

# Data synthesis and bioindicator development for nontidal streams in the interstate Potomac River basin, USA

LeAnne E. Astin \*

*Interstate Commission on the Potomac River Basin, 51 Monroe Suite PE 08, Rockville, MD 20850, USA*

Received 18 March 2005; received in revised form 5 August 2005; accepted 21 August 2005

---

## Abstract

Water resource agencies are increasingly confronted with issues of methods and data comparability when assessing inter-jurisdictional waters. The Interstate Commission on the Potomac River Basin (ICPRB) is developing an assessment framework for the diverse nontidal monitoring data collected by Potomac basin jurisdictions. The ICPRB's goal is to augment the jurisdictions' water quality assessments with uniform, basin-wide evaluations of habitat and biological integrity. Disparate datasets were integrated and used to select and calibrate a common suite of biological indicators of human disturbance. Aggregated physiographic regions were most effective at grouping streams into site classes with similar attributes. Common elements from states' habitat and water quality assessments were used to define regional reference and impairment criteria. Seven macroinvertebrate metrics distinguished reference from impaired sites in most regions: EPT richness, Hilsenhoff Family Biotic Index, percent clingers, percent collectors, percent dominance, percent EPT, and taxonomic richness. Reference communities in the Piedmont region of the basin were dissimilar from those in Highland and Valley regions, suggesting a need for different reference conditions and metric thresholds in these regions. Results confirm that useful bioindicators of aquatic health can be generated from multiple datasets if the synthesis is done with care.

© 2005 Elsevier Ltd. All rights reserved.

*Keywords:* Comparability; Synthesis; Data integration; Bioassessment; Interstate; Nontidal

---

## 1. Introduction

The Interstate Commission on the Potomac River Basin (ICPRB) is an interstate compact agency that assists the Potomac basin states and the federal

government to cooperatively address water quality and related resource problems in the watershed. The Commission relies on monitoring data and information provided by its member jurisdictions (Maryland, Pennsylvania, West Virginia, Virginia, and the District of Columbia) to assess the river and its tributaries. Comprehensive and consistent assessments on a regional or watershed level can be problematic due

---

\* Tel.: +1 301 984 1908; fax: +1 301 984 5841.

*E-mail address:* [lastin@icprb.org](mailto:lastin@icprb.org).

to differences in project design, methods, data analysis, and data management (Methods and Data Comparability Board, 2003).

The ICPRB is developing an assessment framework to merge nontidal monitoring data collected by its member jurisdictions and evaluate them on a common “yardstick.” The framework will be used to assess the health of stream biota in a consistent, basin-wide manner. There are several advantages to integrating scientific information from multiple sources. Pooling reference site data increases confidence in reference conditions, ensuring that assessment criteria reflect the best available conditions for a particular region and stream type (Maxted et al., 2000). Combining information from separate monitoring surveys improves understanding of the biological integrity of riverine systems (Handcock et al., 2002). As watersheds transcend political, social, and economic boundaries, taking a watershed approach ensures greater emphasis on ecological results (US EPA, 2002).

The framework integrates habitat, water quality, and benthic macroinvertebrate data from wadeable, nontidal streams of the Potomac watershed. The basin jurisdictions all use macroinvertebrate communities for biological assessment of their waters. Biological measures are well suited to stream assessment because they are an effective measure of the condition of aquatic life and provide a cost-effective way to assess ecological condition of a large number of streams over large geographic areas (Maxted et al., 2000). Each basin state employs some variation of the widely used and understood US EPA Rapid Bioassessment Protocols (RBP) for Streams and Rivers (Plafkin et al., 1989; Barbour et al., 1999). Therefore, an additive (multi-metric) framework based on this method best suits the heterogeneous data provided by the jurisdictions.

Five steps were involved in characterizing good- and poor-quality ecological conditions and identifying metric thresholds: (1) merge monitoring data provided by the jurisdictions; (2) resolve methods and data compatibility issues; (3) establish uniform habitat evaluation criteria to identify impaired and unimpaired sites; (4) select robust biological indicators of impairment; (5) evaluate the validity of aggregating reference data across geophysical regions. These steps will lead to the development of an integrative index that will facilitate interstate evaluations of biological integrity of Potomac nontidal, wadeable streams.

## 2. Methods

### 2.1. Data integration and integrity

The following datasets were combined in a relational database: Maryland Department of Natural Resources’ (MdDNR) Maryland Biological Stream Survey (MBSS) 1995–1997 ( $n = 290$ ); Pennsylvania Department of Environmental Protection’s (PaDEP) Unassessed Waters/State-Wide Surface Waters Assessment Program (UW/SSWAP) 1997–2001 ( $n = 269$ ); Virginia Department of Environmental Quality’s (VaDEQ) Ambient Water Quality Monitoring program 1994–2000 ( $n = 428$ ); West Virginia Department of Environmental Protection’s (WVDEP) Watershed Assessment Program (WAP) 1996–spring 2001 ( $n = 336$ ). Stream data from the District of Columbia (DC) were unavailable for this analysis.

Resolving the differences inherent in the various datasets to ensure data integrity proved a substantial challenge. This is a common problem when combining diverse data sources into a single database (McLaughlin et al., 2001). Inconsistent, and sometimes incompatible, data formats and file types were resolved to a common format. Considerable effort was spent to ensure that fields with the same names in different datasets contained the same information, and that fields with different names but containing the same data were combined. Errors found during data importation or manual entry into the database were corrected following quality assurance cross-checking; inconsistencies that could not be resolved were deleted. Equivalent NOAA National Oceanographic Data Center (NODC) taxonomic codes and Integrated Taxonomic Information System (ITIS) Taxon Serial Numbers (TSN) were inserted into each data record to facilitate tracking taxa over time regardless of changes in common and scientific nomenclature. Regional and structural classifications (e.g. catchments, ecoregions, hydrologic units) were also inserted to improve geographical references for stream sites. Documentation (metadata) was not provided, so we obtained information on agencies’ collecting, processing, analysis, and interpretation methods from published SOPs, field manuals, reports, and the WWW.

## 2.2. Methods and data comparability issues

Prior to analysis, it was necessary to address various underlying methodological differences and the effects they have on bioassessment data and endpoints. Different methods used by the states to select, sample, and evaluate sites can introduce variability when data are combined, and confound analysis and interpretation. Potential confounding factors and the methods used to resolve them are described below.

### 2.2.1. Study designs

Basin jurisdictions employ targeted/judgment (VaDEQ), random/probabilistic (MdDNR, WVDEP), and census sampling (PaDEP). In judgment sampling, fixed sites are selected based on the best professional judgments of experts. In a probability-based design, stream sites are randomly selected within a region. In census sampling, every unit is assessed for an entire region. Sample sites selected from a combination of study designs may be used to develop biological indices if they are representative of the range of conditions for the area of interest. For example, the US EPA used probabilistic and targeted sites in the mid-Atlantic Highlands Assessment (MAHA) (Davis, 2001; Fore, 2003). Integration of data derived from different sampling designs allows the emergence of information that cannot be obtained from independent efforts (US EPA, 1997; McLaughlin et al., 2001). In the absence of a more comprehensive dataset, we assumed that repeat observations taken at fixed sites were independent (Burton and Gerritsen, 2003), and that all such sites (reference, impaired, and other) were representative of the range of conditions found in the Potomac basin.

### 2.2.2. Index periods

Field collection periods vary among agencies: VaDEQ samples macroinvertebrates and habitat in spring and fall; MdDNR samples macroinvertebrates in spring, habitat in summer; PaDEP samples both in summer; WVDEP samples both nearly year-round (March–December). Although it is not usually advisable (Lenz, 1997), it was necessary to combine datasets collected in different seasons. Burton and Gerritsen (2003) found that classification of Virginia streams into separate index periods was unnecessary due to a high degree of similarity between spring and

fall macroinvertebrate samples. We examined site classifications based on broad “sampling seasons”: spring (3/15–6/14), summer (6/15–9/14), and fall/winter (9/15–3/14) (see below).

### 2.2.3. Sampling and subsampling methods and gears

Different field methods may produce dissimilar data (Kerans et al., 1992; Lenz and Miller, 1996; Carter and Resh, 2001; Stribling and Bressler, 2001). Differences in laboratory sorting and subsampling procedures are also potential sources of bias (Courtemanch, 1996; Vinson and Hawkins, 1996; Grows et al., 1997). It is difficult to quantify the variability associated with different bioassessment methods even when performance characteristics are documented (Diamond et al., 1996; Barbour et al., 1999; Stribling and Bressler, 2001). MdDNR was the only state agency using performance-based methods during this study. In the absence of documented data quality characteristics, we accepted the datasets as they were given to us by the states.

### 2.2.4. Subsample size

Most basin state agencies (except PaDEQ; see below) utilize a fixed-count method of subsampling; the target number of organisms to be “picked” varies among protocols. Some researchers argue against comparing subsamples of different sizes since richness metrics are affected by the numbers of organisms (Barbour and Gerritsen, 1996; Courtemanch, 1996; Grows et al., 1997; Sovell and Vondracek, 1999). Rarefaction is sometimes used to eliminate the effects of differing sample sizes and/or sampling effort. Most basin states use fixed counts  $\leq 200$ , which are inadequate for rarefaction (Vinson and Hawkins, 1996; Larson and Herlihy, 1998). Rarefaction is also known to affect richness metrics through overestimation. Gerritsen et al. (2000b) found that the effects of differing subsample sizes were not great with family level identifications. It was decided not to rarefy the data for our analysis.

### 2.2.5. Habitat assessment protocols and criteria

Although the basin states utilize similar visual-based approaches to characterize physical habitat, these approaches vary in the numbers and kinds of parameters evaluated. Jurisdictions also emphasize

differing attributes when identifying reference and impaired streams. This was addressed by selecting common elements from states' habitat and water quality assessments and equating them to a consistent scale as described below.

#### 2.2.6. *Taxonomic resolution*

MdDNR identifies macroinvertebrates to genus. All other basin states identify to family (or higher). To equalize the taxonomic hierarchy, genus-level identifications were collapsed to family. Hewlitt (2000) and Bailey et al. (2001) found family-level identifications to be sufficient for broad-scale bioassessment programs. And as noted, it is thought to minimize the effect of differing subsample sizes (Gerritsen et al., 2000b).

#### 2.2.7. *Analytical approaches*

Most basin jurisdictions make semi-quantitative counts of abundance. PaDEP's UW/SSWAP program identifies macroinvertebrates to family level in the field, then estimates the relative abundance of each family (Pennsylvania Department of Environmental Protection, 2003). To merge qualitative and semi-quantitative counts, we used the minimum value in each numeric range associated with the narrative categories defined on PaDEP's Unassessed Waters Field Form. A parallel analysis excluding the relative abundance values (RAs) from percentage calculations was also performed to ensure that using the minimum RAs as quantitative counts did not result in a disproportionate weight for metric values.

### 2.3. *Site classification*

Classification is the partitioning of natural variability into homogeneous groups. Candidate reference sites based on minimally degraded physicochemical attributes are used as the basis for stream classification (Barbour et al., 1999). We used the habitat elements collected for basin states' stream assessments to identify candidate reference sites (Table 1). Shared elements (or ecological equivalents) were aggregated into seven parameter categories scored from 0 to 20 (Table 2). Two quantitative water quality measurements (pH and conductivity) were also incorporated. Although the basin states measure dissolved oxygen (DO), DO was not included as a criterion because it

varies significantly with time of day and ambient water temperature.

Numerous statistical techniques are used in bioassessment to explore alternate classifications of reference sites. Ordination analysis, a multivariate method, is a widely used approach to examining the similarity or dissimilarity of biological samples. However, ordination cannot distinguish natural variation from sampling artifacts. We decided that multivariate ordination would not be suitable for identifying and evaluating classification groupings. We applied a variation of the Relative Status Method originally developed by Alden and Perry (1997) and refined by Olson (2002) to identify least-impaired sites. These were then used to confirm site classes using a univariate method.

Stream sites were first stratified a priori by sampling season, primary subwatershed (a.k.a. catchment) derived from USGS 8-digit HUCs, and regional physiographic structure (Fig. 1). We tested several ecoregional schemes (Omernik, 1987; Woods et al., 1999; US EPA, 2000) (Table 3), and determined that aggregated US EPA Level III ecoregions and Level IV subregions represented the most appropriate physiographic framework. Within strata, the distributions of scores for the habitat parameters in Table 2 were divided into percentiles. Upper and lower criteria for the parameter scores and chemical measures were defined. We reviewed quantitative and narrative criteria applied by the basin jurisdictions to determine which abiotic parameters they consider most indicative of stream degradation. A literature review was also conducted to see which habitat measures are thought to be of greater importance to biological communities. Based on these investigations, our habitat parameters were tested iteratively by "strengthening" and "relaxing" criteria. Parameters (or equivalents) reported to be "more precise" in detecting perturbation in mid-Atlantic streams (Kaufmann et al., 1999) were assigned more stringent criteria. Some parameters were assigned more (or less) stringent criteria in certain strata. Quantitative criteria for pH and conductivity were adapted from states' water quality standards. Sites with all habitat and water quality parameters in the most desirable percentiles of their distributions were ranked as candidate reference sites. A similar process was used to identify candidate impaired sites for testing metric

Table 1  
Habitat and water quality (WQ) elements-in-common to Potomac basin states (excluding DC)

	PA (1997–2001)	MD (1995–1997)	VA (1994–2000)	WV (1996–2001)
<b>WQ parameters</b>				
N		x		x
SO		x		x
PO				x
DOC		x		x
TSS				x
Turbidity				? (Rare)
pH	x	x	x	x
DO	x	x	x	x
Temperature	x	x	x	x
Conductivity	x	x	x	x
Alkalinity	? (Rare)			x
Hardness				? (Rare)
Fecal coliforms				x
Salinity			x	
Chloride			x	x
Calcium				x
Asst. metals				x
ANC		x		
Flow (current velocity)	? (Rare)	x		
<b>Habitat parameters</b>				
Instream hab/cover (A) <sup>a</sup>	x	x	x	x
Epifaunal sub (A)	x	x	x	x
Velocity/depth	x	x	x	x
Pool/Eddy/Glide	x	x (Quality)		
Riffle freq (B)	x	x (Quality)	x	x
<b>Chan alteration<sup>b</sup> (B)</b>	x	x	x	x
Bank stability (C)	x	x	x	x
Embeddedness (A)	x	x (%)	x	x
Sediment deposition (B)	x	x	x	x
Chan flow	x	x	x	x
Shade		x		
<b>Riparian buffer</b>	x	x (Width)	x	x
Aesthetic		x		x (Post-98)
Remoteness		x		x (Post-98)
Pool substrate	x (Glide/pool)			
Sinuosity	x (Glide/pool)		x	x (Glide/pool)
Bank veg (C)	x		x	x
<b>Grazing/obv. dist.</b>	x		x	x (Pre-98)
Thalweg depth		x		
Wetted width		x		
Pebble count		x		
Land use(s)	x	x		x
Stream gradient		x		
Woody debris		x		

PA, Pennsylvania DEP; MD, Maryland DNR; VA, Virginia DEQ; WV, West Virginia DEP.

<sup>a</sup> Letters in parenthesis indicate principal categories (primary, secondary, tertiary) of biological significance (Plafkin et al., 1989).

<sup>b</sup> Parameters in bold “more precise” for mid-Atlantic streams (Kaufmann et al., 1999).

Table 2  
Consolidated physical habitat assessment parameters and ecological equivalents

Parameter name	Parameter code	Ecological equivalents
Anthropogenic alterations	<b>ANTHRO_ALT</b>	Grazing score OR aesthetics + remoteness/2
Bank stability	BANKS	Bank stability
Channel alteration	<b>CHAN_ALT</b>	Channel alteration
Habitat heterogeneity	HAB_HETERO	Riffle frequency OR sinuosity OR Pool/Glide/Eddy quality
Instream condition	INSTR_COND	Epifaunal substrate + cover/2 OR Epifaunal substrate/available cover score
Riparian zone	<b>RIP_ZONE</b>	Riparian buffer score OR Riparian width (m) scored to same scale
Substrate quality	SUB_QUAL	Embeddedness OR pool substrate OR %Embeddedness scored to same scale
Total score	TotalScore	Sum of above parameter scores
pH	PH	
Conductivity	CON	

Parameters in bold are considered “more precise” for mid-Atlantic streams (Kaufmann et al., 1999).

discrimination. Criteria for both quartile (less stringent) and decile (more stringent) ranking schemes were defined and tested.

To eliminate extreme outliers, candidate reference and impaired sites were screened by omitting those where the metrics taxa richness (TaxRich) and total abundance (TotAbun) were > or < 2 standard deviations (S.D.) from their means. TaxRich and TotAbun are the “base measures” from which most others are derived: taxa counts (e.g. Ephemeroptera/Plecoptera/

Trichoptera (EPT) taxa, Diptera taxa) are subsets of taxonomic richness; percentage metrics (e.g. percent EPT, percent dominance) are expressed as some proportion of the total abundance of all individuals. While there could be legitimate reasons for a site to have many individuals representing few taxa, or many taxa composed of only a few individuals, sites with extremely inflated (or deflated) values for both these metrics were presumed to be artifacts of the compromises made to equalize the data. This screen-

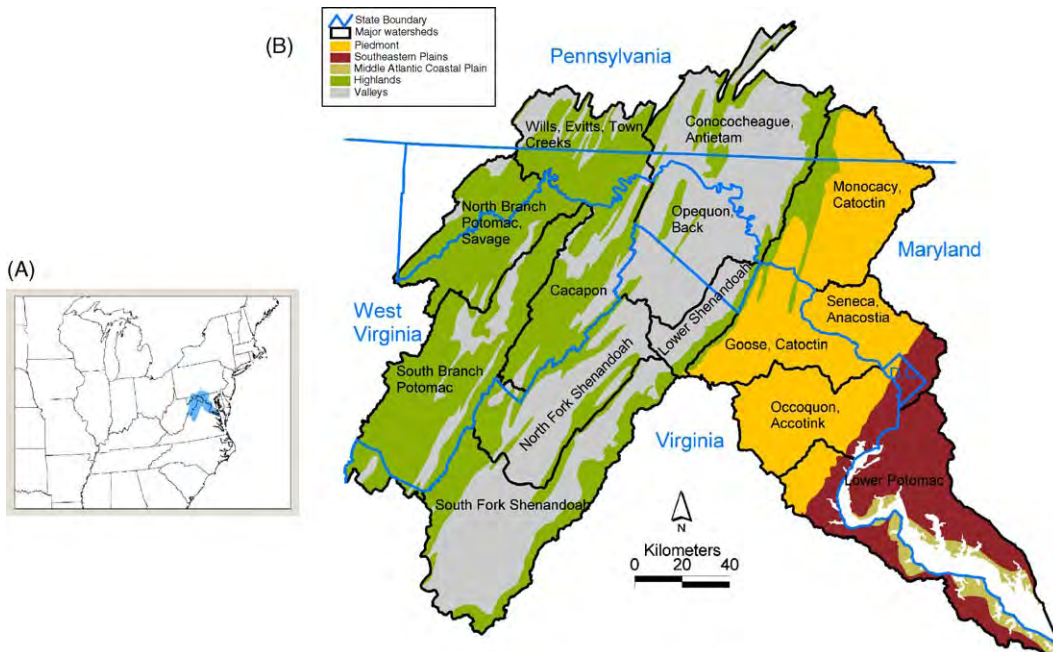


Fig. 1. (A) Location of the Potomac River basin and (B) primary subwatersheds and physiographic regions of the Potomac River basin.

Table 3  
Ecoregions, subregions, and aggregated regions of the Potomac River basin

<b>Omerik (1987) ecoregions</b>	
63	Middle Atlantic Coastal Plain
64	Northern Piedmont
65	Southeastern Plains
66	Blue Ridge Mountains
67	Central Appalachian Ridges and Valleys
69	Central Appalachians
<b>Level III (Woods et al., 1999) ecoregions</b>	
45	Piedmont
63	Middle Atlantic Coastal Plain
64	Northern Piedmont
65	Southeastern Plains
66	Blue Ridge
67	Ridge and Valley
69	Central Appalachians
<b>Level IV (Woods et al., 1999) subregions</b>	
45e	Northern Inner Piedmont
45f	Northern Outer Piedmont
63b	Chesapeake-Albemarle Silty Lowlands and Tidal Marshes
64a	Triassic Lowlands
64b	Diabase and Conglomerate Uplands
64c	Piedmont Uplands
64d	Piedmont Limestone/Dolomite Lowlands
65m	Rolling Coastal Plain
65n	Chesapeake Rolling Coastal Plain
66a	Northern Igneous Ridges
66b	Northern Sedimentary and Metasedimentary Ridges
67a	Northern Limestone/Dolomite Valleys
67b	Northern Shale Valleys
67c	Northern Sandstone Ridges
67d	Northern Dissected Ridges
69a	Forested Hills and Mountains
69b	Uplands and Valleys of mixed land use
<b>MAHA (US EPA, 2000) aggregated ecoregions (highlands only)</b>	
66, 67c, 67d	Ridge and Blue Ridge
67a, 67b	Valley
69a, 69b	North-Central Appalachian Plateau

ing eliminated one candidate reference site (final reference  $n = 87$ ) but no candidate stressed sites (final stressed  $n = 84$ ).

We used one-way analysis of variance (ANOVA) to examine the summed habitat scores (TotalScore) of the screened reference sites to confirm that physiographic region, primary subwatershed, and sampling season were significant grouping factors. A second ANOVA examined select biological characteristics known to be robust indicators (EPT, TaxRich, percent

Haptobenthos (Smith and Voshell, 1997), and Hilsenhoff's Family Biotic Index (Hilsenhoff, 1988)). Strata were tested both independently and "nested." Normalcy tests (Kolmogorov–Smirnov (KS)) on TotalScore and the biological metrics showed that their reference distributions were not significantly different ( $p > 0.05$ ) from the standard normal distribution, so a parametric ANOVA was deemed appropriate.

#### 2.4. Biological metric screening

The ability of a metric to discriminate between known reference sites and known stressed sites is crucial in selecting useful indicators (Barbour et al., 1999). We tested 61 candidate metrics representing 6 structural and functional categories, including 4 abundance, 10 richness, 7 composition, 13 tolerance, 17 trophic, and 10 habit measures (Table 4).

We first eliminated candidates whose values could not be calculated for at least 90% of reference sites. Metrics with a coefficient of variation (C.V.)  $> 0.50$  in the reference site population were also eliminated. As noted, the parallel analysis excluded PaDEP's minimum relative abundance values (RAs) from proportional metric calculations to ensure that using RAs as quantitative counts did not skew metric values; therefore, UW/SSWAP reference sites were not included in these tests in the parallel analysis. Candidate metrics were then evaluated using Student's  $t$ -test to determine if they could effectively discriminate reference sites from stressed sites. To meet the assumption of a normal distribution, three data transformations were tested: arcsine-square-root, base 10 logarithmic, and square-root-square-root (Thorne et al., 1999); untransformed metric values were also evaluated. Kolmogorov–Smirnov goodness-of-fit tests (KS) were performed on transformed and untransformed values to test for departures from normality. Least-squares procedures such as  $t$ -tests perform well even when data are non-normal if there is homogeneity of variance between the populations being tested. We used the parametric  $t$  if transformed or untransformed values were significantly similar to a standard normal distribution and/or if Browne-Forsyth homogeneity of variance (HoV) tests showed that the variances of the reference and stressed populations were similar. We performed two-tailed  $t$ -tests assum-

Table 4  
List of candidate metrics

Metric	ExpResponse	Category	Function
AllEphem	▼	A	Mayfly individuals
AllFilt	—	F	Collector–filterer individuals
AllGather	—	F	Collector–gatherer individuals
AllHapto	▼	H	Haptobenthic (clingers + crawlers) individuals
AllHerpo	▲	H	Herpobenthic (burrowers + sprawlers) individuals
AllSimulids	▲	T	Blackfly individuals
CFTaxa	—	F	Collector–filterer taxa
ClingAllPlecop	▼	H	Clinger and all Plecoptera individuals
ClingTaxa	▼	H	Clinger taxa
DipteraTaxa	▼	T	True fly taxa
EPT_Taxa	▼	R	Ephemeroptera/Plecoptera/Trichoptera taxa
EPTTax_NoHydro	▼	R	EPT taxa excluding Hydropsychidae
HaptoTaxa	▼	H	Haptobenthic (clingers + crawlers) taxa
IndexFBI	▲	T	Family level Hilsenhoff Biotic Index
IndexMargalef	▼	R	Diversity index; richness normalized for sample size
IndexShanWein	▼	R	Diversity index; order/disorder within a community
IndexSimpDiversity	▼	R	Diversity index; richness and proportion of taxa
NoninsectTaxa	▲	T	Non-insect taxa
NumClingInd	▼	H	Clinger individuals
NumCollInd	—	F	Gatherer + filterer individuals
NumEphem	▼	R	Ephemeroptera taxa
NumEPTInd	▼	A	EPT individuals
NumInd_ByTaxon	—	A	Individual abundance in each taxon
NumPredInd	▲	F	Predator individuals
NumScrapInd	▼	F	Scrapper individuals
NumShredInd	▼	F	Shredder individuals
NumTolInd	▲	T	Tolerant individuals (tolerance value > 7)
NumTrichop	▼	R	Caddisfly taxa
PercentChiro	▲	T	Proportion of midges
PercentCling	▼	H	Proportion of clingers
PercentClingand&Plecop	▼	H	Proportion of clingers (including all Plecoptera)
PercentColl	▲	F	Proportion of collectors (gatherers + filterers)
PercentDiptera	▲	C	Proportion of true flies
PercentDom1	▲	T	Percent contribution of dominant taxon
PercentDom2	▲	T	Percent contribution of the 2 dominant taxa
PercentDom5	▲	T	Percent contribution of the 5 dominant taxa
PercentEphem	▼	C	Estimates percent mayfly individuals
PercentEPT	▼	C	Proportion of EPT individuals
PercentEPTTaxa	▼	C	Proportion of EPT taxa
PercentFilter	▲	F	Proportion of collector–filterers
PercentGather	—	F	Proportion of collector–gatherers
PercentHapto	▼	H	Proportion of Haptobenthos (clingers + crawlers)
PercentHerpo	▲	H	Proportion of Herpobenthos (sprawlers + burrowers)
PercentHydro	▲	C	Proportion of Hydropsychids
PercentHydroChiro	▲	C	Proportion of Chironomids + Hydropsychids
PercentPred	▼	F	Proportion of predators
PercentScrap	▼	F	Proportion of scrapers
PercentScrapFilScrap	▼	F	Proportion of scrapers to scrapers + filterers
PercentShred	▼	F	Proportion of shredders
PercentSimulid	▲	T	Proportion of blackflies
PercentTol	▲	T	Proportion of tolerant individuals
PlecopTaxa	▼	R	Stonefly taxa
RatioScrapFilt	▼	F	Proportion of scrapper taxa to collector–filterers

Table 4 (Continued)

Metric	ExpResponse	Category	Function
RatioShredCF	▼	F	Proportion of shredder taxa to collector–filterers
ScrapTaxa	▼	F	Scraper taxa
SensitiveTaxa	▼	T	Intolerant taxa (tolerance value < 3)
TaxRich	▼	R	Total taxa
TolerantTaxa	▲	T	Tolerant taxa (tolerance value > 7)
TotAbun	—	A	Total abundance of individuals
Trich_NoHydro	▲	R	Caddisflies excluding Hydropsychidae

Expected response (ExpResponse): ▼, decreases in response to perturbation; ▲, increases in response to perturbation; —, variable or unknown response to perturbation. Categories: A, abundance; C, composition; F, feeding (trophic); H, habit, R, richness; T, tolerance.

ing equal or unequal variances, depending on HoV results, to determine if differences existed between mean metric values of reference and stressed sites ( $p < 0.05$ ) when grouped by strata. Metrics able to distinguish between reference and stressed sites were retained for further consideration.

Metric discrimination ability was also assessed by examining each metric's discrimination efficiency (DE). DE measures the degree of separation between metric value distributions of reference and stressed sites and is calculated as a percentage using the following equation:

$$DE = 400 \left( \frac{a}{b} \right)$$

For metrics that are expected to decrease in response to impairment, the values for  $a$  and  $b$  are:  $a$ , number of stressed sites with metric values < the 25th percentile of the reference distribution;  $b$ , number of stressed sites. For those metrics expected to increase in response to impairment:  $a$ , number of stressed sites with metric values > the 75th percentile of the reference distribution. A high DE indicates better assessment accuracy (Gerritsen et al., 2000b).

To ensure that each metric contributed independent information to the index, candidates were grouped by structural and functional category and tested for redundancy within and between categories using a Pearson correlation analysis. Metric combinations with  $r \geq 0.75$  were considered redundant, indicating that one of the pair should be eliminated.

Finally, reference communities were compared between physiographic regions. Mean values of core metrics were tested using two-tailed  $t$ -tests assuming unequal variances to determine if the reference

communities of basin regions were dissimilar. Statistical calculations were performed using Total Access Statistics (Version 8.0, FMS, Inc., Vienna, Virginia).

### 3. Results

#### 3.1. Site classification

Sites were stratified by class (region, catchment, season), and a relative measure based upon habitat (Anthropogenic Alteration, Bank Stability, Channel Alteration, Habitat Heterogeneity, Instream Condition, Riparian Zone, Substrate Quality) and water quality (conductivity, pH) parameters was used to identify least- and most-impaired streams. Candidate reference sites identified from quartile and decile data divisions were examined to determine their effectiveness in identifying conditions. Quartiles did not distinguish reference sites (>75th percentile) from impaired sites (<25th percentile), so this scheme was abandoned. The more stringent thresholds of the decile ranking scheme better characterized sites in excellent condition and sites with severe degradation.

Screened reference sites were used to confirm the necessity of grouping data by site classes to account for natural variability. Group means for the four selected biological metrics (EPT, TaxRich, percent Haptobenthos, Hilsenhoff's Family Biotic Index) and TotalScore were significantly different from each other ( $p \leq 0.05$ ) when apportioned by aggregated physiographic region. Grouping by primary subwatershed appeared to have a significant effect ( $p < 0.01$ ), but the  $n$  in many subwatersheds was insufficient to fully address classification by catch-

Table 5  
Selection criteria used to identify reference sites for each ecoregion ranked by deciles

Reference criteria (all of the following)			
Parameter	P	V	H
ANTHRO_ALT	>70%	>70%	>70%
BANK_STAB	>70%	>50%	>50%
CHAN_ALT	>70%	>50%	>50%
HAB_HETERO	>10%	>10%	>10%
INSTR_COND	>10%	>10%	>10%
RIP_ZONE	>50%	>70%	>70%
SUB_QUAL	>70%	>70%	>70%
And (both)			
pH	Between 6 and 9		
CON	<500		

Ecoregions: P, Piedmont; V, Valleys; H, Highlands. Numbers in columns indicate threshold centiles.

ment. Sampling seasons were nonsignificant as grouping variables due to programmatic differences in index period. Tables 5 and 6 list the habitat and water quality parameters and selection criteria (threshold centiles) used to identify reference and stressed sites for the Piedmont (ecoregions 45 and 64), Highland (ecoregions 66, 69, and subregions 67c and 67d), and Valley (subregions 67a and 67b) regions of the nontidal Potomac basin. We were not able to identify sufficient reference sites in the Southeastern Plains (ecoregion 65) for analysis using the relative method described. The Middle Atlantic Coastal Plains

Table 6  
Selection criteria used to identify stressed sites for each ecoregion ranked by deciles

Impairment criteria (three or more of the following)			
Parameter	P	V	H
ANTHRO_ALT	<10%	<10%	<10%
BANK_STAB	<10%	<10%	<10%
CHAN_ALT	<10%	<10%	<10%
HAB_HETERO	<10%	<10%	<10%
INSTR_COND	<10%	<10%	<10%
RIP_ZONE	<10%	<10%	<10%
SUB_QUAL	<10%	<10%	<10%
Or (either)			
pH	<4.5		
CON	>1000		

Ecoregions: P, Piedmont; V, Valleys; H, Highlands. Numbers in columns indicate threshold centiles.

(ecoregion 63) portion of the basin is tidally influenced and was not evaluated (Fig. 1).

### 3.2. Biological metric screening

Twenty-five metrics survived the initial screening and were selected as potential indicators. Candidate metrics were eliminated from consideration if their DE was less than 50% for all stressed sites or if they could not differentiate between reference and stressed sites ( $p < 0.05$ ) based on Student's  $t$  (Tables 7 and 8). Data transformation to approximate normality proved unnecessary; candidates were either significantly similar to a standard normal distribution ( $p > 0.05$ ), and/or variances between their reference and stressed populations were similar. Candidate metrics were also evaluated for redundancy among and between

Table 7  
Results of  $t$ -tests between metric values of reference and impaired sites

Metric	P	V	H	Selected
CFTaxa (CF)	**	**	**	
ClingTaxa (CL)	**	**	**	
DipteraTaxa (DIP)	NS	NS	NS	
EphemTaxa (E)	NS	**	**	
EPT_Taxa (EPT)	**	**	**	x
EPTTax_NoHydro (ENH)	**	**	**	
HaptoTaxa (HAP)	**	**	**	
IndexFBI (FBI)	**	**	**	x
IndexMargalef (MAR)	**	**	**	
IndexShanWein (SW)	**	**	**	
IndexSimpDiversity (SIM)	**	**	*	
NumTrichop (T)	**	**	NS	
PercentCling (%CL)	**	**	*	x
PercentColl (%CO)	**	**	*	x
PercentDom1 (%D1)	**	**	**	x
PercentDom2 (%D2)	**	**	**	
PercentDom5 (%D5)	**	**	**	
PercentEphemeroptera (%E)	**	**	**	
PercentEPT (%EPT)	**	**	**	x
PercentEPTTaxa (%ETX)	*	**	NS	
PercentFilter (%FIL)	*	*	NS	
PercentHapto (%HAP)	**	**	NS	
ScraperTaxa (SCR)	**	**	**	
SensitiveTaxa (S)	**	**	**	
TaxRich (TR)	**	**	**	x

Data are arranged according to ecoregion: P, Piedmont; V, Valleys; H, Highlands. For all metrics: (\*\*) indicates highly significant result ( $p < 0.01$ ); (\*) indicates significant result ( $p < 0.05$ ); NS indicates nonsignificant result. Metrics selected for the index are specified.

Table 8  
Percent of stressed sites correctly assigned as stressed (<25th percentile of the reference site distribution) for all metrics

Metric	Discrimination efficiencies (as %)				
	P	V	H	All	Selected
CFTaxa (CF)	–	6	3	3	
ClingTaxa (CL)	100	94	69	82	
DipteraTaxa (DIP)	–	17	21	15	
EphemTaxa (E)	86	71	81	75	
EPT_Taxa (EPT)	86	91	81	83	x
EPTTax_NoHydro (ENH)	86	91	81	83	
HaptoTaxa (HAP)	100	82	78	80	
IndexFBI (FBI)	100	85	81	83	x
IndexMargalef (MAR)	86	65	75	70	
IndexShanWein (SW)	86	62	83	72	
IndexSimpDiversity (SIM)	86	59	75	68	
NumTrichop (T)	100	65	53	63	
PercentCling (%CL)	86	65	67	67	x
PercentColl (%CO)	79	65	67	66	x
PercentDom1 (%D1)	86	88	83	83	x
PercentDom2 (%D2)	100	82	81	82	
PercentDom5 (%D5)	100	79	89	84	
PercentEphemeroptera (%E)	100	85	83	84	
PercentEPT (%EPT)	93	79	64	72	x
PercentEPTTaxa (%ETX)	71	82	56	67	
PercentFilter (%FIL)	21	26	36	29	
PercentHapto (%HAP)	64	59	44	52	
ScraperTaxa (SCR)	86	41	47	49	
SensitiveTaxa (S)	100	97	81	87	
TaxRich (TR)	100	94	89	90	x

Data are arranged according to ecoregion: P, Piedmont; V, Valleys; H, Highlands. Metrics selected for the index are specified.

categories (Table 9). Priority was given to selecting at least 1 metric from each category.

In the parallel analysis, using RAs as quantitative counts did not significantly increase variability or decrease the discriminatory power of percentage metrics. Out of 25 proportional metrics (other than diversity indices), 10 were selected as candidates when RAs were included in calculations: percent clingers; percent collectors; percent dominance; percent 2 dominance; percent 5 dominance; percent EPT; percent EPT taxa; percent Ephemeroptera; percent filterers; percent Haptobenthos. The same metrics (except percent dominance and percent filterers) were also selected when RAs were excluded. Most candidate percentage metrics calculated using RAs (except percent filterers and percent Haptobenthos) could distinguish reference from impaired sites in most regions ( $p < 0.05$ ). Metrics that discriminated poorly in particular regions were no

more discriminating when RAs were excluded than when RAs were considered. In other words, most weak metrics (poor discrimination) remained weak whether or not RAs were considered; most strong metrics (clear discrimination) remained strong whether or not RAs were considered.

Of the richness measures, TaxRich was selected for the index because of its ubiquitousness and ecological importance as a descriptor of aquatic assemblages. This metric showed highly significant differences ( $p < 0.01$ ) between reference and stressed sites in all regions, and DE exceeded 80% in all regions. EPT was strongly correlated with the metric EPT taxa excluding Hydropsychidae, a tolerant caddisfly (EPT\_NoHydro), and both discriminated equally well. EPT was selected because it is a more widely used indicator than EPT\_NoHydro and to maintain continuity with basin states' assessment indices. Number of Ephemeroptera taxa and number of Trichoptera taxa are not recommended for family level identification (Barbour et al., 1999). Both metrics discriminated poorly in our analysis and were eliminated.

When RAs were included in diversity index calculations, Simpson's, Shannon–Weiner, and Margalef's intercorrelated with each other, and Margalef's and Shannon–Weiner were redundant with TaxRich. These indices distinguished reference from stressed streams in all regions. Excluding RAs increased homogeneity of variance between indices' reference and stressed populations and rendered all three highly redundant with TaxRich. Diversity indices also correlated negatively with percent dominance metrics. Based on these results as well as on concerns about the merits of these measures (see Section 4), we decided not to include a diversity index.

The compositional metrics percent EPT and percent EPT taxa did not intercorrelate. Percent EPT taxa was eliminated due to marginal discrimination in the Highland region and because it was redundant with EPT and EPT\_NoHydro. Percent EPT (%EPT) discriminated well in all regions and was retained.

Of the habit measures, Clinger taxa and Haptobenthic taxa discriminated effectively in all regions but were redundant with TaxRich and EPT, as well as with each other, and were rejected. Percent clingers (%Cling) and percent Haptobenthos (%Hapto) also intercorrelated. %Cling distinguished between refer-

Table 9

Pearson correlation matrix of *r* values for candidate metrics

Metric	CF	CL	DIP	E	EPT	ENH	HAP	FBI	MAR	SW	SIM	T	%CL	%CO	%D1	%D2	%D5	%E	%EPT	%ETX	%FIL	%HAP	SCR	S
CF																								
CL	0.74																							
DIP	0.29	0.33																						
E	0.54	0.71	0.24																					
EPT	0.57	0.86	0.31	0.80																				
ENH	0.52	0.84	0.31	0.79	0.99																			
HAP	0.67	0.94	0.40	0.73	0.90	0.89																		
FBI	-0.55	-0.67	-0.07	-0.51	-0.64	-0.62	-0.71																	
MAR	0.69	0.81	0.25	0.66	0.68	0.65	0.83	-0.73																
SW	0.67	0.76	0.22	0.65	0.66	0.62	0.78	-0.77	0.93															
SIM	0.56	0.62	0.21	0.52	0.55	0.51	0.63	-0.75	0.80	0.93														
T	0.60	0.74	0.28	0.44	0.78	0.76	0.75	-0.50	0.55	0.52	0.47													
%CL	0.51	0.56	0.00	0.38	0.42	0.38	0.50	-0.59	0.44	0.47	0.43	0.40												
%CO	-0.25	-0.39	0.03	-0.17	-0.33	-0.33	-0.40	0.64	-0.56	-0.55	-0.56	-0.26	-0.21											
%D1	-0.58	-0.66	-0.18	-0.52	-0.59	-0.56	-0.66	0.69	-0.82	-0.95	-0.97	-0.51	-0.46	0.56										
%D2	-0.63	-0.72	-0.18	-0.56	-0.63	-0.60	-0.72	0.69	-0.87	-0.96	-0.89	-0.54	-0.47	0.59	0.93									
%D5	-0.63	-0.77	-0.22	-0.60	-0.66	-0.64	-0.78	0.63	-0.88	-0.90	-0.70	-0.51	-0.44	0.53	0.76	0.88								
%E	0.53	0.59	0.05	0.64	0.54	0.53	0.59	-0.55	0.61	0.64	0.54	0.33	0.29	-0.16	-0.52	-0.56	-0.55							
%EPT	0.48	0.67	0.08	0.58	0.73	0.72	0.70	-0.81	0.57	0.63	0.63	0.58	0.60	-0.36	-0.58	-0.55	-0.50	0.60						
%ETX	0.31	0.57	0.19	0.56	0.82	0.83	0.62	-0.53	0.27	0.33	0.37	0.65	0.33	-0.22	-0.37	-0.34	-0.30	0.33	0.70					
%FIL	0.43	0.28	0.03	0.21	0.18	0.15	0.21	-0.31	0.15	0.15	0.15	0.25	0.70	0.16	-0.13	-0.13	-0.06	0.16	0.39	0.17				
%HAP	0.35	0.49	0.08	0.32	0.49	0.48	0.56	-0.73	0.42	0.46	0.46	0.42	0.77	-0.45	-0.47	-0.46	-0.43	0.24	0.66	0.49	0.46			
SCR	0.44	0.59	0.04	0.54	0.45	0.43	0.53	-0.39	0.64	0.55	0.43	0.40	0.35	-0.25	-0.45	-0.51	-0.54	0.44	0.33	0.16	0.19	0.18		
S	0.63	0.88	0.23	0.71	0.88	0.88	0.92	-0.74	0.79	0.76	0.60	0.65	0.49	-0.43	-0.62	-0.70	-0.77	0.62	0.72	0.63	0.19	0.54	0.49	
TR	0.73	0.86	0.30	0.74	0.74	0.71	0.88	-0.68	0.95	0.87	0.69	0.60	0.46	-0.41	-0.71	-0.78	-0.85	0.63	0.57	0.33	0.22	0.41	0.68	0.82

$r_{crit} = 0.75$ . See Tables 7 and 8 for metric abbreviations.

ence and impaired sites in all regions, although DE was poorer in the Highland and Valley regions. %Hapto also discriminated poorly in the Highlands. The unreliability of mode-of-existence designations at the family level (since all genera within a family may not have the same habit) may have influenced the discriminatory ability of these metrics (Gerritsen et al., 2000b). %Cling was retained for the index.

Of the tolerance measures, Hilsenhoff's Family Biotic Index (FBI) clearly separated reference from impaired sites in all regions. Although FBI correlated negatively with %EPT, this metric was selected because of its known reliability over a wide geographic range (Stribling et al., 1998; Maxted et al., 2000). Sensitive taxa was redundant with EPT and TaxRich and was eliminated.

Percent dominance of the single dominant taxon (%Dom1), percent dominance of the 2 dominant taxa (%Dom2) and percent dominance of the 5 dominant taxa (%Dom5), a metric used for the MAHA (Klemm et al., 2001), all distinguished reference from stressed sites in all regions. %Dom5 had the highest overall DE, followed by %Dom1 and 2. %Dom2 and 5 negatively correlated with TaxRich, while %Dom1 was not redundant. All three dominance measures were intercorrelated. %Dom1 was selected for the index.

Of the trophic metrics, percent collectors (proportion of gatherers plus filterers) differentiated reference from impaired sites in all regions and was retained for our index. Percent filterers (proportion of filterers) discriminated poorly in Piedmont streams and was rejected. Both collector–filterer taxonomic richness (CF taxa) and Scraper taxa found significant differences between reference and stressed sites but were eliminated due to poor DE. Percent Simuliidae was considered as a trophic metric but proved unreliable (C.V. > 0.9).

Separation between interquartile ranges (no overlap between 25th and 75th percentiles) was used to further evaluate metric discrimination (Fig. 2). There was clear separation for all seven selected metrics in the Piedmont region, for four metrics (EPT, FBI, %Dom1, TaxRich) in the Highlands, and for five metrics (EPT, FBI, %EPT, %Dom1, TaxRich) in the Valleys. Although %Coll showed some overlap of the interquartile ranges (IQRs) of reference and stressed samples in the Highland and Valley regions, the

medians of either population of samples were well outside the IQRs of the other population of samples for this metric. This was also the case for %EPT in the Highlands.

Results of *t*-tests comparing regional reference communities using core metrics are presented in Table 10. Reference communities in the Piedmont region appeared significantly different ( $p < 0.05$ ) from those in the Highlands for EPT and %EPT, and from the Valleys for EPT, %Cling, and %EPT. Reference communities in the Highland and Valley regions were not significantly different from each other for any core metric at  $p \leq 0.05$ , although there were moderate differences at  $p \leq 0.10$ .

#### 4. Discussion

Improved collaboration and coordination among state monitoring entities promotes data sharing and improves ecological inferences. However, agencies and programs will likely continue to select bioassessment protocols that are compatible with their own monitoring objectives. The proposed approach for synthesizing and evaluating disparate datasets provides a means for assessing interstate wadeable stream habitats and associated macroinvertebrate communities in a consistent, scientifically defensible manner. Data from dissimilar methods were successfully combined to produce common reference and impairment benchmarks for the Potomac River basin. Bioindicators sensitive to the differences between reference and degraded habitats were identified. This study confirms that information useful to managers can be generated from diverse datasets.

##### 4.1. Habitat metrics

Of the habitat parameters (or equivalents) collected by basin states, Anthropogenic Alteration, Riparian Zone, and Channel Alteration were best at distinguishing high-quality sites from those with severe degradation. These findings are compatible with those reported by Kaufmann et al. (1999) for the mid-Atlantic. Bank Stability and Substrate Quality were also useful measures. These parameters are considered important by the basin states for establishing reference and impairment criteria. MdDNR's MBSS, WVDEP's

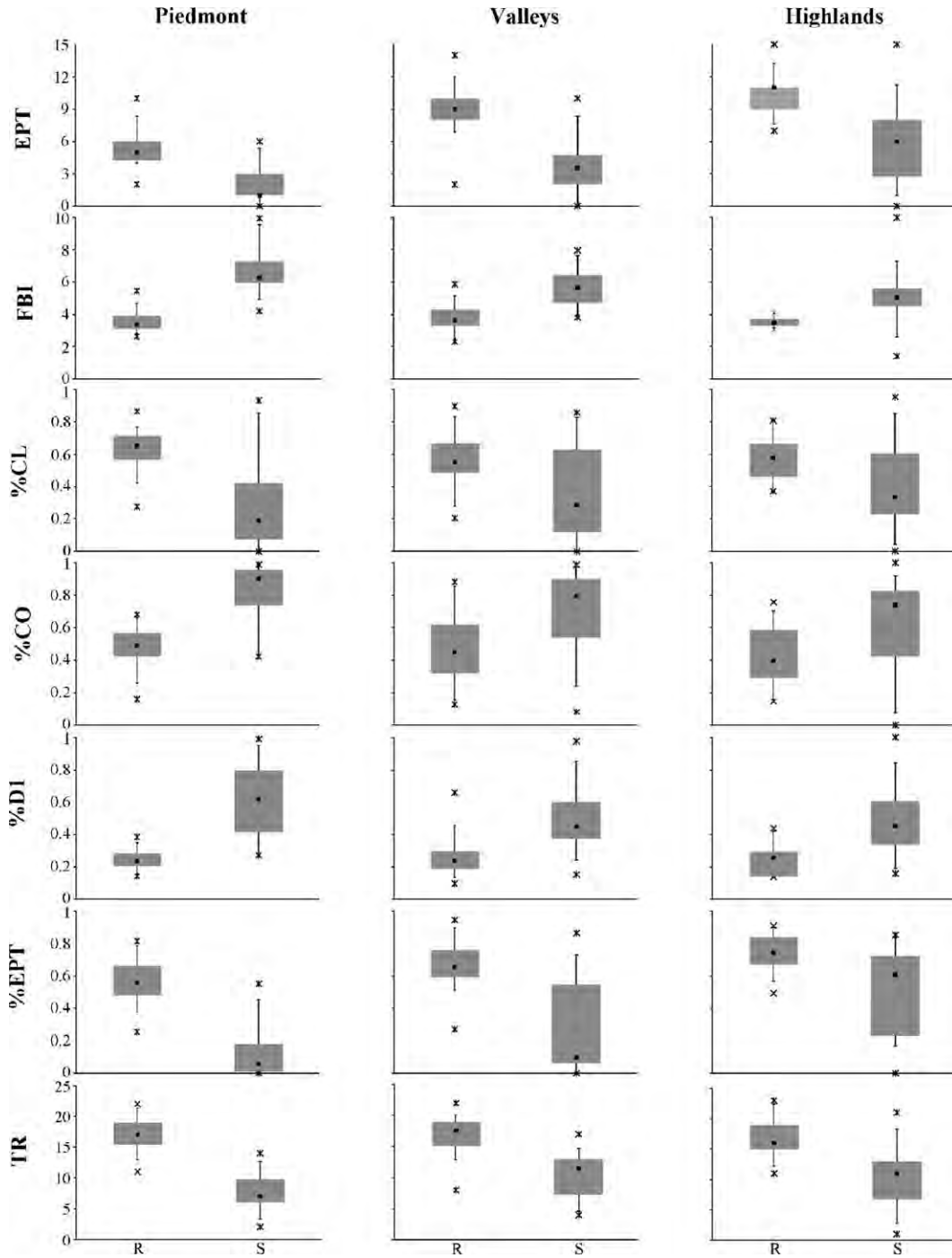


Fig. 2. Distributions of values for core metrics. Boxes represent the 25th and 75th percentiles; whiskers represent the 5th and 95th percentiles; squares represent median values; asterisks represent maximum and minimum values. R, reference sites; S, stressed sites.

Stream Condition Index (WVSCI), and the Stream Condition Index (SCI) under development for VaDEQ all emphasize stream channel changes, bank erosion, and extent of riparian buffer as selection criteria (Gerritsen et al., 2000b; Roth et al., 2001; Burton and Gerritsen, 2003). MdDNR uses percent embeddedness (an equivalent of Substrate Quality) as an additional criterion (Roth et al., 2001). Agencies also use these parameters in field habitat assessments. PaDEP performs field assessments using these parameters but does not apply selection criteria. The basin jurisdictions also consider evidence of human activity and impacts when evaluating habitat quality. “Grazing or Disruptive Pressure,” “Aesthetics/Remoteness,” and their equivalents rate the level of anthropogenic effects or disturbance (Roth et al., 1997; Lazorchak et al., 1998; Wirts, personal communication, 2001). The other parameters are geomorphic factors that are themselves subject to anthropogenic stress.

#### 4.2. Determination of strata

The use of ecological regions as natural geographic units for classifying waterbodies is well documented (Omernik and Griffith, 1991; Hughes et al., 1994; Omernik, 1995). Biological indices calibrated for ecoregions or watersheds are valuable assessment tools (Barbour et al., 1995; Stribling et al., 1998). Significant biotic variation related to ecoregion was reported by Feminella (2000) in the Southeastern US, Pan et al. (2000) in the mid-Atlantic Highlands, and Van Sickle and Hughes (2000) in Oregon. Gerritsen et al. (2000a) found that aggregating ecoregions into mountainous and nonmountainous groupings strengthened the classification of Wyoming streams.

We considered several ecoregional schemes: Omernik (1987); Woods et al. (1999); US EPA (2000) (Table 3). The framework developed by Omernik (1987) for US EPA has evolved significantly since its introduction. After the 1:7,500,000-scale resolution of Omernik (1987) ecoregions was found to be inadequate for states’ needs, Woods et al. (1999) refined the Level III ecoregions of the mid-Atlantic to a 1:250,000 scale and delineated more detailed Level IV subregions. As a result of these refinements, ecoregion 64 “lost” stream sites to ecoregions 65 and 45, but also “gained” sites from the former ecoregion 66. Ecoregion 45 was under-represented in the basin

dataset ( $n = 3$ ), so we aggregated ecoregions 64 and 45 into a Piedmont grouping. Many sites formerly in ecoregions 66 and 69 were reclassified by Woods et al. (1999) in ecoregion 67. Consequently, the basin portion of ecoregion 67 increased substantially in area, resulting in higher site  $n$  and greater between-site variability, while ecoregions 66 and 69 lacked sufficient  $n$  for analysis. For the MAHA, the US EPA (2000) combined similar Level III ecoregions, and separated ecoregion 67 into its ridge and valley components based on Level IV subregions after Waite et al. (2000). We adapted this classification to produce “Ridge” and “Valley” groupings of basin streams.

The Potomac portion of ecoregion 69 forms a relatively small and narrow plateau in the western basin. Some of its waters originate in the far western portion of ecoregion 67 behind the plateau and dissect the plateau into steep-sided valleys. Other streams (including the North Branch Potomac) originate on the surface of the plateau. Streams re-enter ecoregion 67 east of the Allegheny Front. While we were able to define reference conditions for ecoregion 69, only 1 site met the criteria. This was clearly inadequate for statistical significance. We explored whether ecoregion 69 could feasibly be combined with other Potomac regions, or should be evaluated as a distinct site class using some other means. At the broad regional scale, US EPA’s MAHA considered ecoregion 69 to be more similar to other montane ecoregions (not within the basin) than to either the Ridge or Valley aggregated regions (US EPA, 2000). At the jurisdictional scale, Burton and Gerritsen (2003) suggested that Virginia’s portion of ecoregion 69 might be distinct from ecoregions 66 and 67, but could not be certain as this region was under-represented in VaDEQ data; they recommended using the WVSCI to evaluate ecoregion 69 sites. On the other hand, neither WVDEP (Gerritsen et al., 2000b) nor MdDNR (Stribling et al., 1998; Boward et al., 1999) found ecoregion 69 distinct enough to warrant separate classification. We decided to combine sites in the basin’s portion of ecoregion 69 with those in the Ridges grouping (“Highlands”) and compare them to sites in the Valley and Piedmont groupings (Fig. 1). These ecoregion-based groupings effectively partitioned the variability of Potomac reference sites.

While the classificatory value of primary sub-watersheds could not be fully evaluated, our results

suggest subwatersheds represent ecologically “real” subunits of the basin. Smaller catchments (e.g. 3rd–4th order catchments) can be as useful as ecoregions for geographic classifications (Waite et al., 2000), and in some cases more so (McCormick et al., 2000; Omernik, personal communication, 2004). Ecoregions (or groupings of ecoregions) and catchments can be used together in a hierarchical approach to develop a system of regional reference sites and improve understanding of the attainable quality, integrity, and health of aquatic ecosystems (Omernik and Bailey, 1997; Hawkins et al., 2000). Classification of Potomac streams by subwatershed tiered with region (and other classification strata, e.g. soil type) will be evaluated in the future as more data are incorporated.

#### 4.3. *Biological metrics*

Diversity indices such as Simpson’s, Shannon–Weiner, and Margalef’s are widely used in bioassessment. Simpson’s index accounts for the richness of each taxon within a community, while Shannon–Weiner accounts for the relationship between richness and evenness (Gove et al., 1996). As these measures are not independent of taxonomic richness, sample sizes must be the same (or rarefaction techniques applied) to compare these metrics between communities (Boyle et al., 1990; Norris and Georges, 1993). Sampling methods, habitat types, and area sampled must also be consistent (Norris and Georges, 1993). These presented problems in our analysis due to the disparities among the datasets. As expected, Shannon–Weiner was redundant with TaxRich, although Simpson’s was not. Margalef’s minimizes the effect of sample size bias without using rarefaction. However, it is a function of taxonomic richness as well, and was redundant with TaxRich in our analysis, particularly when RAs were excluded from calculations. Diversity indices also intercorrelated with the percent dominance metrics. Percent dominance is often redundant with diversity indices when dominance by pollution-tolerant organisms is high (Barbour et al., 1996). The biological relevance of diversity indices has been strongly questioned because their values can be influenced by study design, and because the underlying assumptions of some are invalid (Hurlbert, 1971; Boyle et al., 1990; Norris and

Georges, 1993; Norris, 1995; Lydy and Crawford, 2000). For these reasons, we elected not to include a diversity measure.

Numerous trophic measures were tested to ensure that this aspect of the community was captured by the index. As noted, percent collectors (%Coll) was able to differentiate reference from impaired sites in all basin regions. Some disagreement exists regarding this and certain other feeding metrics’ responses to perturbation. For example, Gerritsen et al. (2000b) predicted that %Coll would decrease in response to stress, but the WVSCI analysis found the trend to be opposite from expected. Others (e.g. Barbour et al., 1999) define the expected response for this metric as “Variable” or “Unknown.” We expected that collectors, generalists that feed on suspended detritus, would increase in response to perturbation. This was confirmed by the IQR plots for %Coll, which showed significantly higher values for this metric at stressed sites compared to reference sites in all regions.

Responses of feeding metrics are sometimes tied to dominance by a particular taxon. Grouping taxa by feeding mode at reference and impaired sites in each region revealed high proportions of filtering blackflies (Simuliidae) at stressed sites. Percent Simuliidae exhibited high variability in the reference population and was not a useful indicator.

Collector–filter taxonomic richness (CF taxa) is used in the Macroinvertebrate Biotic Integrity Index (MBII) developed by Klemm et al. (2001) for the MAHA. Although both CF taxa and Scraper taxa found significant differences between reference and stressed sites, statistical significance alone does not make a metric useful for bioassessment. A metric with very low variation will have high statistical significance, but a metric demonstrating little variation in response to change is not a useful indicator (Marshall, 2001). On closer examination, reference and stressed sites identified to be significantly different differed by only one or two families. This was supported by the poor DE’s for these metrics. The MBII relies on taxonomic identifications to genus/species. Neither CF taxa nor Scraper taxa demonstrated functional differences at the family level.

Percent clingers (%Cling) is recommended as a “best candidate metric” for the habit category (Barbour et al., 1999). We found %Cling to be an

excellent indicator in the Piedmont, but less so in the Highland and Valley regions. As previously noted, habit metrics may be unreliable for family-level data (Gerritsen et al., 2000b). On the other hand, habit metrics have been reported to be more robust than trophic metrics in some instances (Fore et al., 1996). We decided to retain %Cling, although its inclusion will be re-evaluated during index development.

We could not select and test metrics for the Potomac portion of ecoregion 65 due to the lack of reference *n*. All ecoregion 65 data in our dataset were collected by MdDNR. MdDNR's family level benthic index of biotic integrity (B-IBI) for Maryland's Coastal Plain province (Stribling et al., 1998) was developed using these data. We will use the Coastal Plain B-IBI to assess sites in the ecoregion 65 portion of the basin.

#### 4.4. Reference condition comparisons

We identified a significant segregation between the reference communities of the Piedmont and reference communities in the remaining regions, with Piedmont and Valley communities showing the greatest dissimilarity. The MBSS stratifies streams into Coastal Plain and non-Coastal Plain bioregions (Stribling et al., 1998), and this classification has been proposed for the Virginia SCI (Burton and Gerritsen, 2003). However, Burton and Gerritsen (2003) found moderate separation of Northern Piedmont (64) biota from samples in other regions of Virginia, and suggested that Piedmont streams were distinct from those in montane ecoregions. Because of our interstate approach, we had more data from the Piedmont for our analysis. Our results corroborate Burton and Gerritsen's (2003) recommendation.

Our results also indicated that reference communities in the Highland and Valley regions were moderately similar. These findings agree with Stribling et al. (1998), Smith and Voshell (1997), Gerritsen et al. (2000b), and others in the mid-Atlantic. However, as previously noted, the MAHA treated these as distinct regions. We expected them to differ based on general land use patterns, and because our aggregated regions were adapted from MAHA ecoregions. We expected these differences to be reflected in both habitat quality and macroinvertebrate community health.

We examined reference communities and habitat conditions across land use types to determine whether

Valley and Ridge streams were different enough to warrant evaluating them separately, and to confirm that streams in the basin portion of ecoregion 69 were similar enough to those in the Ridges to warrant combining them. Using GIS, ICPRB's Potomac basin map file was intersected with land use characterizations based on the Multi-Resolution Land Characteristics Consortium (MRLC)'s 2001 National Land Cover Database (NLCD) for Region 3 ([http://www.mrlc.gov/mrlc2k\\_nlcd.asp](http://www.mrlc.gov/mrlc2k_nlcd.asp)). Individual NLCD land use classes (e.g. low intensity urban, coniferous forest) were combined into four aggregated classes: developed/urban; forested/vegetated; agricultural/open; water (Mercurio et al., 1999). We considered the 30 m scale of the NLCD (in other words, the dominant surrounding land use which would be more or less visible from the stream) acceptable for assigning a land use class to a site. Sites were grouped by land use class within regions, and habitat reference values (derived from regional reference criteria) and reference communities were compared between regions (Tables 10 and 11). The western basin contains few urban areas, so we focused on the agricultural/open and forested/vegetated classes. Both land use classes were present in the Valley and Ridge/Highland regions, though the Ridges/Highlands had fewer agricultural sites. Good- and poor-quality sites were found in both classes in both regions. Habitat reference values were generally higher at forested sites than agricultural sites. Reference values of Ridge/Highland sites were generally higher than Valley sites for agricultural, forested, and combined land uses. We found no significant differences between Ridge/Highland and Valley reference communities at the family level at  $p \leq 0.05$  regardless of land use class, although there were slight differences for some core metrics (EPT, %EPT, FBI; also %Dom1 at Forested sites) at  $p \leq 0.10$ . A lower level of taxonomy (genus/species) might have revealed landscape patterns with stronger statistical power (Waite et al., 2000). Due to insufficient *n* in ecoregion 69, we could not determine if a distinct enough segregation of reference communities occurred to warrant separation of this ecoregion from other basin regions for bioassessment. However, ecoregion 69 reference values were very similar to those in both the Ridges and the Highlands (Ridges + ecoregion 69) across land use types. These findings suggest that habitat quality in Ridge/Highland and Valley streams differs even when land use types are

Table 10  
Results of *t*-tests between core metric values of reference sites grouped by ecoregion: P, Piedmont; V, Valleys, H, Highlands

Ecoregion	Ecoregion	EPT	FBI	%CL	%CO	%D1	%EPT	TR
P	H	**	ss	ss	ss	ss	**	ss
P	V	**	ss	*	ss	ss	**	ss
H	V	ss	ss	ss	ss	ss	ss	ss

Reference communities were considered significantly similar (ss) at  $p > 0.05$ ; (\*) indicates significantly different ( $0.01 < p < 0.05$ ); (\*\*) indicates highly significantly different ( $p < 0.01$ ).

similar. We will continue to evaluate these regions separately. Ecoregion 69 streams appear sufficiently similar to Ridge streams to support their consolidation into a single “Highlands” bioregion in the Potomac basin.

#### 4.5. Can relative abundances be used as quantitative counts?

Our results suggest that at family-level taxonomic resolution, census data using RAs appear comparable

to data collected using semiquantitative (e.g. fixed-count) sampling and subsampling methods and can be analyzed quantitatively. As a caution, it should be emphasized that the datasets used in this analysis were collected and processed using variants of the same basic methodology (RBP I, II, or III), and that combining data collected by more widely divergent methodologies may not be feasible. But this approach may offer an acceptable compromise between using only quantitative counts while discarding some datasets, and aggregating all data to the lowest

Table 11  
Land use comparisons in the western regions of the Potomac River basin: V, Valleys, H, Highlands, R, Ridges, 69 = Central Appalachian ecoregion

Parameter code	69	H	R	V
Agriculture/open	( <i>n</i> = 3)	( <i>n</i> = 82)	( <i>n</i> = 79)	( <i>n</i> = 276)
ANTHRO_ALT	–	15	15	15
BANK_STAB	–	14	14	13
CHAN_ALT	–	14	14	15
HAB_HETERO	–	9	10.8	8
INSTR_COND	–	8.6	9.4	7.75
RIP_ZONE	–	10	10	10
SUB_QUAL	–	16	16	15
Forested/vegetated	( <i>n</i> = 47)	( <i>n</i> = 293)	( <i>n</i> = 246)	( <i>n</i> = 221)
ANTHRO_ALT	18	18	18	16.5
BANK_STAB	16	16	16	15
CHAN_ALT	16	16	16	16
HAB_HETERO	10.1	10.5	10.5	10
INSTR_COND	7.9	9.1	9.5	9
RIP_ZONE	18	19	20	17
SUB_QUAL	15.2	17	17	16
All land use classes	( <i>n</i> = 53)	( <i>n</i> = 391)	( <i>n</i> = 338)	( <i>n</i> = 553)
ANTHRO_ALT	17.4	17.5	17.5	16
BANK_STAB	16	16	16	14
CHAN_ALT	16	16	16	16
HAB_HETERO	9.1	9.5	10	8
INSTR_COND	7	9	9.5	8
RIP_ZONE	17.4	18	18	12
SUB_QUAL	16	17	17	15

Parameter values are derived from regional reference criteria in Table 5. Dashes indicate insufficient *n* for statistical significance.

common denominator (e.g. presence–absence) when combining datasets from multiple sources.

#### 4.6. Management and policy implications

Under the Clean Water Act (40 CFR 130–131), each state must produce water quality inventories, identify bodies of water that are not meeting applicable water quality standards, and report results to the EPA. Inconsistencies in the way states assess water quality lead not only to reporting inconsistencies but also to difficulties in identifying impaired waters within a region or nationwide (US GAO, 2002). Poor cooperation between neighboring states impedes data sharing and can result in duplicated sampling efforts or the underutilization of available information (Lenz and Miller, 1996).

The Methods and Data Comparability Board, a multiagency partnership under the auspices of the National Water Quality Monitoring Council, has identified numerous potential benefits to states to work cooperatively to assess their waters and share their data, including:

- increases spatial, temporal, or taxonomic scale of an assessment beyond a single limited dataset;
- increases the pool of reference sites available to any one state;
- enhances ability to make scientifically defensible judgments on the condition of the nation's waters;
- provides decision makers with more complete and reliable information with which to devise and implement monitoring strategies;
- increases the ability to use data produced by other programs, which will encourage collaboration;
- helps reduce costs on many levels, from helping managers efficiently design and implement more effective programs, to reducing the number of sites requiring sampling.

Our results confirm that bioassessment data from multiple sources can be combined to produce scientifically sound information about the quality of aquatic resources. Data synthesis can be a conceptually sophisticated and resource-intensive task. Merely combining multiple datasets within a database does not make the data comparable. Underlying differences and the compromises made to resolve

them can affect confidence in the results of the analysis (Lenz and Miller, 1996; Lenz, 1997; Sovell and Vondracek, 1999; Hale, 2000; McLaughlin et al., 2001; Stribling and Bressler, 2001). Considerable care and effort were required to address methods and data variability during this analysis. Our high degree of statistical significance for most tests indicates that our “corrective measures” were effective, and that underlying methods and data differences did not adversely affect our results.

#### Acknowledgements

The author would like to thank the following state agencies who contributed their data for this project: Maryland Department of Natural Resources, Pennsylvania Department of Environmental Protection, Virginia Department of Environmental Quality, and West Virginia Department of Environmental Protection. Funds from the EPA 106 Grant and the Interstate Commission on the Potomac River Basin supported this analysis effort. The comments and suggestions of Jerry Diamond, David Rosenberg, Richard Norris, Carlton Haywood, Jim Omernik, and two anonymous reviewers greatly improved this manuscript and are sincerely appreciated. Claire Buchanan's encouragement and support are gratefully acknowledged.

#### References

- Alden III, R.W., Perry, E., 1997. Presenting Measurements of Status: Report to the Chesapeake Bay Program Monitoring Subcommittee's Data Analysis Workgroup. Chesapeake Bay Program Annapolis, Maryland.
- Bailey, R.C., Norris, R., Reynoldson, T., 2001. Taxonomic resolution of benthic macroinvertebrate communities in bioassessments. *J. North Am. Benthol. Soc.* 20 (2), 280–286.
- Barbour, M.T., Gerritsen, J., 1996. Subsampling of benthic samples: a defense of the fixed-count method. *J. North Am. Benthol. Soc.* 15 (3), 386–391.
- Barbour, M.T., Gerritsen, J., Griffith, G., Frydenborg, R., McCarron, E., White, J., Bastian, M., 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. *J. North Am. Benthol. Soc.* 15 (2), 185–211.
- Barbour, M.T., Gerritsen, J., Snyder, B., Stribling, J., 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton Benthic Macroinvertebrates and Fish, second ed. EPA 841-B-99-002. US Environmental Protection Agency, Office of Water, Washington, DC.

- Barbour, M.T., Stribling, J., Karr, J., 1995. Multimetric approach for establishing biocriteria and measuring biological condition. In: Davis, W.S., Simon, T.P. (Eds.), *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making*. Lewis Publishers, Boca Raton, FL, pp. 63–77.
- Boward, D., Kazyak, P., Stranko, S., Hurd, M., Prochaska, T., 1999. From the Mountains to the Sea: The State of Maryland's Freshwater Streams. EPA 903-R-99-023. Maryland Department of Natural Resources, Monitoring and Non-tidal Assessment Division, Annapolis, Maryland.
- Boyle, T.P., Smillie, G., Anderson, J., Beeson, D., 1990. A sensitivity analysis of nine diversity and seven similarity indices. *Res. J. WPCF* 62 (6), 749–762.
- Burton, J., Gerritsen, J., 2003. A Stream Condition Index for Virginia Non-Coastal Streams. Tetra Tech, Inc., Owings Mills, Maryland.
- Carter, J.L., Resh, V., 2001. After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. *J. North Am. Benthol. Soc.* 20 (4), 658–682.
- Courtemanch, D.L., 1996. Commentary on the subsampling procedures used for rapid bioassessment. *J. North Am. Benthol. Soc.* 15 (3), 381–385.
- Davis, W., 2001. Best practices for the development and interpretation of biological indices: lessons learned in MAIA. In: *Mid-Atlantic Water Pollution Biology Workshop*, 2001 March 29–30, Cacapon, West Virginia.
- Diamond, J.M., Barbour, M., Stribling, J., 1996. Characterizing and comparing bioassessment methods and their results: a perspective. *J. North Am. Benthol. Soc.* 15 (4), 713–727.
- Feminella, J.W., 2000. Correspondence between stream macroinvertebrate assemblages and 4 ecoregions of the southeastern USA. *J. North Am. Benthol. Soc.* 19 (3), 442–461.
- FMS, Inc., 1997. Total Access Statistics for Microsoft Access: Statistical Analysis for Microsoft Access Data. Financial Modeling Specialists, Inc., Vienna, Virginia.
- Fore, L., Karr, J., Wiseman, R., 1996. Assessing invertebrate responses to human activities: evaluating alternative approaches. *J. North Am. Benthol. Soc.* 15 (2), 212–231.
- Fore, L., 2003. Developing Biological Indicators: Lessons Learned from Mid-Atlantic Streams. Report prepared for EPA under Contract No. 50-CMAA-900065. EPA 903/R-003/003. U.S. Environmental Protection Agency, Office of Environmental Information and Mid-Atlantic Integrated Assessment Program, Region 3, Ft. Meade, Maryland.
- Gerritsen, J., Barbour, M., King, K., 2000a. Apples, oranges, and ecoregions: on determining pattern in aquatic assemblages. *J. North Am. Benthol. Soc.* 19 (3), 487–496.
- Gerritsen, J., Burton, J., Barbour, M., 2000b. A Stream Condition Index for West Virginia Wadeable Streams. Tetra Tech, Inc., Owings Mills, Maryland.
- Gove, J.H., Patil, G., Taillie, C., 1996. Diversity measurement and comparison with examples. In: Szaro, R.C., Johnston, D.W. (Eds.), *Biodiversity in Managed Landscapes*. Oxford University Press, New York, pp. 157–175.
- Growth, J., Chessman, B., Jackson, J., Ross, D., 1997. Rapid assessment of Australian rivers using macroinvertebrates: cost and efficiency of 6 methods of sample processing. *J. North Am. Benthol. Soc.* 16 (3), 682–693.
- Hale, S., 2000. How to Manage Data Badly (Part 2). *Bulletin of the Ecological Society of America*, January 2000, pp. 101–103.
- Handcock, M.S., Sedransk, J., Olsen, A., 2002. Statistical methods for ecological assessment of riverine systems by combining information from multiple sources. In: *Proceedings of the Section on Environmental Statistics of the American Statistical Societies Meetings*. American Statistical Association, Alexandria, Virginia.
- Hawkins, C.P., Norris, R., Gerritsen, J., Hughes, R., Jackson, S., Johnson, R., Stevenson, R., 2000. Evaluation of the use of landscape classifications for the prediction of freshwater biota: synthesis and recommendations. *J. North Am. Benthol. Soc.* 19 (3), 541–556.
- Hewlitt, R., 2000. Implications of taxonomic resolution and sample habitat for stream classification at a broad geographic scale. *J. North Am. Benthol. Soc.* 19 (2), 352–361.
- Hilsenhoff, W.L., 1988. Rapid field assessment of organic pollution with a family-level biotic index. *J. North Am. Benthol. Soc.* 7 (1), 65–78.
- Hughes, R.M., Heiskarsky, S., Matthews, W., Yoder, C., 1994. Use of ecoregions in biological monitoring. In: Loeb, S., Spacie, A. (Eds.), *Biological Monitoring of Aquatic Systems*. Lewis Publishers, Boca Raton, FL, pp. 125–150.
- Hurlbert, S.H., 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52 (4), 577–586.
- Kaufmann, P.R., Levine, P., Robison, E., Seeliger, C., Peck, D., 1999. Quantifying Physical Habitat in Wadeable Streams. EPA/620-R-99/003. US Environmental Protection Agency, Washington, DC.
- Kerans, B.L., Karr, J., Ahlstedt, S., 1992. Aquatic invertebrate assemblages: spatial and temporal differences among sampling protocols. *J. North Am. Benthol. Soc.* 11 (4), 377–390.
- Klemm, D.J., Blocksom, K., Fulk, F., Herlihy, A., Hughes, R., Kaufmann, P., Peck, D., Stoddard, J., Thoeny, W., Griffith, M., Davis, W., 2001. A macroinvertebrate biotic integrity index (MBII) for regional assessment of mid-Atlantic highland streams. In: *Society for Environmental Toxicology and Chemistry 22nd Annual Meeting*, 2001 November 11–15, Baltimore, Maryland.
- Larson, D.P., Herlihy, A., 1998. The dilemma of sampling streams for macroinvertebrate richness. *J. North Am. Benthol. Soc.* 17 (3), 359–366.
- Lazorchak, J.M., Klemm, D., Peck, D. (Eds.), 1998. *Environmental Monitoring and Assessment Program—Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Wadeable Streams*. EPA/620-R-94/004F. US Environmental Protection Agency, Washington, DC.
- Lenz, B.N., Miller, M., 1996. Comparison of Aquatic Macroinvertebrate Samples Collected Using Different Field Methods. Fact Sheet FS-216-96. US Geological Survey, National Water Quality Assessment Program, Madison, Wisconsin.
- Lenz, B.N., 1997. Feasibility of Combining Two Aquatic Benthic Macroinvertebrate Community Databases for Water-Quality Assessment. Fact Sheet FS-132-97. US Geological Survey,

- National Water Quality Assessment Program, Madison, Wisconsin.
- Lydy, M.J., Crawford, C., Frey, J., 2000. A comparison of selected diversity, similarity, and biotic indices for detecting changes in benthic-invertebrate community structure and stream quality. *Arch. Environ. Contam. Toxicol.* 39, 469–479.
- Marshall, B.D., 2001. An evaluation of the sensitivity of a macro-invertebrate biomonitoring study in headwater streams of New River Gorge National River. *J. Freshwater Ecol.* 16 (3), 415–428.
- Maxted, J.R., Barbour, M., Gerritsen, J., Poretti, V., Primrose, N., Silvia, A., Penrose, D., Renfrow, R., 2000. Assessment framework for mid-Atlantic coastal plain streams using benthic macroinvertebrates. *J. North Am. Benthol. Soc.* 19 (1), 128–144.
- McCormick, F.H., Peck, D., Larsen, D., 2000. Comparison of geographic classification schemes for mid-Atlantic stream fish assemblages. *J. North Am. Benthol. Soc.* 19 (3), 385–404.
- McLaughlin, R.L., Carl, L., Middel, T., Ross, M., Noakes, D., Hayes, D., Baylis, J., 2001. Potentials and pitfalls of integrating data from diverse sources: lessons from a historical database for Great Lakes stream fishes. *Fisheries* 26 (7), 14–23.
- Mercurio, G., Chaillou, J., Roth, N., 1999. Guide to Using 1995–1997 Maryland Biological Stream Survey Data. Prepared for Maryland Department of Natural Resources under Contract No. PR-96-055-001. Maryland Department of Natural Resources, Monitoring and Non-Tidal Assessment Division, Annapolis, MD.
- Methods and Data Comparability Board, 2003. Fact Sheet: The Value of Data Comparability. ([http://wi.water.usgs.gov/methods/about/publications/valcomp\\_fs.pdf](http://wi.water.usgs.gov/methods/about/publications/valcomp_fs.pdf)).
- Norris, R.H., 1995. Biological monitoring: the dilemma of data analysis. *J. North Am. Benthol. Soc.* 14 (3), 440–450.
- Norris, R.H., Georges, A., 1993. Analysis and interpretation of benthic macroinvertebrate surveys. In: Rosenberg, D., Resh, V. (Eds.), *Freshwater Biomonitoring and Benthic Macroinvertebrates*. Routledge, Chapman and Hall, Inc., New York, pp. 234–286.
- Olson, M., 2002. Benchmarks for nitrogen, phosphorus, chlorophyll and suspended solids in Chesapeake bay. Chesapeake Bay Program Technical Report Series, Chesapeake Bay Program, Annapolis, Maryland.
- Omerik, J.M., 1987. Ecoregions of the conterminous United States. *Ann. Assoc. Am. Geographers* 77, 118–125.
- Omerik, J.M., 1995. Ecoregions: a spatial framework for environmental management. In: Davis, W.S., Simon, T.P. (Eds.), *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making*. Lewis Publishers, Boca Raton, FL, pp. 49–62.
- Omerik, J.M., Bailey, R., 1997. Distinguishing between watersheds and ecoregions. *J. Am. Water Res. Assoc.* 33 (5), 935–949.
- Omerik, J.M., Griffith, G., 1991. Ecological regions vs. hydrologic units: frameworks for managing water quality. *J. Soil Water Conserv.* 46, 334–340.
- Pan, Y., Stevenson, R., Hill, B., Herlihy, A., 2000. Ecoregions and benthic diatom assemblages in mid-Atlantic Highlands streams, USA. *J. North Am. Benthol. Soc.* 19 (3), 518–536.
- Pennsylvania Department of Environmental Protection, 2003. Pennsylvania's State-Wide Surface Waters Assessment Program: 2003 Update. 9 May 2003. Pennsylvania Department of Environmental Protection, Office of Water Management, 22 July 2003. (<http://www.dep.state.pa.us/dep/deputate/watermgmt/Wqp/WQStandards/UnassesWater.htm>).
- Plafkin, J.L., Barbour, M., Porter, K., Gross, S., Hughes, R., 1989. Rapid Bioassessment Protocols for Use in Streams and Rivers: Benthic Macroinvertebrates and Fish. EPA 440-4-89-001. US Environmental Protection Agency, Office of Water Regulations and Standards, Washington, DC.
- Roth, N.E., Southerland, M., Mercurio, G., Volstad, J., 2001. Maryland Biological Stream Survey 2000–2004. Volume I: Ecological Assessment of Watersheds Sampled in 2000 CBWP-MANTA-EA-01-5. Maryland Department of Natural Resources, Chesapeake Bay and Watershed Programs, Monitoring and Non-Tidal Assessment Division, Annapolis, Maryland.
- Roth, N.E., Southerland, M., Chaillou, J., Volstad, J., Weisberg, S., Wilson, H., Heimbuch, D., Seibel, J., 1997. Maryland Biological Stream Survey: Ecological Status of Non-Tidal Streams in Six Basins Sampled in 1995. CBWP-MANATA-EA-97-2. Maryland Department of Natural Resources, Chesapeake Bay and Watershed Programs, Monitoring and Non-Tidal Assessment Division, Annapolis, Maryland.
- Smith, E.P., Voshell, R., 1997. Studies of Benthic Macroinvertebrates and Fish in Streams Within EPA Region 3 for Development of Biological Indicators of Ecological Condition. Part I. Benthic Macroinvertebrates. Virginia Tech, Blacksburg, Virginia.
- Sovell, L.A., Vondracek, B., 1999. Evaluation of the fixed-count method for rapid bioassessment protocol III with benthic macroinvertebrate metrics. *J. North Am. Benthol. Soc.* 18 (3), 420–426.
- Stribling, J.B., Bressler, D., 2001. Defining Analytical Truth and Evaluation of Accuracy in Biological Assessments, 3 July 2001. Methods and Data Comparability Board, 30 November 2001. ([http://wi.water.usgs.gov/methodsboard/madison/accuracy\\_manuscript.htm](http://wi.water.usgs.gov/methodsboard/madison/accuracy_manuscript.htm)).
- Stribling, J.B., Jessup, B., White, J., Boward, D., Hurd, M., 1998. Development of a Benthic Index of Biotic Integrity for Maryland Streams. CBWP-EA-98-3. Maryland Department of Natural Resources, Chesapeake Bay and Watershed Programs, Monitoring and Non-Tidal Assessment Division, Annapolis, Maryland.
- Thorne, R.St.J., Williams, W., Cao, Y., 1999. The influence of data transformations on biological monitoring studies using macroinvertebrates. *Water Res.* 33 (2), 343–350.
- US EPA (US ENVIRONMENTAL PROTECTION AGENCY), 1997. Guidelines for the Preparation of the Comprehensive State Water Quality Assessments (305[b] Reports). EPA-841-B-97-002A. US Environmental Protection Agency, Office of Water, Washington, DC.
- US EPA (U S ENVIRONMENTAL PROTECTION AGENCY), 2000. Mid-Atlantic Highlands Stream Assessment: Technical Support Document. EPA/903/B-00/004. US Environmental Protection Agency, Office of Research and Development and Region 3, Mid-Atlantic Integrated Assessment Program, Ft. Meade, Maryland.
- US EPA (US ENVIRONMENTAL PROTECTION AGENCY), 2002. The Watershed Approach. Watersheds. US Environmental

- Protection Agency, Office of Wetlands, Oceans, and Watersheds. (<http://www.epa.gov/owow/watershed/wal.html>).
- US GAO (US GOVERNMENT ACCOUNTABILITY OFFICE), 2002. Inconsistent State Approaches Complicate Nation's Efforts to Identify its Most Polluted Waters. GAO-02-186. US General Accounting Office, Washington, DC.
- Van Sickle, J., Hughes, R., 2000. Classification strengths of ecoregions, catchments, and geographic clusters for aquatic vertebrates in Oregon. *J. North Am. Benthol. Soc.* 19 (3), 385–404.
- Vinson, M.R., Hawkins, C., 1996. Effects of sampling area and subsampling procedure on comparisons of taxa richness among streams. *J. North Am. Benthol. Soc.* 15 (3), 392–399.
- Waite, I., Herlihy, A., Larsen, D., Klemm, D., 2000. Comparing strengths of geographic and nongeographic classifications of stream benthic macroinvertebrates in the Mid-Atlantic Highlands, USA. *J. North Am. Benthol. Soc.* 19 (3), 429–441.
- Woods, A., Omernik, J., Brown, D., 1999. Level III and IV Ecoregions of Delaware, Maryland, Pennsylvania, Virginia, and West Virginia. U.S. Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Corvallis, Oregon. Report with map supplement, Scale 1:1,000,000, September 1999.