

## Striving for consistency in a national assessment: the challenges of applying a reference-condition approach at a continental scale

**Alan T. Herlihy<sup>1</sup>**

*Department of Fisheries and Wildlife, Oregon State University, Corvallis, Oregon 97331 USA*

**Steven G. Paulsen<sup>2</sup>, John Van Sickle<sup>3</sup>, AND John L. Stoddard<sup>4</sup>**

*National Health and Environmental Effects Research Laboratory, Western Ecology Division, US Environmental Protection Agency, Corvallis, Oregon 97333 USA*

**Charles P. Hawkins<sup>5</sup>**

*Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Watershed Sciences, Utah State University, Logan, Utah 84322-5210 USA*

**Lester L. Yuan<sup>6</sup>**

*National Center for Environmental Assessment, Office of Research Development, US Environmental Protection Agency, Washington, DC 20460 USA*

**Abstract.** One of the biggest challenges when conducting a continental-scale assessment of streams is setting appropriate expectations for the assessed sites. The challenge occurs for 2 reasons: 1) tremendous natural environmental heterogeneity exists within a continental landscape and 2) reference sites vary in quality both across and within major regions of the continent. We describe the process used to set expectations for the multimetric index of biotic integrity (MIBI) and observed/expected (O/E) indices generated from predictive models used to assess stream condition for the US Wadeable Streams Assessment (WSA). The assessment was based on a reference-site approach, in which the least-disturbed sites in each region of the US were used to establish benchmarks for assessing the condition of macroinvertebrate assemblages at other sites. Reference sites were compiled by filtering WSA sample sites for disturbance using a series of abiotic variables. Additional reference sites were needed and were obtained from other state, university, and federal monitoring programs. This pool of potential reference sites was then assessed for uniformity in site quality and comparability of macroinvertebrate sample data. Ultimately, 1625 sites were used to set reference expectations for the WSA. Reference-site data were used to help define 9 large ecoregions that minimized the naturally occurring variation in macroinvertebrate assemblages associated with continental-wide differences in biogeography. These ecoregions were used as a basis for developing MIBI and O/E indices and for reporting results. A least-disturbed definition of reference condition was used nationally, but we suspect that the quality of the best extant sites in ecoregions, such as the Northern Plains and Temperate Plains, was lower than that of sites in other ecoregions. For the MIBI assessment, we used a simple modeling approach to adjust scores in ecoregions where gradients in reference-site quality could be demonstrated conclusively. The WSA provided an unparalleled opportunity to push the limits of our conceptual and technical understanding of how to best apply a reference-condition approach to a real-world need. Our hope is that we have learned enough from this exercise to improve the technical quality of the next round of national assessments.

**Key words:** reference condition, reference sites, regionalization, biological condition gradient, regional assessments.

<sup>1</sup> E-mail addresses: alan.herlihy@oregonstate.edu

<sup>2</sup> paulsen.steve@epa.gov

<sup>3</sup> vansickle.john@epa.gov

<sup>4</sup> stoddard.john@epa.gov

<sup>5</sup> chuck.hawkins@usu.edu

<sup>6</sup> yuan.lester@epa.gov

The National Wadeable Streams Assessment (WSA; USEPA 2006) was an ecological assessment with the primary goals of evaluating the biological condition of the streams in the US and ranking the stressors that

might affect them. The biological assessment was done by analyzing the macroinvertebrate assemblages in each stream with a multimetric index of biotic integrity (MIBI; Stoddard et al. 2008) and observed/expected (O/E) indices derived from River InVertebrate Prediction And Classification System (RIVPACS)-type modeling (Yuan et al. 2008). Both of these assessment techniques rely on a reference condition approach (Bailey et al. 2004) to set expectations for assemblages in individual streams (Stoddard et al. 2006). Expectations should be region or even site specific because of natural variation in environmental conditions. Reference condition for an assessed site is approximated from information collected at reference sites within a region. This information provides the benchmark against which the ecological conditions of all other streams in a region are measured. Thus, in the WSA, reference sites were used for 2 major purposes: 1) to develop and calibrate the MIBI and O/E assessment models, and 2) to set the thresholds used to divide continuous assessment variables into good, fair, or poor condition classes.

Identification of reference sites is difficult and time consuming in any assessment. Ideal sites would be unaffected by human activities. However, it is doubtful that many such sites exist in the conterminous US. Indeed, such sites probably would be difficult to locate anywhere in the world. Sometimes locations can be identified that have experienced a minimal degree of human influence (e.g., wilderness areas). Streams in these areas generally are thought to be in a minimally disturbed reference condition (Stoddard et al. 2006). However, these locations also are rare and often provide a comparison for only a specific subset of stream types (e.g., high-gradient, high-altitude streams). In most other cases, least-disturbed reference sites must be used, and assessments are done by comparing sites to be assessed with the highest-quality sites within the study area. The selection of least-disturbed sites is usually based on best professional judgment or is done by screening abiotic data (Whittier et al. 2007).

For the WSA, the problem of identifying appropriate and comparable reference sites was greatly complicated by the continental scale of the assessment. US streams are extremely heterogeneous with respect to many natural environmental attributes. Thus, reference sites had to be selected to characterize this range of natural heterogeneity. Furthermore, the degree of landscape and waterway alteration that has occurred in different regions of the US varies greatly, so the availability and overall quality of reference sites also varies among regions.

The 1<sup>st</sup> challenge was to acquire enough reference-

site data within the time frame of the study to characterize reference conditions sufficiently for all assessed streams. A 2<sup>nd</sup> challenge was to develop a regionalization scheme that would minimize the effect of natural environmental variation on indicator values while providing large-enough sample sizes to allow statistically valid assessments within ecologically meaningful subregions of the US. The WSA was based on a probabilistic survey design, and in many parts of the US, the number of reference sites in the sample of randomly chosen sites was too small to characterize reference condition within a region. Therefore, WSA leaders considered using reference-site data that had been collected previously during different state, university, and federal programs. However, the field and laboratory protocols used in these different programs differed to varying degrees from WSA protocols, and the effects of these differences on data comparability had to be assessed. A 3<sup>rd</sup> challenge was the variability in reference-site quality among and within different regions of the country. For example, agricultural regions of the country (e.g., the Midwest) have undergone such extensive landscape and waterway degradation through time that even the best remaining streams are far removed from historical conditions (e.g., Karr et al. 1985). Test sites assessed relative to degraded reference sites are likely to appear to be in better condition than would test sites in regions with less-altered reference systems. Therefore, assessments had to be adjusted for differences in reference-site quality to ensure comparability among regions.

We describe our efforts to define consistent reference conditions for aquatic macroinvertebrates sampled during the WSA. First, we discuss our approach for screening and assembling a reference-site data set large enough to characterize the wadeable streams in the conterminous US. Next, we present our scheme for partitioning the effects of natural environmental heterogeneity on assemblage structure and composition. Third, we discuss our efforts to identify and resolve issues of data consistency associated with different sources of reference-site data. Last, we present an approach that can be used to compensate for differences in the absolute quality of reference sites across very large geographical regions.

## Methods

### *Data collection*

Sources of reference-site data were inherently constrained by the funds available for sampling new sites and by the comparability of previously collected

reference-site data with data collected during the WSA.

*WSA sample design.*—The WSA was a product of 2 surveys (Olsen and Peck 2008). All flowing waters (streams and rivers) in 12 western states (Washington, Oregon, California, Nevada, Arizona, Idaho, Montana, Wyoming, Utah, North Dakota, South Dakota, and Colorado) were sampled during the summers of 2000 to 2004 as part of Environmental Monitoring and Assessment Program (EMAP) Western Pilot Survey (EMAP-West; Stoddard et al. 2005b). The 841 EMAP-West sites that were wadeable (i.e., could be sampled safely by field crews wading the stream) were used in the WSA analyses. In a 2<sup>nd</sup> survey (WSA-East), another 551 wadeable sites in the remaining 36 conterminous states were sampled during summer 2004 (USEPA 2006). In both surveys, sites were chosen on the basis of the probabilistic EMAP sampling design applied to the digital stream network depicted on 1:100,000-scale US Geological Survey (USGS) topographic maps (Herlihy et al. 2000, Stevens and Olsen 2004, Olsen and Peck 2008). The probabilistic survey design ensured that sampled sites were representative of the streams in the region surveyed so that statistically valid population percentiles could be estimated. In addition to the randomly chosen sites (probability sites), 333 hand-picked wadeable streams in the western states and 143 hand-picked streams in the eastern states were sampled during WSA in an attempt to augment the number of least-disturbed sites in the data set. Hand-picked sites were not used to make national population estimates, but both hand-picked and probability sites were used as potential reference sites if they passed reference screening criteria (see *Identifying reference sites* below). Identical field sampling and laboratory protocols were used in both surveys (Peck et al. 2006).

*WSA field and laboratory methods.*—Benthic macroinvertebrates in WSA and EMAP surveys were sampled with kick nets (0.09 m<sup>2</sup>, 595- $\mu$ m mesh) at 11 transects laid out systematically along a reach around the random sampling point (Peck et al. 2006, Hughes and Peck 2008). Samples from the 11 transects were pooled into a single composite sample, and macroinvertebrates in the composite samples were counted (target 500 count) in the laboratory. All individuals were identified to the genus level except for oligochaetes and arachnids, which were identified to family level, and nematodes and platyhelminths, which were identified to phylum.

*Identifying reference sites.*—We assembled a data set from as many reference sites as possible. Data from reference sites that had been sampled during other studies were included if the methods used to collect

the data were similar to those used for the WSA. The key characteristics of potential reference-site data were that: 1) macroinvertebrate samples were collected either from riffle or multiple habitats, 2)  $\geq 250$  individual organisms were identified from each sample, and 3) most taxonomic groups were identified to the genus level. WSA protocols specified genus-level identifications for chironomids. This requirement eliminated several data sets in which chironomids had been identified to family, subfamily, or tribe.

We used 2 approaches to identify reference sites from the assembled data (Table 1). First, we screened the physical-habitat and water-quality data from existing national or regional macroinvertebrate probabilistic-survey data sets collected in conjunction with EMAP and regional EMAP (REMAP) programs. All of these data sets included relatively similar and comparable measurements of environmental factors that could be used to screen for least-disturbed sites. We developed a series of ecoregion-specific, site-quality filters (Table 2). We removed any site that failed any one of the filters from consideration as a reference site. The filtering process and its rationale have been described for the Mid-Atlantic EMAP survey (Waite et al. 2000), several REMAP surveys (Herlihy et al. 2006), and the EMAP-West survey (Stoddard et al. 2005a). The list of screening criteria in Table 2 is not a complete list of all factors that ideally would be used to identify reference sites. For example, we did not possess direct information on hydrologic alteration, presence of invasive species, and contaminants (metals, pesticides). In this sense, our screening probably allowed some sites affected by important anthropogenic stressors to pass the screening process. We were able to minimize these potential problems by inspecting satellite and aerial photographs of western EMAP watersheds for evidence of water diversions, mining activity, and other alterations that probably would have compromised the natural integrity of a site. Although imperfect, we think the 9 disturbance factors and watershed-level screening produced a reasonably broad characterization of the physical/chemical environment at a site. We used similar criteria to screen sites sampled during EMAP-West, WSA-East, and the USGS National Water Quality Assessment (NAWQA) and compiled a database of 792 reference sites (Table 1).

Second, we obtained data directly from sources that already had identified (primarily on the basis of best professional judgment) sites in least-disturbed condition. Complete environmental data generally were not available for these sites, so we were unable to apply the screening approach used for the EMAP, WSA, and NAWQA sites. We obtained data from macroinvertebrate monitoring programs operated by state environ-

TABLE 1. Sources of macroinvertebrate reference-site data used in the US Environmental Protection Agency (EPA) Wadeable Streams Assessment (WSA). Data sources included EPA Environmental Monitoring and Assessment Program (EMAP), EMAP Western Pilot Study (EMAP-West), Regional Environmental Monitoring and Assessment Programs (REMAP), US Geological Survey National Water Quality Assessment Program (NAWQA), and Utah State University (USU). *N* = number of different reference sites from survey used in WSA analyses, RW = reach-wide composite, TR = targeted riffle, MH = multihabitat, RBP = EPA Rapid Bioassessment Protocol (Barbour et al. 1999).

Survey	<i>N</i>	Field sampling protocol	Reference <sup>a</sup>
EMAP			
Mid-Atlantic	98	RW	Klemm et al. 2003
WSA: EMAP-West	204	RW	Stoddard et al. 2005a
WSA: Eastern US	130	RW	Hughes and Peck 2008
REMAP			
New Mexico Chama/Gila basins	10	RW	Joseph 2004
Kansas/Missouri/Nebraska	9	RW	Kansas DWP 2002
NAWQA	341	MH	Moulton et al. 2002
USU (western US)	476	TR	
State agency			
Alabama	25	MH	
Illinois	70	MH	
Kentucky	80	RBP	Kentucky DEP 2002
Minnesota	55	MH	
Mississippi	21	RBP	
Missouri	23	MH	
Nebraska	12	MH	
South Carolina	30	RBP	
Tennessee	1	RBP	Arnwine and Denton 2001
Texas	54	RW	Linam et al. 2002
Vermont	16	TR	Vermont DEC 2002
Total	1655		

<sup>a</sup> Information for uncited surveys acquired by personal communication

mental agencies and from Utah State University (USU). The USU data were from reference sites throughout the western US that had been sampled during the summers of 2002 and 2003 (CPH, unpublished data). We identified 863 additional reference sites. The combined reference-site database consisted

of data from 1655 sites (Table 1). In comparison, the *total* number of probability sites sampled as part of the WSA and used to produce the final estimates of population percentiles was only 1392. The need for high-quality reference sites is significant and should not be underestimated in future surveys.

TABLE 2. Criteria for 9 chemical and physical-habitat filters used to identify least-disturbed sites in the Environmental Monitoring and Assessment Program (EMAP) Wadeable Streams Assessment (WSA) in aggregated ecoregions in the central and eastern US. Any site that exceeded any one of the listed criteria was eliminated from the reference set. EMAP physical habitat metrics are defined in Kaufmann et al. (1999). NAP = Northern Appalachians, SAP = Southern Appalachians, CPL = Coastal Plain, UMW = Upper Midwest, TPL = Temperate Plains, SPL = Southern Plains, ANC = acid neutralizing capacity, DOC = dissolved organic C, RBP = Rapid Bioassessment Protocol. – indicates filter criterion was not used in that ecoregion.

Filter criterion	NAP	SAP	CPL	UMW	TPL	SPL
Total P (µg/L)	>20	>20	>30	>30	>150	>150
Total N (µg/L)	>750	>750	>1000	>1000	>4500	>4500
Cl <sup>-</sup> (µeq/L)	>250 <sup>a</sup>	>200	–	>300	>2000	>1000
SO <sub>4</sub> <sup>2-</sup> (µeq/L)	>250	>400	>600	>400	–	–
ANC (µeq/L) + DOC (mg/L) <sup>b</sup>	<50 + <5	<50 + <5	<50 + <5	<50 + <5	<50 + <5	<50 + <5
Turbidity (NTU)	>5	>5	>10	>5	>50	>50
Mean RBP habitat score	<15	<15	<12.5	<15	<12.5	<12.5
Riparian disturbance index	>2	>2	>2	>2	>2	>2
% fine substrate	>25	>25	>50	>25	>80	>90

<sup>a</sup> Cl<sup>-</sup> criterion not applied in Northeastern Coastal Zone ecoregion 59 sites

<sup>b</sup> Filter was specific for inorganic acidity; site had to exceed both criteria to fail

Field sampling protocols differed among data sources. For example, protocols for the EMAP and WSA surveys called for a reach-wide sample that was a composite of samples from 11 equal-interval transects (Hughes and Peck 2008), whereas protocols for other surveys called for samples collected from multiple habitats (multihabitat) or from a single habitat type (targeted riffle). Most data from state surveys were obtained with the US Environmental Protection Agency rapid bioassessment protocol, which specifies either a multihabitat or single-habitat approach (Barbour et al. 1999). Most EMAP, REMAP, and WSA sites were sampled during summer, but Mid-Atlantic EMAP sites were sampled in spring. Samples for most of the other data sets, except those from Alabama and Mississippi, were collected in summer. Mississippi and Alabama samples were collected in winter (see *Reference data comparability* below for a description of our efforts to quantify and reconcile differences among sampling protocols).

We combined all data into a single master database. Misspelled taxonomic names were corrected according to information in the Integrated Taxonomic Information System (ITIS; <http://www.itis.gov/>). We used valid names in ITIS as the master list to ensure that taxon names were consistent across samples and to adjust for synonymous names and variations in upper-level taxonomic naming among data sources. We adjusted all samples with >300 individuals to a fixed 300-individual count by randomly resampling without replacement. We dropped sites from consideration as reference sites if they were represented by samples with <250 individuals. Ninety-two percent of the candidate reference sites were represented by samples with >300 individuals, and <50% of sites were represented by samples with >500 individuals.

### *Regionalization*

Appropriate expectations for reference condition must be adjusted for the natural conditions at a site. The 2 general approaches used to make such adjustments are classification (or regionalization) and direct modeling (e.g., Moss et al. 1987, Hawkins et al. 2000, Verdonchot and Nijboer 2004). Regionalization is often used to control for natural variability during development of biological indices, such as most multimetric indices, whereas the direct modeling of E (the expected number of taxa at a site) within regions is used to adjust O/E indices for site-specific differences in natural factors. We used a 2-stage approach. We developed a regionalization scheme to help account for continental-scale differences in macroinvertebrate assemblage composition associated with regional

differences in water chemistry and physical habitat. We used these ecoregions when we established reference expectations for the multimetric indices (Stoddard et al. 2008). We also used these ecoregions when we developed the RIVPACS-type predictive models used to estimate O/E (Yuan et al. 2008). Ecoregions also served as reporting units in the national assessment.

Our goal was to define ecoregions that maximized within-ecoregion similarity in macroinvertebrate assemblages and that were ecologically and geopolitically useful for reporting assessment outcomes. The number of ecoregions was constrained by the number of probability sites that could be used to draw inferences regarding stream condition within an ecoregion. The WSA survey design called for 50 sites within an ecoregion to provide a regional assessment with acceptably low uncertainty. Only 1392 probability sites were sampled during the WSA, and this number set the upper limit on the number of possible ecoregions and the lower limit on the size of an ecoregion. Enough reference sites to provide a statistically valid characterization of expectations had to be available within an ecoregion. Last, the number of ecoregions was limited by logistical problems associated with setting different condition thresholds, building separate models, and reporting conditions separately for a large number of ecoregions. These considerations led us to attempt to identify ~10 ecoregions across the country.

Many existing hydrological, ecological, and physiographic classification schemes are available for the US. We began by considering level II and level III ecoregional frameworks developed by Omernik (1987). The level II regionalization scheme divides the US into 20 ecoregions, but in several of these ecoregions (e.g., Everglades, Western Sierra Madre Piedmont) we did not have enough probability sites to assess condition. The level III scheme divides the US into 84 ecoregions, and we combined these ecoregions to define aggregated ecoregions that had enough probability sites and reference sites to enable us to draw statistically valid inferences of condition within an ecoregion. We used a 2-dimensional nonmetric multidimensional scaling (NMS) ordination based on macroinvertebrate relative abundances at all available reference sites to guide our regionalization decisions. Before ordination, we composited macroinvertebrate data across all sites in each level III ecoregion to create a single composite sample that consisted of relative abundances of the macroinvertebrate assemblage for each individual level III ecoregion. We dropped 37 level III ecoregions that had <10 sites each from the analysis, and retained 47 level III ecoregions. We dropped rare taxa (those found in

<5% of the level III ecoregions) from the analysis, and retained 501 taxa. We dropped rare taxa because we were interested in whole-assemblage patterns, and rare taxa can add noise to such analyses (McCune and Grace 2002). We applied the NMS ordination to Bray–Curtis distances calculated between all pairs of ecoregions. We used the NMS results, past experience, consultation with local benthologists, and best professional judgment to aggregate the 84 level III ecoregions into 9 aggregated ecoregions.

We used EMAP survey data to analyze the classification strength of the set of 9 aggregated ecoregions (Van Sickle and Hughes 2000). We calculated Sørensen similarity between macroinvertebrate assemblages for all possible pairs of reference sites. For each aggregated ecoregion, we calculated the difference between mean within-ecoregion similarity and the overall mean between-ecoregion similarity. Greater differences indicate greater within-ecoregion homogeneity of assemblages, and hence greater strength of the 9-ecoregion classification.

#### *Reference data comparability*

Compiling data from many different sources into a single database created the potential for data-source artifacts in the final results. Data sets differed in how samples were collected (e.g., reach-wide vs multi-habitat samples; Table 1), how many organisms were identified from each sample, sample season, taxonomic naming, and the taxonomic resolution applied to different groups. Any of these differences could produce systematic differences (artifacts) in estimates of macroinvertebrate assemblage composition and structure. These artifacts could have an adverse effect on MIBI and O/E model development and on the calculation of condition-class thresholds. We expected that the strength of these artifacts would differ depending on the methods used to assess biological condition.

The MIBI and predictive modeling approaches differ with respect to their sensitivity to specific taxonomic identities. The MIBI approach assesses biological condition by combining the values of different metrics, most of which are based on either taxon autecology (e.g., tolerance values, feeding group) or taxonomic richness within specific taxonomic groups. In contrast, the predictive modeling approach assesses biological condition by comparing the identities of taxa observed at a site to the identities that are expected to occur under reference conditions. Taxon identities are most often resolved to the genus level. We expected that the MIBI approach would be less sensitive to differences in data sources than would be predictive modeling

because the autecology of closely related taxa is often similar.

We looked for data-source artifacts that might affect the MIBI by examining box plots of some common bioassessment metric distributions (total richness; Ephemeroptera, Plecoptera, and Trichoptera [EPT] richness; Shannon diversity) across data sources within each of the WSA ecoregions. Within ecoregion, similar metric distributions among different data sources would suggest that the data-source effect was relatively weak, and that the data from multiple sources could be included in analyses. If any particular data source showed strong deviation from the overall ecoregional distribution, we would drop all sites from that data source.

We looked for data-source artifacts that might affect O/E by evaluating reference data from state agencies, USU, and NAWQA to determine whether macroinvertebrate assemblage structure in data from these sources differed systematically from data collected in EMAP/REMAP surveys. We computed an NMS ordination (on the basis of a Sørensen dissimilarity matrix) on the basis of data from EMAP/REMAP and other data sources (Table 1) within each WSA ecoregion. In the absence of data-source artifacts, macroinvertebrate assemblage compositions would be similar for both data sets at similar geographical locations, and we expected that portions of the 2 data sets would be randomly interspersed in ordination space. If the 2 data sets occupied distinct areas of ordination space, then systematic differences in assemblage composition might have arisen from differences in sampling protocols, sampling season, or stream size.

#### *Condition classes and their adjustment for reference-site quality*

For WSA reporting, the continuous MIBI and O/E index scores were separately assigned to 3 classes of assemblage condition. Good, fair, and poor condition classes were defined to represent index values that were, respectively, “not different from,” “somewhat different from,” and “markedly different from” the values at least-disturbed reference sites (Stoddard et al. 2006, Van Sickle et al. 2006, Van Sickle and Paulsen 2008). O/E and MIBI scores were assigned to condition classes by comparing the scores to percentiles of the distribution of scores observed at reference sites. For both indices, lower scores indicate greater departure from reference conditions. Thus, a site at which the indicator score was <5<sup>th</sup> percentile of the distribution of reference-site scores was classified as in poor condition, a site at which the indicator score was

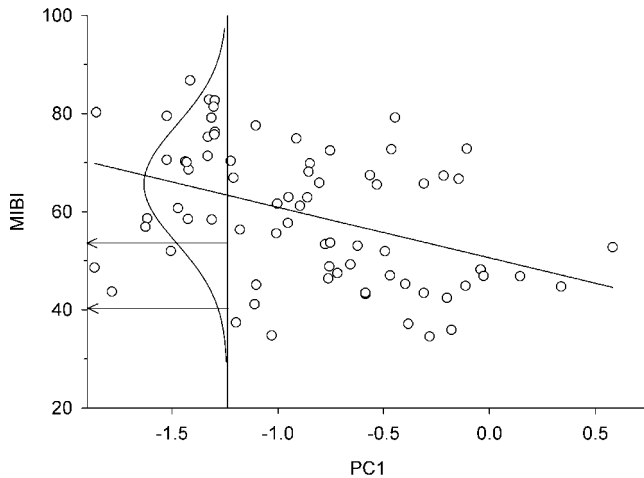


FIG. 1. Example of adjusting the multimetric index of biotic integrity (MIBI) reference distribution within an aggregated ecoregion to represent higher-quality conditions. Vertical line is the 25<sup>th</sup> percentile of the disturbance-factor principal components analysis axis-1 (PC1) scores for the ecoregion. The sloping line is the regional regression of MIBI on PC1. The bell-shaped curve is the regression error distribution centered at the point predicted from the 25<sup>th</sup> percentile of PC1, with standard deviation equal to the pooled residual standard error for the regression. The 25<sup>th</sup> percentile (upper horizontal arrow) and 5<sup>th</sup> percentile (lower horizontal arrow) of the error distribution identify thresholds for fair/good and poor/fair condition classes, respectively.

>5<sup>th</sup> and <25<sup>th</sup> percentile was classified as in fair condition, and a site at which the indicator score was >25<sup>th</sup> percentile was classified as in good condition. Class descriptors have no regulatory meaning; they were chosen to facilitate communication with the public and policy makers.

When we compiled our reference-site database, we included sites that were selected by a variety of methods and by a variety of people. In addition, expectations of reference-site condition and the ease with which reference sites were found varied among different parts of the country. Thus, we expected that the quality of reference sites might vary substantially among and within ecoregions. Therefore, we estimated class thresholds (5<sup>th</sup> and 25<sup>th</sup> percentiles of reference distributions) separately for each aggregated ecoregion.

We used screening criteria to select many of our reference sites (Table 2), but these screening criteria were a compromise between specifying conditions that would ensure high-quality sites and conditions that were lenient enough to provide enough sites to permit statistical estimates of reference expectations. The criteria selected enough sites to estimate reference variability statistically, but selected sites probably

varied in the degree to which they represented minimally impaired reference conditions. We partially corrected for these variations in reference-site quality by applying a post hoc adjustment to specify condition-class thresholds for biotic index values that represented the higher-quality reference conditions within each ecoregion.

We did a principal components analysis (PCA) of the 9 disturbance variables that were used to screen for reference condition (Table 2), but we substituted pH for acid neutralizing capacity to avoid the problem of transforming negative values. The PCA was based on data from 250 of the 334 sites sampled during WSA (Table 1) at which we had complete data for all 9 variables. We  $\log_{10}(x)$ -transformed all variables except pH, which is log-transformed by definition, and % fine substrate, which we  $\arcsine\sqrt{(x)}$ -transformed. We interpreted the 1<sup>st</sup> principal component (PC1) of the PCA as a generalized disturbance gradient. Lower PC1 scores represented lower disturbance. We selected the theoretical 25<sup>th</sup> percentile (assuming a normal distribution) of each ecoregion's PC1 scores to represent a benchmark of higher-quality (lower disturbance) reference conditions for that region (Fig. 1).

We regressed the raw MIBI scores for WSA reference sites on PC1. Low sample sizes in some aggregated ecoregions (<15 sites) did not permit separate regional regressions. Therefore, we carried out a single regression on all 250 sites, using regional dummy variables (Montgomery et al. 2001) to allow some degree of aggregated ecoregional specificity in intercepts and slopes of the PC1 effect. We applied the fitted regression model separately within each aggregated ecoregion, and we assumed that the error distribution of the theoretical model at the 25<sup>th</sup> percentile of PC1 would characterize the distribution of reference-site index scores at higher-quality sites within the aggregated ecoregion (Fig. 1). This normal distribution of higher-quality index scores has a mean equal to the model-predicted index at the 25<sup>th</sup> percentile of PC1, and we assumed that its SD was given by the pooled residual standard error for the model. We used the 5<sup>th</sup> and 25<sup>th</sup> percentiles of this higher-quality index distribution as poor/fair and fair/good thresholds, respectively, for the index condition classes used in the WSA (Fig. 1).

## Results

### Regionalization

On the basis of the NMS ordination of level III ecoregions (Fig. 2) and other information, we aggregated the 84 level III ecoregions into 9 WSA aggregated ecoregions (Table 3, Fig. 3). The 2-dimen-

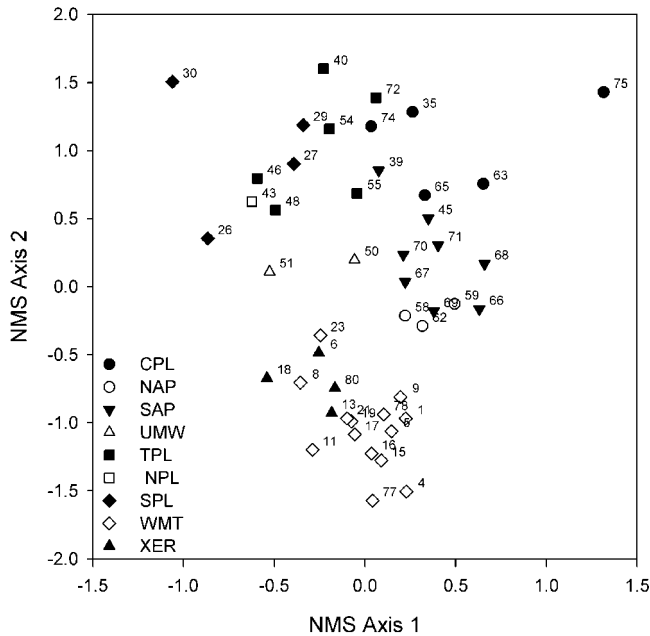


FIG. 2. Nonmetric multidimensional scaling (NMS) ordination plot of aggregated Omernik level III ecoregion reference-site macroinvertebrate data. Numbers in the figure correspond to the level III ecoregion code (Table 3). Symbols indicate Wadeable Stream Assessment aggregated ecoregion classification of the Omernik level III ecoregion.

sional NMS ordination of level III ecoregions (stress = 11.8) explained 88% of the variation in distances among ecoregions. In general, clusters of level III ecoregions formed relatively distinct and coherent groups (Fig. 2). The level III ecoregions that made up the Xeric West and Western Mountain aggregated ecoregions clustered by themselves in the lower left quadrant of the NMS plot with the Western Mountain ecoregions mostly in  $\frac{1}{2}$  of the cluster. With the exception of the Ozark Highlands (ecoregion 39), the level III ecoregions that made up the Northern and Southern Appalachians aggregated ecoregions were strongly clustered in the middle of the ordination plot, with the Northern Appalachians ecoregions in one corner of the cluster. The level III ecoregions that made up the Northern, Southern, Temperate, and Coastal Plain aggregated ecoregions were all in the upper portion of the ordination plot, with Coastal Plain ecoregions on the right, the more arid Northern and Southern Plains ecoregions on the left, and the Temperate Plains ecoregions in the middle.

Many of the aggregation decisions shown in Fig. 3 were fairly straightforward. For example, past analyses suggested that macroinvertebrate assemblages differed little among the mountainous ecoregions in western Oregon (Cascades, Coast Range, and Klamath Mountains) (Herlihy et al. 2005). A few ecoregions

presented greater challenges. For example, the Piedmont ecoregion in the southeastern US is intermediate in topography in between the flat Atlantic Coastal Plain to the east and the Appalachian Mountains to the west. The Piedmont ecoregion was too small to define as a separate ecoregion in the WSA, so it had to be combined with either the Southern Appalachians ecoregion or the Coastal Plain ecoregion. After consulting with local experts and examining ordination plots of macroinvertebrate data, we decided to combine the Piedmont (e.g., ecoregion 45; Fig. 2) with the Southern Appalachians ecoregion.

For the 9 aggregated ecoregions, the average between-ecoregion similarity of reference-site assemblages was 0.24 on a Sørensen similarity scale of 0 (no taxa in common) to 1 (all taxa in common). Average within-ecoregion similarities exceeded the between-ecoregion similarity by minima of 0.02 and 0.03 (Northern Plains and Southern Plains ecoregions, respectively) and maxima of 0.17 and 0.14 (Northern Appalachians and Western Mountains ecoregions, respectively). The overall classification strength was 0.10.

#### Reference-site comparability

In general, we found no strong effect of data source on metric values. For example, in the Northern Plains, EPT richness distributions were similar among sites in the USU, EMAP, and Nebraska data sets (Fig. 4A). However, EPT richness was significantly higher at Southern Appalachians (Fig. 4B) and Coastal Plain (not shown) sites in the South Carolina state data set than at sites in the other data sets. We removed the South Carolina state data set from subsequent analyses to avoid the possibility of bias. We did not use the South Carolina state data for model development for the MIBI or O/E, nor did we use it to set condition-class thresholds in the WSA. We did not find any evidence to justify removing any other data sources from the MIBI analyses, and all other reference data ( $n = 1625$ ) were used to construct the WSA MIBI.

We found strong effects of data source on assemblage composition. In most ecoregions, sites from state data sets tended to have different assemblage composition than did sites from WSA data sets where sampling was done with EMAP protocols. The NMS ordination for sites in the Temperate Plains ecoregion is a typical example (Fig. 5). In contrast, sites from the USU and NAWQA data sets and sites from data sets obtained with EMAP sampling protocols generally occupied similar regions in ordination space. Therefore, only EMAP, REMAP, USU, and NAWQA data

TABLE 3. Aggregation scheme used in the Wadeable Streams Assessment to define 9 aggregated ecoregions (abbreviations in parentheses) from Omernik (1987) level III ecoregions (codes in parentheses).

Coastal Plain (CPL)	Northern Appalachians (NAP)	Southern Appalachians (SAP)	Upper Midwest (UMW)
East Central Texas Plains (33)	Northeastern Highlands (58)	Piedmont (45)	Northern Minnesota Wetlands (49)
South Central Plains (35)	North Central Appalachians (62)	Northern Piedmont (64)	Northern Lakes and Forests (50)
Southeastern Plains (65)	Laurentian Plains and Hills (82)	Interior Plateau (71)	North Central Hardwood Forests (51)
Mississippi Valley Loess Plains (74)	Northeastern Coastal Zone (59)	Ouachita Mountains (36)	
Middle Atlantic Coastal Plain (63)	N. Appalachian Plateau and Uplands (60)	Arkansas Valley (37)	
Mississippi Alluvial Plain (73)	Erie Drift Plain (61)	Boston Mountains (38)	
Southern Coastal Plain (75)	E. Great Lakes and Hudson Lowlands (83)	Ozark Highlands (39)	
Atlantic Coastal Pine Barrens (84)		Blue Ridge (66)	
Western Gulf Coastal Plain (34)		Ridge and Valley (67)	
Southern Florida Coastal Plain (76)		Southwestern Appalachians (68)	
		Central Appalachians (69)	
		Western Allegheny Plateau (70)	

TABLE 3. Extended

Temperate Plains (TPL)	Northern Plains (NPL)	Southern Plains (SPL)	Western Mountains (WMT)	Xeric West (XER)
Southeastern Wisconsin Till Plains (53)	Northwestern Glaciated Plains (42)	Nebraska Sand Hills (44)	Blue Mountains (11)	Columbia Plateau (10)
Central Corn Belt Plains (54)	Northwestern Great Plains (43)	High Plains (25)	Northern Rockies (15)	Snake River Plain (12)
Eastern Corn Belt Plains (55)		Southwestern Tablelands (26)	Idaho Batholith (16)	Central Basin and Range (13)
Huron/Erie Lake Plains (57)		Central Great Plains (27)	Middle Rockies (17)	Wyoming Basin (18)
Interior River Valleys and Hills (72)		Cross Timbers (29)	Wasatch and Uinta Mountains (19)	Colorado Plateaus (20)
Central Irregular Plains (40)		Edwards Plateau (30)	Southern Rockies (21)	Arizona/New Mexico Plateau (22)
Northern Glaciated Plains (46)		Texas Blackland Prairies (32)	Cascades (4)	Northern Basin and Range (80)
Western Corn Belt Plains (47)		Southern Texas Plains (31)	Canadian Rockies (41)	Mojave Basin and Range (14)
Lake Agassiz Plain (48)			Blue Mountains (11)	Sonoran Basin and Range (81)
Flint Hills (28)			Northern Rockies (15)	Chihuahuan Deserts (24)
			Idaho Batholith (16)	California Chaparral and Oak Woodlands (6)
			Sierra Nevada (5)	Central California Valley (7)
			North Cascades (77)	Madrean Archipelago (79)
			Klamath Mountains (78)	
			E. Cascades Slopes and Foothills (9)	
			Coast Range (1)	
			Puget Lowland (2)	
			Willamette Valley (3)	

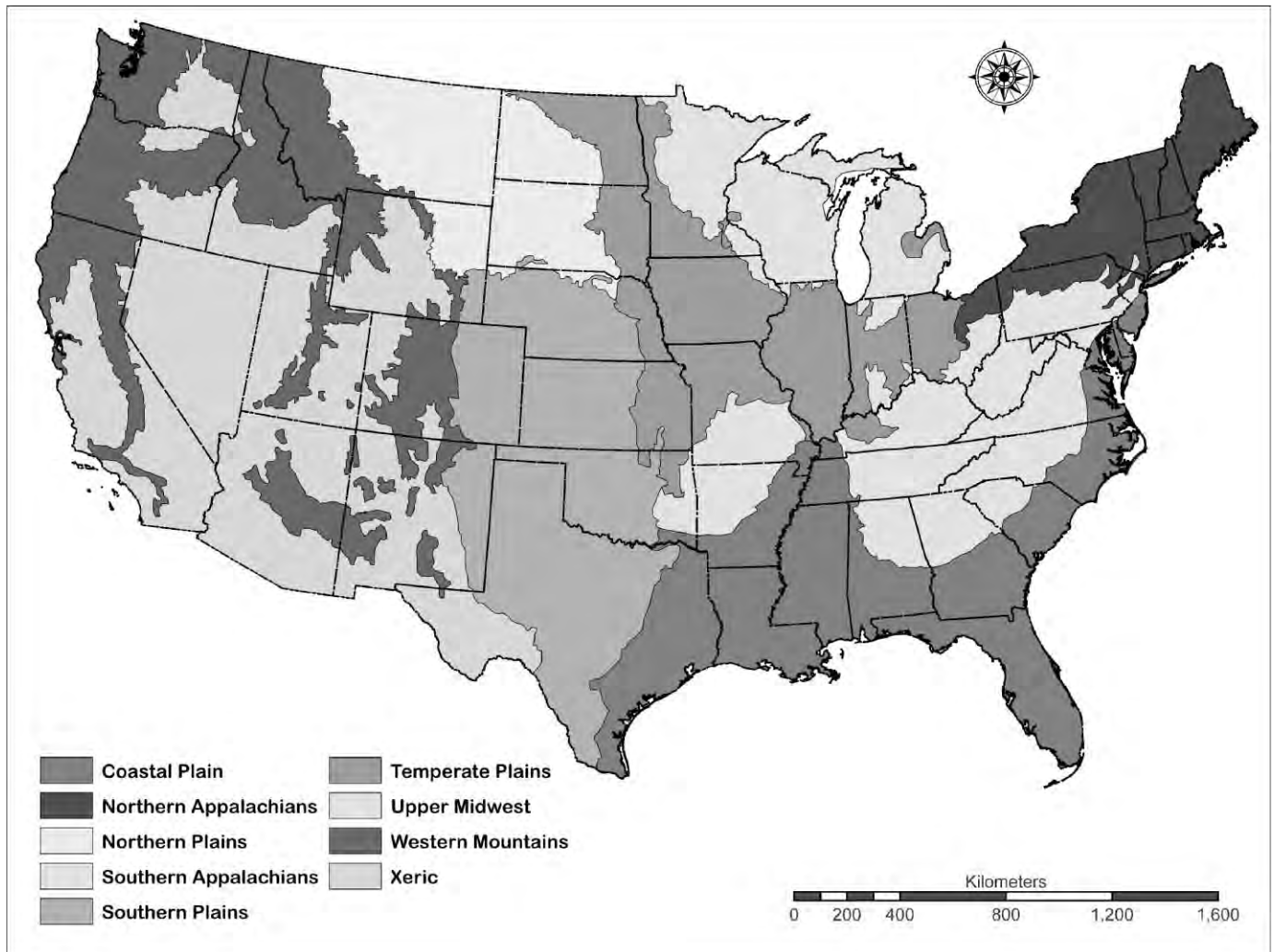


FIG. 3. Nine aggregated ecoregions (Table 3) used in the Wadeable Streams Assessment.

sets were used to supply reference sites for predictive model development.

#### *Condition classes and their adjustment for reference-site quality*

Condition-class thresholds for MIBI and O/E indices varied substantially across the 9 aggregated regions (Table 4). The range of unadjusted MIBI 5<sup>th</sup>-percentile thresholds estimated directly from the full reference-site database (excluding South Carolina state sites; Table 1) was 31 MIBI units (15–46) across the 9 regions (Table 4). O/E 5<sup>th</sup>-percentile thresholds ranged from 0.82 in the Northern Appalachians to 0.38 in the Temperate Plains. This variability and the low thresholds in the Temperate and Southern Plains motivated our attempts to examine the quality of reference sites in more depth.

PC1 was highly correlated with 8 of the 9

disturbance variables used in the analysis and explained 44% of the variance in the stressors (Table 5). The 9<sup>th</sup> variable, pH, was strongly correlated with PC2. Acidification issues generally are confined to a small percentage of streams in specific parts of the eastern US (Kaufmann et al. 1991). Therefore, we focused on PC1 as an overall index of disturbance. The variables that loaded positively on PC1 were those that increase with increased disturbance (e.g., nutrients, turbidity). The single variable that loaded negatively on PC1 (rapid habitat assessment score) is an inverse variable, and higher scores indicate less disturbance. Thus, higher PC1 scores indicated overall increased disturbance.

MIBI reference distributions were biased downward in >1 aggregated ecoregion, i.e., they included relatively more disturbed sites with lower MIBI values. Over the 9 aggregated ecoregions combined, MIBI

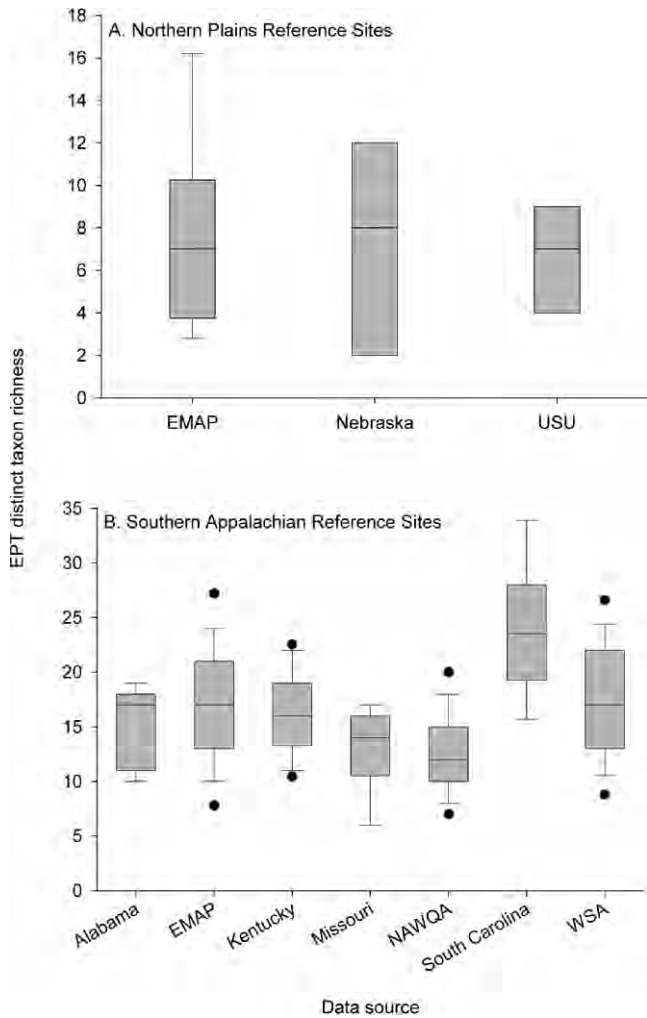


FIG. 4. Box-and-whisker plots of Ephemeroptera, Plecoptera, Trichoptera (EPT) distinct taxon richness scores by data source for the Northern Plains (A) and Southern Appalachians (B) aggregated ecoregions. Lines in boxes show ecoregion medians, boxes show interquartile ranges, and whiskers show 5<sup>th</sup> and 95<sup>th</sup> percentiles. EMAP = Environmental Monitoring and Assessment Program, NAWQA = National Water Quality Assessment Program, WSA = Wadeable Streams Assessment, USU = Utah State University.

scores decreased weakly with increasing PC1 scores (Pearson correlation,  $r = -0.32$ ,  $p < 0.0001$ ; Fig. 6A). PC1 scores were higher at all but one of the Temperate Plains sites than at Southern Appalachians sites (difference in ecoregional mean score = 2.0 PC1 units, 95% CI: 1.8–2.2; Fig. 6B). The mean MIBI score for sites in the Southern Appalachians also exceeded the mean MIBI score for sites in the Temperate Plains by 8.1 units (95% CI: 0.2–16.0). Within-region variation in MIBI and PC1 scores also was substantial (Fig. 6B), but both ecoregions show declining MIBI scores with increasing disturbance (PC1).

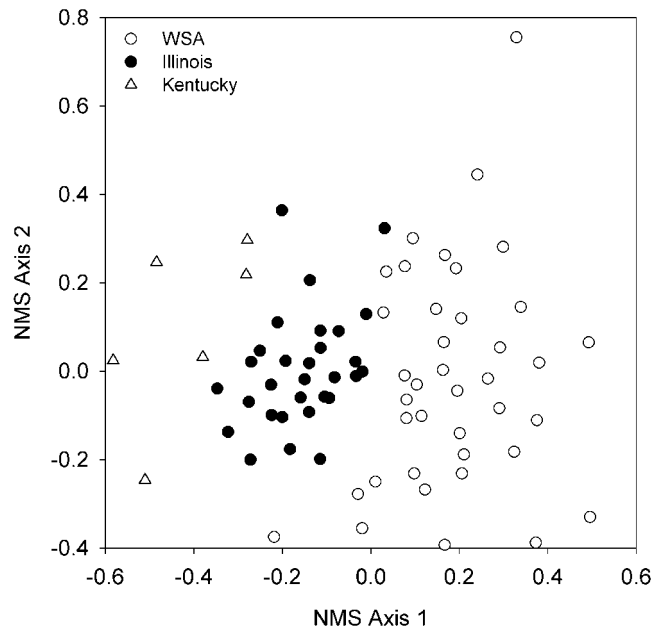


FIG. 5. Nonmetric multidimensional scaling (NMS) ordination plot contrasting macroinvertebrate reference-site data from Indiana, Kentucky, and Wadeable Streams Assessment (WSA) data sources for the Temperate Plains aggregated ecoregion.

We chose a single parsimonious model for PC1 to explain as much MIBI variance as possible. This model specified different intercepts for each aggregated ecoregion, a PC1 slope of 0 for 4 aggregated ecoregions, a common non-0 PC1 slope for 4 other aggregated ecoregions, and a separate slope for the remaining aggregated ecoregion (Table 4). The residual standard error for the model was 14.02.

We illustrate the application of the regression model to adjust the MIBI reference distribution in the Western Mountains aggregated ecoregion (Fig. 1). The mean and SD of PC1 scores in that ecoregion were  $-0.87$  and  $0.54$ , respectively. These values imply a theoretical 25<sup>th</sup>-percentile PC1 score of  $-1.24$ , assuming normality (mean  $- 0.68[SD]$ ). Inserting this theoretical score into the regression model for the Western Mountains (Table 4) yields a predicted MIBI of  $(51.0 - 9.8[-1.24]) = 63.2$ , which we defined as the mean of the adjusted MIBI distribution for higher-quality conditions. We also assumed that the adjusted MIBI distribution was normally distributed with  $SD = 14.02$ , the model's residual standard error. The 25<sup>th</sup> percentile of the reference-sites scores for this distribution, assuming normality, is  $(63.2 - 0.68[14.02]) = 53.7$ , and a similar calculation yields the 5<sup>th</sup> percentile. The same approach was followed for the other 4 regions with non-0 PC1 slopes. In the remaining 4 regions, we found no evidence of a PC1 effect on MIBI (0 PC1 slopes; Table 4), and we made no upward adjustment for quality. In

TABLE 4. Percentiles of multimetric index of biotic integrity (MIBI) and observed/expected (O/E) scores at Wadeable Streams Assessment reference sites for each aggregated ecoregion. Percentiles were used as thresholds to separate poor/fair (5<sup>th</sup> percentile) and fair/good (25<sup>th</sup> percentile) condition classes. MIBI percentiles are shown before and after adjustment for site quality. Intercepts and slopes are given for the MIBI adjustment regression model ( $R^2 = 0.261$ , residual standard error = 14.02, overall model  $F_{10,239} = 8.46$ ,  $p < 0.0001$ ). O/E and adjusted MIBI percentiles are theoretical percentiles calculated from the ecoregional mean and standard deviation. Unadjusted MIBI percentiles are empirical percentiles, estimated without distributional assumptions. MIBI condition classes were based on adjusted percentiles. The O/E modeling used a slightly different subset of reference sites than the MIBI (see Yuan et al. 2008 for details).  $N$  = number of reference sites used in the analyses.

Aggregated ecoregion	O/E			Unadjusted MIBI			Adjusted MIBI			Coefficients for MIBI adjustment	
	5 <sup>th</sup>	25 <sup>th</sup>	$N$	5 <sup>th</sup>	25 <sup>th</sup>	$N$	5 <sup>th</sup>	25 <sup>th</sup>	$N$	Intercept	Slope
Coastal Plain	0.74	0.93	10	36	54	113	42	56	14	65.5	0
Northern Appalachians	0.82	0.93	25	46	61	141	49	63	26	72.5	0
Northern Plains	0.70	0.93	42	30	41	28	41	55	14	69.7	-9.8
Southern Appalachians	0.66	0.83	31	32	46	389	37	51	35	53.4	-9.8
Southern Plains	0.48	0.71	18	15	34	77	30	44	16	57.6	-9.8
Temperate Plains	0.38	0.94	38	24	44	162	31	45	32	76.1	-16.9
Upper Midwest	0.58	0.81	12	37	54	80	34	48	12	57.5	0
Western Mountains	0.73	0.91	157	31	49	464	40	54	73	51.0	-9.8
Xeric West	0.64	0.84	36	36	52	171	40	53	28	62.9	0

the 4 0-slope regions, the MIBI reference distribution was assumed to be normal, with a mean given by the regression intercept and  $SD = 14.02$ .

The quality adjustment increased MIBI condition-class thresholds, relative to unadjusted thresholds, in 8 of the 9 regions (Table 4). Upward adjustments ranged from 2 to 15 MIBI points. In the Upper Midwest ecoregion, the adjustment reduced the thresholds slightly, even though the region had a 0 PC1 slope. This result occurred because reference-site MIBI scores in the Upper Midwest ecoregion had an observed  $SD < 14.02$ , the value used for the adjusted regional distribution. Across the 9 ecoregions, the adjusted MIBI 5<sup>th</sup>-percentile thresholds ranged across 19 MIBI units (30–49).

Condition-class thresholds for O/E scores were not adjusted for reference-site quality because O/E scores

TABLE 5. Correlation of disturbance variables with principal components analysis (PCA) axis-1 (PC1) and -2 (PC2) scores for Wadeable Streams Assessment reference sites ( $n = 250$ ).

Metric	PC1	PC2
Total P	0.765	-0.077
Total N	0.770	-0.284
Cl <sup>-</sup>	0.779	0.140
SO <sub>4</sub> <sup>2-</sup>	0.734	0.445
pH	0.181	0.905
Turbidity	0.758	-0.435
RBP habitat score	-0.526	-0.017
Riparian disturbance index	0.675	0.028
% fine substrate	0.564	-0.035
% variance explained	44.3	14.6

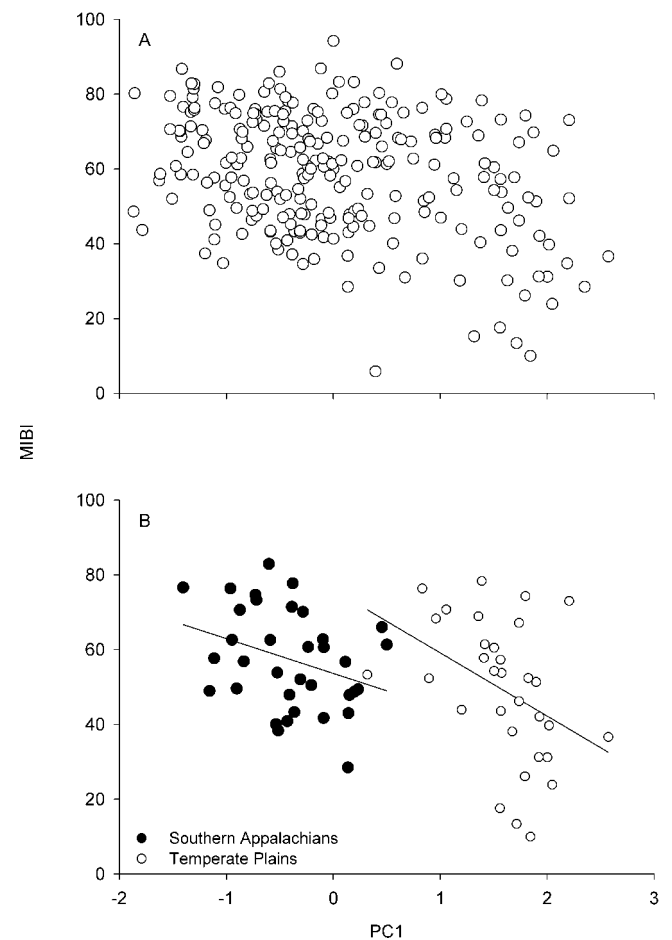
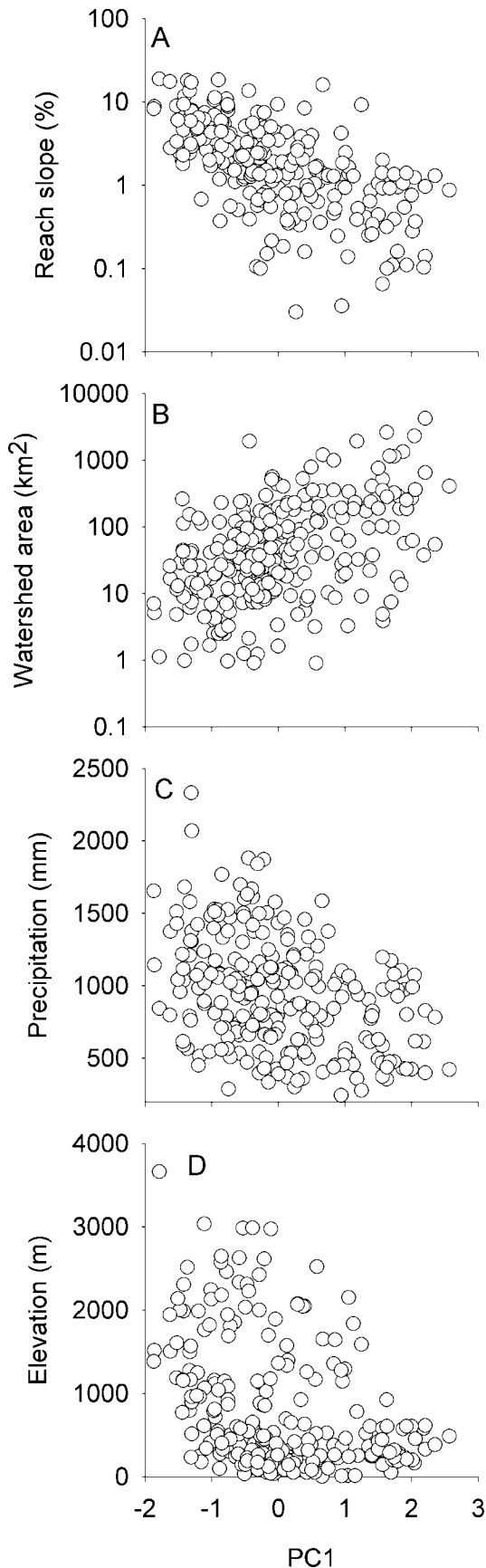


FIG. 6. Multimetric index of biotic integrity (MIBI) scores vs principal components analysis axis-1 (PC1) disturbance scores for 250 Wadeable Streams Assessment reference sites having complete data for disturbance variables (A), and reference sites in 2 example aggregated ecoregions (B).



were not related to PC1 scores ( $r^2 = +0.0053$ ). In addition, the O/E-PC1 correlations were not statistically significant ( $p < 0.05$ ) in any of the 9 aggregated ecoregions. O/E condition-class thresholds were based on theoretical 5<sup>th</sup> and 25<sup>th</sup> percentiles of distributions of scores at WSA sites only. Thresholds were calculated separately for each ecoregion, assuming a normal distribution and the ecoregion's mean and SD of reference-site O/E scores (Table 4).

The magnitude of disturbance at a site, as represented by PC1, covaried with naturally occurring factors. For example, PC1 was correlated with reach slope (Pearson correlation,  $r = -0.49$ ; Fig. 7A), watershed area ( $r = 0.38$ ; Fig. 7B), annual precipitation ( $r = -0.35$ ; Fig. 7C), and elevation ( $r = -0.36$ ; Fig. 7D). However, these correlations were not evident within some aggregated ecoregions. For example, watershed area and PC1 were not significantly correlated ( $r < 0.1$ ) at sites within the Coastal Plain or Southern Appalachians ecoregions, and reach slope and PC1 were not significantly correlated at sites within the Upper Midwest. On the other hand, correlations between naturally occurring variables and PC1 were much stronger at sites within some aggregated ecoregions. For example, elevation and PC1 were strongly negatively correlated in the Northern Plains ( $r = -0.82$ ).

## Discussion

### Regionalization

Partitioning the effects of natural factors from the effects of anthropogenic stressors is a critical component of nearly all bioassessment programs. Such partitioning should result in more accurate and precise specification of the reference condition for each assessed site. Modeling of continuous environmental gradients, as is done in RIVPACS-type modeling, is perhaps the most direct approach to specifying site-specific reference conditions (e.g., Moss et al. 1987). However, an equally, if not more, common approach is to assign sites to classes on the basis of either spatially defined ecoregions (e.g., Omernick 1987, Moog et al. 2004, Metzeling et al. 2006) or reach-level ecotypes (e.g., Ehlert et al. 2002, Verdonschot and Nijboer 2004, Sandin and Verdonschot 2006, Hatton-Ellis 2008, Turak and Koop 2008).

FIG. 7. Relationship between principal components analysis axis-1 (PC1) disturbance scores for 250 Wadeable Streams Assessment reference sites and the slope of the sample reach (A), watershed area (B), mean annual precipitation (C), and elevation (D).

Our use of large-scale ecoregions was useful in partitioning some of the marked variation in stream invertebrate assemblages that occurred at a continental scale (shown by the NMS ordination and classification strength analyses). Such regionalization appears to be most effective for accounting for biotic variation between ecoregions that differ markedly in topography and climate (cf. Feminella 2000, Hawkins et al. 2000, Hawkins and Vinson 2000, Rabeni and Doisy 2000, Waite et al. 2000, Sandin and Verdonschot 2006) and between ecoregions that differ significantly in watershed geology (Moog et al. 2004). The (within-between) differences in mean assemblage similarity for the 9 aggregated ecoregions were very similar to within-between differences that have been observed for other, smaller ecoregions (Hawkins et al. 2000, Hawkins and Vinson 2000). The relatively small within-between difference observed in the Northern and Southern Plains ecoregions might be a manifestation of streams that are inherently more temporally variable than are mountain streams (e.g., Northern Appalachians and Western Mountains) or a result of more variable reference-site quality in the plains ecoregions.

#### *Reference-site comparability*

Our test for comparability among sources of reference-site data used to develop the MIBI arose directly from the assumption that reference-site metric values were relatively homogeneous among sites within aggregated ecoregions. Thus, a simple comparison of the distributions of metric values across data sources was sufficient to establish whether different data sets were comparable. These analyses led us to remove only one state data set when constructing the WSA MIBI. Further investigation found differences in laboratory counting protocols that would be expected to yield higher numbers of EPT taxa than would EMAP protocols.

Data comparability requirements are more stringent when O/E predictive models are used to assess biological condition because the models compare the identity of taxa observed at test sites with the expected identity of taxa predicted from comparable reference sites. Thus, the comparability of the taxonomic composition of full assemblage data must be considered across data sources. Furthermore, predictive models make no assumptions regarding the homogeneity of data within ecoregions. Instead, variations in assemblage composition are modeled across continuous natural gradients. Thus, to test for comparability of data for use in predictive models, we considered whether assemblage composition from different data

sources occupied similar or different locations in ordination space. Our results suggested that most state data sets differed systematically from EMAP data sets, but the reasons for these differences were not clear. Some differences undoubtedly stemmed from differences in the natural characteristics of streams that were sampled (e.g., stream size) and the time of sampling among the different data surveys. In theory, these differences could have been represented accurately in the predictive model. However, differences in sampling protocols also might have contributed to the compositional differences between state and EMAP data, and these differences could not be modeled. We chose to omit the state data sets from the reference-site database used to develop the predictive model.

Assemblage composition at reference sites in the NAWQA data set did not appear to differ systematically from that at reference sites in the WSA data set. However, the spatial extent of the NAWQA data was much larger (covering all of the eastern US) than the spatial extent of state data sets. Therefore, the NMS test based on the NAWQA data set was not as robust as the tests based on the smaller-scale state data sets. That is, the range of natural environmental heterogeneity spanned by the NAWQA and EMAP reference sites in ordination space could have been large enough to mask differences that might have been apparent under a more controlled set of environmental conditions. We decided to include the NAWQA data in predictive model development for 2 reasons. First, we had no strong empirical evidence that NAWQA samples differed systematically from those collected by EMAP surveys. Second, inclusion of NAWQA samples allowed us to build predictive models for areas in which we would otherwise have had too few samples. Subsequent analyses confirmed the general validity of including NAWQA data with some caveats. Carlisle and Hawkins (2008) refined the predictive model used in the WSA for the western US by adding data from 88 additional NAWQA reference sites. Comparison of mean reference-site O/E values between western NAWQA and EMAP/USU samples showed that the mean O/E values for the 2 sets of sites did differ in a manner consistent with a difference in field sampling protocols. NAWQA protocols require sampling a larger area of the site than do EMAP and USU sampling protocols. E was thus underestimated for the NAWQA samples. However, the bias was small, and NAWQA site scores could be adjusted to remove this bias.

#### *Adjusting for reference-site quality*

Our PCA-based regression model compensated for some of the within- and between-aggregated ecoregion

variation observed in reference-site quality (Fig. 6). The model adjusted the MIBI condition-class thresholds upward in 8 of the 9 aggregated ecoregions, thereby reflecting a higher-quality subset of reference conditions within each ecoregion. Model adjustments also partly compensated for between-region differences in reference-site quality because the range of 5<sup>th</sup>-percentile thresholds across ecoregions was reduced from 31 MIBI units (before adjustment) to 19 units (after adjustment). The range of 25<sup>th</sup>-percentile thresholds was similarly reduced from 37 units (before adjustment) to 19 units (after adjustment) (Table 4).

PC1-based adjustments reduced the absolute between-region differences in condition-class thresholds for MIBI, but substantial differences still remain for MIBI and for O/E (Table 4). For this reason, the assessed MIBI or O/E condition class (good, fair, or poor) for any site must be interpreted in relative terms, as reflecting similar, somewhat different, and very different, respectively, condition from the least-disturbed conditions within the aggregated ecoregion that encompasses that site. Our condition classes are relative measures *within ecoregions* and cannot be used to provide a meaningful between-ecoregion comparison of biological condition. Further research is needed to harmonize regional differences in reference-site quality, so that useful between-ecoregion comparisons of condition can be made.

Our adjustment model depended on the assumption that PC1 accurately represented a gradient of human disturbance and that lower PC1 values corresponded to lower levels of disturbance. This assumption might be true, but further research is required to verify that such statistical constructs do indeed correspond with a human disturbance gradient for different parts of the country. Some of the variation of environmental conditions along PC1 might originate from natural variability among stream types (Fig. 7A–D). Thus, we might have confounded human disturbance with natural variability when we set benchmarks by adjusting thresholds to modeled 25<sup>th</sup> percentiles on the basis of PC1 scores.

Conversely, the adjustment for reference-site quality might have *underestimated* reference conditions. In parts of the US, stream biological communities probably have undergone substantial losses of fauna, and pristine or minimally disturbed systems no longer exist. Thus, our estimates of reference condition from least-disturbed sites might bear little resemblance to pristine conditions. Our final estimates of biological condition might still be biased by differences in the degree of pervasive human alteration of the landscape across the country because the adjustment for refer-

ence-site quality that we used was bounded by the data in our reference-site database.

One of the most contentious issues regarding the results of the WSA was that assessments of individual sites and regional inferences derived from those site assessments might not agree with state and local assessments of biotic conditions (e.g., Ode et al. 2008). Such differences between assessments can be caused by use of different reference criteria, use of differentially sensitive indicators, or use of sampling methods that differ in how effectively various taxa are collected. The biological condition gradient (BCG) (Davies and Jackson 2006) was developed as a conceptual framework to help managers map values of different indicators onto a commonly understood model of how sets of taxa (and thus, indicators based on those taxa) respond to stress. Use of the BCG requires that managers be able to judge how the quality of the reference sites used in indicator development and application compares with that of truly natural sites. If we know the quality of reference sites in terms of their location along the BCG, we should be able to adjust assessments to a common benchmark, i.e., the natural condition. For the BCG to be practically useful, we must quantify the degree to which reference sites differ from the ideal natural benchmark. However, quantifying the quality of reference sites is difficult (as we have demonstrated). The application of PCA model to the WSA reference sites was an important step in the direction of quantifying reference-site quality, but we made no attempt to factor out the effects of any natural factors that might co-occur with the stressors examined. Doing so might improve our ability to establish more uniform reference conditions across regions. Considerable work remains to be done on this issue, especially in regard to understanding the degree to which natural and stressor variables are confounded and to what extent we can separate their effects on indicator response.

Confounding between natural and disturbance gradients within reference sites could be minimized in at least 2 ways. First, more sophisticated modeling of the effects of natural gradients might improve our ability to set reference expectations at individual sites (e.g., Cao et al. 2007, Cutler et al. 2007). This approach will require a large number of reference sites so that models have enough degrees of freedom to detect strong and subtle relationships between taxa and the many environmental gradients that vary at different spatial scales. Second, if modeling applied to adjust for natural gradients during index development is not fully successful, post hoc modeling of index values on the natural gradients might allow adjustment for remaining biases.

*Synthesis and concluding thoughts*

Our work with the WSA reference-condition approach showed 2 distinct problems with variable reference-site quality. First, enough uniformly good-quality reference sites is difficult to find in some regions. Second, some regions are more degraded (across the whole range) than others. We attempted to ameliorate the 2<sup>nd</sup> problem by developing ecoregion-specific MIBI and O/E models and ecoregion-specific good/fair/poor class thresholds that were adjusted for reference-site quality. A consequence of setting ecoregion-specific thresholds (and models) is that one loses some ability to compare directly the overall biological condition of sites in different ecoregions. Each ecoregion is graded against its own least-disturbed condition, which might differ greatly from those in other ecoregions. In the future, differences in regional definitions of least-disturbed condition could be addressed by placing them in context with the BCG or by narrative descriptions.

The amount of variability of reference-site conditions within aggregated ecoregions could have been reduced by using more and smaller ecoregions. Increasing the number of ecoregions also might have reduced the degree to which human disturbance and natural gradients were confounded because the magnitude of natural variability within smaller ecoregions would have been reduced. However, our choice of ecoregion size was constrained by the number of available reference sites. Use of more ecoregions will require more ecoregional reference sites.

Over the last 2 decades, significant progress has been made toward development of the conceptual underpinnings for a reference-site approach to the assessment of freshwater ecosystems and application of that approach to real-world management issues (Hughes et al. 1986, Moss et al. 1987, Brinson and Rheinhardt 1996, Reynoldson and Wright 2000, Bailey et al. 2004, Chessman and Royal 2004, Sandin and Verdonschot 2006, Stoddard et al. 2006, Cao et al. 2007). However, major challenges remain in refining the reference-condition approach. These challenges fall into 2 general categories. First, reliable ways must be developed to adjust for variation in reference-site quality, especially given that the best conditions in some regions are far from historical conditions. Second, statistical and other analytical tools must be developed that can accurately and precisely estimate reference condition at specific sites. We have taken small but significant steps toward addressing both of these challenges. The WSA provided an unparalleled opportunity to push the limits of our conceptual and

technical understanding of how to apply a reference-condition approach to a real-world need. Our hope is that we have learned enough from this exercise that we will be able to further improve the technical quality of the next round of national assessments.

**Acknowledgements**

This work was funded by grants R-829498-01 (ATH), R-828637-01 (CPH), and R-830594-01 (CPH) from the National Center for Environmental Research (NCER) Science to Achieve Results (STAR) Program of the US Environmental Protection Agency (EPA) and cooperative agreement CR831682-01 between Oregon State University and the US EPA National Health and Environmental Effects Research Laboratory—Western Ecology Division. We thank the many people, including Seva Joseph, Daren Carlisle, Vickie Hulcher, Mike Compton, Greg Pond, Robert Murzyn, Debbie Baker, Natalie Guedon, Jim Glover, Debbie Arnwine, Charles Bayer, Tom Archdeacon, Ben Rich, Steve Fisk, Jennifer Pitt, and Dave Peck, who graciously compiled and shared their data and metadata for use in our study. We thank all the people involved with the EMAP-West and the WSA for their insights and work in collecting the survey data and Colleen Johnson for making the ecoregion map. The views expressed in this paper are those of the authors and do not represent those of the US EPA.

**Literature Cited**

- ARNWINE, D. H., AND G. M. DENTON. 2001. Development of regionally-based numeric interpretations of Tennessee's narrative biological integrity criteria. Tennessee Department of Environmental and Conservation, Nashville, Tennessee.
- BAILEY, R. C., R. H. NORRIS, AND T. B. REYNOLDS. 2004. Bioassessment of freshwater ecosystems: using the reference condition approach. Kluwer Academic Publishers, New York.
- BARBOUR, M. T., J. GERRITSEN, B. D. SNYDER, AND J. B. STRIBLING. 1999. Rapid bioassessment protocols for use in streams and wadeable rivers. EPA/841/B-99/002. US Environmental Protection Agency, Washington, DC.
- BRINSON, M. M., AND R. RHEINHARDT. 1996. The role of reference wetlands in functional assessment and mitigation. *Ecological Applications* 6:69-76.
- CAO, Y., C. P. HAWKINS, J. OLSON, AND M. A. KOSTERMAN. 2007. Modeling natural environmental gradients improves the accuracy and precision of diatom-based indicators. *Journal of the North American Benthological Society* 26:566-585.
- CARLISLE, D. M., AND C. P. HAWKINS. 2008. Land use and the structure of western US stream invertebrate assemblages: predictive models and ecological traits. *Journal of the North American Benthological Society* 27:986-999.

- CHESSMAN, B. C., AND M. J. ROYAL. 2004. Bioassessment without reference sites: use of environmental filters to predict natural assemblages of river macroinvertebrates. *Journal of the North American Benthological Society* 23: 599–615.
- CUTLER, D. R., T. C. EDWARDS, K. H. BEARD, A. CUTLER, K. T. HESS, J. GIBSON, AND J. J. LAWLER. 2007. Random forests for classification in ecology. *Ecology* 88:2783–2792.
- DAVIES, S. P., AND S. K. JACKSON. 2006. The biological condition gradient: a conceptual model for interpreting detrimental change in aquatic ecosystems. *Ecological Applications* 16:1251–1266.
- EHLERT, T., D. HERING, U. KOENZEN, T. POTTSGLIESSER, H. SCHUHMACHER, AND G. FRIEDRICH. 2002. Typology and type specific reference conditions for medium-sized streams and large rivers in North Rhine-Westphalia: methodological and biological aspects. *International Review of Hydrobiology* 87:151–163.
- FEMINELLA, J. W. 2000. Correspondence between stream macroinvertebrate assemblages and 4 ecoregions of the southeastern USA. *Journal of the North American Benthological Society* 19:442–461.
- HATTON-ELLIS, T. 2008. The hitchhiker's guide to the water framework directive. *Aquatic Conservation: Marine and Freshwater Ecosystems* 18:111–116.
- HAWKINS, C. P., R. H. NORRIS, J. GERRITSEN, R. M. HUGHES, S. K. JACKSON, R. K. JOHNSON, AND R. J. STEVENSON. 2000. Evaluation of the use of landscape classification for the prediction of freshwater biota: synthesis and recommendations. *Journal of the North American Benthological Society* 19:541–556.
- HAWKINS, C. P., AND M. R. VINSON. 2000. Weak correspondence between landscape classifications and stream invertebrate assemblages: implications for bioassessment. *Journal of the North American Benthological Society* 19:501–517.
- HERLIHY, A. T., W. J. GERTH, J. LI, AND J. L. BANKS. 2005. Macroinvertebrate community response to natural and forest harvest gradients in western Oregon headwater streams. *Freshwater Biology* 50:905–919.
- HERLIHY, A. T., R. M. HUGHES, AND R. C. SIFNEOS. 2006. National clusters of fish species assemblages in the conterminous United States and their relationship to existing landscape classification schemes. Pages 87–112 in R. M. Hughes, L. Wang, and P. W. Seelbach (editors). *Influences of landscapes on stream habitats and biological assemblages*. American Fisheries Society, Bethesda, Maryland.
- HERLIHY, A. T., D. P. LARSEN, S. G. PAULSEN, N. S. URQUHART, AND B. J. ROSENBAUM. 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.
- HUGHES, R. M., D. P. LARSEN, AND J. M. OMERNIK. 1986. Regional reference sites: a method for assessing stream potentials. *Environmental Management* 10:629–635.
- HUGHES, R. M., AND D. V. PECK. 2008. Acquiring data for large aquatic resource surveys: the art of compromise among science, logistics, and reality. *Journal of the North American Benthological Society* 27:837–859.
- JOSEPH, S. 2004. Regional Environmental Monitoring and Assessment Program (REMAP): assessment of biological integrity in select river basins of New Mexico. New Mexico Environment Department, Santa Fe, New Mexico.
- KANSAS DWP (KANSAS DEPARTMENT OF WILDLIFE AND PARKS). 2002. Measuring the status and trends of biological resources in Kansas using Environmental Monitoring and Assessment Program probability based sampling design (R-EMAP). Kansas Department of Wildlife and Parks, Pratt, Kansas.
- KARR, J. R., L. A. TOTH, AND D. R. DUDLEY. 1985. Fish communities of midwestern rivers: a history of degradation. *BioScience* 35:90–95.
- KAUFMANN, P. R., A. T. HERLIHY, M. E. MITCH, J. J. MESSER, AND W. S. OVERTON. 1991. Stream chemistry in the eastern United States: 1. Synoptic survey design, acid-base status and regional patterns. *Water Resources Research* 27:611–627.
- KAUFMANN, P. R., P. LEVINE, E. G. ROBISON, C. SEELIGER, AND D. V. PECK. 1999. Quantifying physical habitat in wadeable streams. EPA 620/R-99/003. Office of Research and Development, US Environmental Protection Agency, Washington, DC.
- KENTUCKY DEP (KENTUCKY DEPARTMENT OF ENVIRONMENTAL PROTECTION). 2002. Methods for assessing biological integrity of surface waters in Kentucky. Kentucky Department of Environmental Protection, Frankfort, Kentucky.
- KLEMM, D. J., K. A. BLOCKSOM, F. A. FULK, A. T. HERLIHY, R. M. HUGHES, P. R. KAUFMANN, D. V. PECK, J. L. STODDARD, W. T. THOENY, M. B. GRIFFITH, AND W. S. DAVIS. 2003. Development and evaluation of a macroinvertebrate biotic integrity index (MBII) for regionally assessing Mid-Atlantic Highlands streams. *Environmental Management* 31:656–669.
- LINAM, G., L. J. KLEINSASSER, AND K. B. MAYES. 2002. Regionalization of the index of biotic integrity for Texas streams. *River Studies Report* 17. Texas Parks and Wildlife Department, Austin, Texas.
- MCCUNE, B., AND J. B. GRACE. 2002. Analysis of ecological communities. MjM Software Design, Gleneden Beach, Oregon.
- METZELING, L., D. TILLER, P. NEWALL, F. WELLS, AND J. REED. 2006. Biological objectives for the protection of rivers and streams in Victoria, Australia. *Hydrobiologia* 572:287–299.
- MONTGOMERY, D. C., E. A. PECK, AND G. G. VINING. 2001. Introduction to linear regression. 3<sup>rd</sup> edition. John Wiley and Sons, New York.
- MOOG, O., A. SCHMIDT-KLOIBER, T. OFENBÖCK, AND J. GERRITSEN. 2004. Does the ecoregion approach support the typological demands of the EU 'Water Framework Directive'? *Hydrobiologia* 516:21–33.
- MOSS, D., M. T. FURSE, J. F. WRIGHT, AND P. D. ARMITAGE. 1987. The prediction of the macroinvertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. *Freshwater Biology* 17:41–52.

- MOULTON, S. R., J. G. KENNEN, R. M. GOLDSTEIN, AND J. A. HAMBROOK. 2002. Revised protocols for sampling algal, invertebrate, and fish communities as part of the National Water-Quality Assessment Program. U.S. Geological Survey Open-File Report 02-150. US Geological Survey, Reston, Virginia.
- ODE, P. R., C. P. HAWKINS, AND R. D. MAZOR. 2008. Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. *Journal of the North American Benthological Society* 27:967-985.
- OLSEN, A. R., AND D. V. PECK. 2008. Survey design and extent estimates for the Wadeable Streams Assessment. *Journal of the North American Benthological Society* 27:822-836.
- OMERNIK, J. M. 1987. Ecoregions of the conterminous United States. *Annals of the Association of American Geographers* 77:118-125.
- PECK, D. V., A. T. HERLIHY, B. H. HILL, R. M. HUGHES, P. R. KAUFMANN, D. J. KLEMM, J. M. LAZORCHAK, F. H. MCCORMICK, S. A. PETERSON, P. L. RINGOLD, T. MAGEE, AND M. R. CAPPAERT. 2006. Environmental Monitoring and Assessment Program—Surface Waters Western Pilot Study: field operations manual for wadeable streams. EPA 600/R-06/003. Office of Research and Development, US Environmental Protection Agency, Washington, DC.
- RABENI, C. F., AND K. E. DOISY. 2000. Correspondence of stream benthic invertebrate assemblages to regional classification schemes in Missouri. *Journal of the North American Benthological Society* 19:419-428.
- REYNOLDS, T. B., AND J. F. WRIGHT. 2000. The reference condition: problems and solutions. Pages 293-303 in J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). *Assessing the biological quality of fresh waters: RIV-PACS and other techniques*. The Freshwater Biological Association, Far Sawrey, UK.
- SANDIN, L., AND P. F. M. VERDONSCHOT. 2006. Stream and river typologies: major results and conclusions from the STAR project. *Hydrobiologia* 566:33-37.
- STEVENS, D. L., AND A. R. OLSEN. 2004. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* 99:262-278.
- STODDARD, J. L., A. T. HERLIHY, D. V. PECK, R. M. HUGHES, T. R. WHITTIER, AND E. TARQUINIO. 2008. A process for creating multimetric indices for large-scale aquatic surveys. *Journal of the North American Benthological Society* 27:878-891.
- STODDARD, J. L., D. P. LARSEN, C. P. HAWKINS, R. K. JOHNSON, AND R. H. NORRIS. 2006. Setting expectations for the ecological condition of running waters: the concept of reference condition. *Ecological Applications* 16:1267-1276.
- STODDARD, J. L., D. V. PECK, A. R. OLSEN, D. P. LARSEN, J. VAN SICKLE, C. P. HAWKINS, R. M. HUGHES, T. R. WHITTIER, G. LOMNICKY, A. T. HERLIHY, P. R. KAUFMANN, S. A. PETERSON, P. L. RINGOLD, S. G. PAULSEN, AND R. BLAIR. 2005a. Environmental Monitoring and Assessment Program (EMAP): western streams and rivers statistical summary. EPA 620/R-05/006. US Environmental Protection Agency, Washington, DC.
- STODDARD, J. L., D. V. PECK, S. G. PAULSEN, J. VAN SICKLE, C. P. HAWKINS, A. T. HERLIHY, R. M. HUGHES, P. R. KAUFMANN, D. P. LARSEN, G. LOMNICKY, A. R. OLSEN, S. A. PETERSON, P. L. RINGOLD, AND T. R. WHITTIER. 2005b. An ecological assessment of western streams and rivers. EPA 620/R-05/005. US Environmental Protection Agency, Washington, DC.
- TURAK, E., AND K. KOOP. 2008. Multi-attribute ecological river typology for assessing ecological condition and conservation planning. *Hydrobiologia* 603:83-104.
- USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2006. Wadeable Streams Assessment: a collaborative survey of the nation's streams. EPA 841-B-06-002. Office of Research and Development and Office of Water, US Environmental Protection Agency, Washington, DC.
- VAN SICKLE, J., AND R. M. HUGHES. 2000. Classification strengths of ecoregions, catchments, and geographic clusters for aquatic vertebrates in Oregon. *Journal of the North American Benthological Society* 19:370-384.
- VAN SICKLE, J., AND S. G. PAULSEN. 2008. Assessing the attributable risks, relative risks, and regional extents of aquatic stressors. *Journal of the North American Benthological Society* 27:920-931.
- VAN SICKLE, J., J. L. STODDARD, S. G. PAULSEN, AND A. R. OLSEN. 2006. Using relative risk to compare the effects of aquatic stressors at a regional scale. *Environmental Management* 38:1020-1030.
- VERDONSCHOT, P. F. M., AND R. C. NIJBOER. 2004. Testing the European stream typology of the Water Framework Directive for macroinvertebrates. *Hydrobiologia* 516:35-54.
- VERMONT DEC (VERMONT DEPARTMENT OF ENVIRONMENTAL CONSERVATION). 2002. Wadeable stream biocriteria development for fish and macroinvertebrate assemblages in Vermont streams and rivers. Vermont Department of Environmental Conservation, Waterbury, Vermont.
- WAITE, I. R., A. HERLIHY, D. P. LARSEN, AND D. J. KLEMM. 2000. Comparing strengths of geographic and nongeographic classifications of stream benthic macroinvertebrates in the Mid-Atlantic Highlands, USA. *Journal of the North American Benthological Society* 19:429-441.
- WHITTIER, T. R., J. L. STODDARD, D. P. LARSEN, AND A. T. HERLIHY. 2007. Selecting reference sites for stream biological assessments: best professional judgment or objective criteria. *Journal of the North American Benthological Society* 26:349-360.
- YUAN, L. L., C. P. HAWKINS, AND J. VAN SICKLE. 2008. Effects of regionalization decisions on an O/E index for the national assessment. *Journal of the North American Benthological Society* 27:892-905.

Received: 14 May 2008

Accepted: 28 August 2008