bioassessment methods: Comparable evaluations of biotic integrity from contrasting procedures. *Journal of the North American Benthological Society, 25*, 513–530.

Herlihy, A. T., Larsen, D. P., Paulsen, S. G., Urquhart, N. S., & Rosenbaum, B. J. (2000). Designing a spatially balanced, randomized site selection process for regional stream surveys: The EMAP Mid-Atlantic pilot study. *Environmental Monitoring and Assessment, 63*, 95–113.

Houston, L., Barbour, M. T., Lenat, D., & Penrose, D. (2002). A multi-agency comparison of aquatic macroinvertebrate-based stream bioassessment methodologies. *Ecological Indicators, 1*, 279–292.

ITFM (1995a). The Strategy for Improving Water Quality Monitoring in the U.S. Report #OFR95-742. Reston: U.S. Geological Survey.

ITFM (1995b). Performance-based approach to water quality monitoring. In: Strategy for improving water quality monitoring in the U.S., Appendix M, Report #OFR95-742, Interagency Task Force on Monitoring Water Quality. Reston: U.S. Geological Survey.

Jessup, B., & Gerritsen, J. (2006). Data and assessment comparability among stream bioassessment methods: EPA-NEWS methods and New England State methods. Prepared for: Susan Holdsworth, USEPA Office of Water, Office of Watersheds, Oceans, and Wetlands, Washington, DC , Tetra Tech, Inc., Montpelier, VT, March 2006.

Karr, J. R., & Chu, E. W. (1999). *Restoring life in running waters: Better biological monitoring*. Washington, D.C.: Island Press.

Mackey, E. (2002). The state of the nation's ecosystems: measuring the lands, waters, and living resources of the United States. The H. John Heinz III Center for Science, Economics, and the Environment, Washington, D.C., Cambridge University Press. http://www.heinzctr.org/ecosystems/index.shtml.

NWQMC. (2001). *Towards a definition of performance-based laboratory methods. National Water Quality Monitoring Council Technical Report 01–02*. Reston: U.S. Geological Survey.

Peterson, D., & Zumberge, J. (2006). *Comparison of macroinvertebrate community structure between two riffle-based sampling protocols in Wyoming, Colorado, and Montana, 2000–2001. Scientific Investigations Report 2006–5117*. Reston: U.S. Geological Survey.

Reynoldson, T. B., Norris, R. H., Resh, V. H., Day, K. E., & Rosenberg, D. M. (1997). The reference condition: A comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. *Journal of the North American Benthological Society, 16*(4), 833–852.

Rosenberg, D., & Resh, V. (1993). *Freshwater biomonitoring and benthic macroinvertebrates*. New York: Chapman and Hall.

Stoddard, J., Larsen, D., Hawkins, C., Johnson, R., & Norris, R. (2006). Setting expectations for the ecological condition of streams: The concept of reference condition. *Ecological Applications, 16*, 1267–1276.

Stribling, J., Moulton, S. II, & Lester, G. (2003). Determining the quality of taxonomic data. *Journal of the North American Benthological Society, 22*, 621–631.

Stribling, J. B., Pavlik, K. L., Holdsworth, S. M., & Leppo, E. W. (2008a). Data quality, performance, and uncertainty in taxonomic identification for biological assessments. *Journal of the North American Benthological Society, 27*(4), 906–919.

Stribling, J. B., Jessup, B. K., & Feldman, D. L. (2008b). Precision of benthic macroinvertebrate indicators of stream condition in Montana. *Journal of the North American Benthological Society, 27*(1), 58–67.

Tetra Tech, Inc. (2006). Southern California Coastal stream tiered aquatic life uses: Pilot study. Prepared for Los Angeles Regional Water Quality Control Board and USEPA, Washington, DC. Tetra Tech, Inc., Owings Mills, MD.

Wiley, M. J., Seelbach, P. W., Wehrly, K., & Martin, J. S. (2003). Regional ecological normalization using linear models: a meta-method for scaling stream assessment indicators. In T. Simon (Ed.), *Biological response signatures: Indicator patterns using aquatic communities* (pp. 201–224). Boca Raton: CRC.

Yuan, L. L., & Norton, S. B. (2003). Comparing responses of macroinvertebrate metrics to increasing stress. *Journal of the North American Benthological Society, 22*(2), 308–322.