Calibration and validation of a regionally and seasonally stratified macroinvertebrate index for West Virginia wadeable streams

Gregory J. Pond · Jeffrey E. Bailey ·
Benjamin M. Lowman · Michael J. Whitman

Abstract Multimetric indices (MMIs) are routinely used by federal, state, and tribal entities to assess the quality of aquatic resources. Because of their diversity, abundance, ubiquity, and sensitivity to environmental stress, benthic macroinvertebrates are well suited for MMIs. West Virginia has used a statewide family-level stream condition index (WVSCI) since 2002. We describe the development, validation, and application of a geographically- and seasonally partitioned genus-level index of most probable stream status (GLIMPSS) for West Virginia wadeable streams. Natural classification strata were evaluated with reference site communities using mean similarity analysis and non-metric multidimensional scaling ordination. Forty-one metrics spanning six ecological categories (richness, composition, tolerance, dominance, trophic groups, and habits) were evaluated for sensitivity, responsiveness, redundancy, range and variability across seasonal (spring and summer) and regional (mountains and plateau) strata. Through a step-wise metric selection process, 8–10 metrics were aggregated to comprise four stratum-specific GLIMPSS models (mountain/plateau and spring/summer). A comparison of GLIMPSS with WVSCI exhibited marked improvements where GLIMPSS detecting greater stream impacts. A variation of the GLIMPSS, which differs only in the family-level taxonomic identification of Chironomidae (GLIMPSS (CF)), was comparable to the full GLIMPSS. These MMIs are robust yet practical tools for evaluating impacts to water quality, instream and riparian habitat, and aquatic wildlife in wadeable riffle-run streams based on sensitivity, responsiveness, precision, and independent validation. These models may be used effectively to detect degradation of the naturally occurring benthic community, assess causes of biological degradation, and plan and evaluate remediation of damaged stream ecosystems.

Keywords Macroinvertebrates · Bioassessment · Multimetric index · Reference condition · Taxonomic resolution · IBI

G. J. Pond (✉)
Office of Monitoring and Assessment, Environmental Assessment and Innovation Division, U.S. Environmental Protection Agency, Region III,
1060 Chapline St,
Wheeling, WV 26003, USA
e-mail: pond.greg@epa.gov

J. E. Bailey · B. M. Lowman · M. J. Whitman
Division of Water and Waste Management, Watershed Branch, West Virginia Department of Environmental Protection,
601 57th Street S.E.,
Charleston, WV 25304-2345, USA
e-mail: Jeffrey.E.Bailey@wv.gov

e-mail: Ben.M.Lowman@wv.gov

e-mail: Michael.J.Whitman@wv.gov

Introduction

Standardized freshwater bioassessments are used worldwide to ascertain the relative health and quality

of aquatic resources. Community structure-based assessment models such as multimetric indices (MMIs) are typically calibrated within states or regions, but some models have been developed nationally (e.g., Stoddard et al. 2008) and continentally (e.g., Hering et al. 2006). For decades, benthic macroinvertebrates have served as excellent indicators of biological health in flowing waters (Resh and Jackson 1993); they are the most widely studied group and are typically considered one of the most sensitive to anthropogenic disturbance. Macroinvertebrates can quickly recolonize habitats under improving chemical or physical conditions and are well-suited for use with MMIs. In the USA, the Federal Water Pollution Control Act, also known as the Clean Water Act (CWA) of 1972, is a comprehensive statute with the stated Congressional objective to "restore and maintain the chemical, physical, and biological integrity of the Nation's waters." Many federal, state, and tribal entities in the USA use multimetric indices to assess this biological integrity for CWA programs. Since 2002, WVDEP has used a statewide macroinvertebrate multimetric index (developed using family-level macroinvertebrate data) for biological assessments called the West Virginia Stream Condition Index (WVSCI) (Gerritsen et al. 2000). The index uses six metrics (total no. of families, no. of Ephemeroptera+Plecoptera+Trichoptera (EPT) families, family-level Hilsenhoff Biotic Index (HBI), % 2 dominant families, and % Chironomidae individuals). WVDEP has used the WVSCI to characterize patterns of stream condition and to measure biological attainment of the aquatic life use (ALU). In the development of the WVSCI, Gerritsen et al. (2000) found no distinct natural classification patterns at the family-level (based on an original set of 109 reference sites), and thus the WVSCI is not tailored to geographic or seasonal variation and has been used statewide over one broad index period (April–Oct). However, Gerritsen et al. (2000) recommended a re-evaluation of classification and metric suitability when more detailed genus-level data became available.

While the scientific debate over the cost-effectiveness of finer taxonomic resolution continues (Jones 2008), it is widely accepted that genus or species-level data more accurately represent the composition of the aquatic community and increases our ability to detect a variety of impacts. In the USA, Carter and Resh (2001) found that most government agencies favored genus- and species-level taxonomy. Some research

acknowledges that coarser family-level data can detect severe and obvious degradation in streams (Lenat and Resh 2001; Bailey et al. 2001), but genus- and species-level information can detect more subtle and complex effects and impairments as well (Lenat and Resh 2001; Schmidt-Kloiber and Nijboer 2004; Arscott et al. 2006). Several states adjacent to WV currently use genus-level MMIs (KY, OH, MD, and PA). Internationally, genus- or species-level taxonomy is widely used in MMIs (e.g., Vlek et al. 2004; Baptista et al. 2007; Verandes and Cortes 2010) but family-level taxonomy is frequently used for multivariate predictive models (e.g., Reynoldson et al. 2001). In southern WV, Pond et al. (2008) compared family- and genus-level metrics and taxonomy in assessing the effects of surface coal mining and reported that genus-level metrics detected impacts more effectively than WVSCI and its component metrics.

Our objectives were to develop and validate a genus-level index of most probable stream status (GLIMPSS) for bioassessment of West Virginia streams. This paper describes the refinement of WV's assessment tools, the goal of which was to create robust and easy-to-apply MMIs that account for natural variation and can accurately distinguish between reference and environmentally stressed benthic communities within WV streams.

Methods

Environmental setting and data collection

Woods et al. (1996) described the WV landscape as highly forested and dissected by four level III ecoregions: Blue Ridge (ecoregion 66), Ridge and Valley (ecoregion 67), central Appalachian Mountains (ecoregion 69), and Allegheny Plateau (ecoregion 70). WVDEP estimates that over 88,000 km of streams (1:24,000 scale National Hydrography Dataset) occur within the state. Because of the high relief and topographic dissection, much of this stream length is comprised of 1st- and 2nd-order headwater streams (1:24,000 scale). Three quarters of WV is composed of sedimentary sandstones and shales (ecoregions 69 and 70); much of the remainder higher relief is folded sedimentary and metamorphic rocks in the eastern part (ecoregion 67) and a very small portion of the state (<2 %) in ecoregion 66. WV contains diverse temperate

mesophytic and hardwood forests where the mineral resource extraction (e.g., coal, oil, and gas), silviculture, agriculture, and urban development are some of the most common sources of impacts to aquatic resources.

All data collection followed WVDEP standard operation procedures (WVDEP 2011). Briefly, biologists collected macroinvertebrates from riffle/run habitats within a 100-m reach using a 0.5-m-wide rectangular frame kick net (500 μm mesh). A composite of four 0.25 m^2 kick samples represented approximately 1 m^2 of stream bottom substrate. Taxonomists identified organisms to genus or lowest level practical (e.g., Nematoda) from sorted random 200-organism (±20 %) subsamples. Taxa were coded for various traits (pollution tolerance value, functional feeding group, habit) based on published literature (e.g., Lenat 1993; Barbour et al. 1999; Merritt et al. 2008) and further adapted from neighboring state lists. Within the 100-m reach, field crews measured in situ water quality (dissolved oxygen, pH, specific conductance, and temperature) and assessed habitat using US EPA Rapid Bioassessment Protocol (RBP) Habitat Assessment procedures following Barbour et al. (1999). This latter procedure qualitatively evaluates habitat components such as epifaunal substrate quantity and quality, degree of embeddedness, sediment deposition, channel alteration, stream bank stability, bank vegetative protection, and riparian zone width. Water chemistry data availability varied depending on WVDEP's programmatic survey types (e.g., probabilistic, long-term trend monitoring, targeted total maximum daily load monitoring). Metals were collected at 69 % of all sites (concurrent with benthic sampling) while nutrients were collected at 54 % of all sites. In contrast, fecal coliform bacteria samples were analyzed at 90 % of all sites. A full suite of all water chemistry parameters (e.g., nutrients, metals, bacteria, ions) were collected at approximately 50 % of all sites at the time of benthic sampling.

Dataset

The genus-level dataset (*n*=3,737) represented 2,354 unique streams and 3,411 unique stations sampled from March–October, 1999–2009. Those streams that had more than one station located on them averaged a distance of 3.9 km between stations. Overall, the dataset consisted of approximately 33.3 % probabilistically selected sites and 66.7 % target sites. We removed samples

containing <100 individuals from the development dataset. While most of these low abundance sites indicate severe chemical or habitat impairment, they might also include samples influenced by drought or spate conditions. As standard protocol, WVDEP reviews all available information to help evaluate whether samples with <100 individuals are related to natural or anthropogenic factors prior to assessing the site; conservatively, we excluded these low abundance sites for MMI development. Therefore, benthic samples used in this analysis ranged from 100 to 240 individuals. We also omitted larger rivers (>150 km^2) and true limestone streams (which make up <1 % of stream kilometers in the state). We excluded same day duplicate samples, and any additional samples collected from the same site within a 5-year period from the development dataset as a means to reduce bias and confounding. To estimate within-site variability due to spatial patchiness, annual changes and method error, we analyzed duplicates and annual revisits separately. We modeled a stressor gradient using only those samples with a full suite of water chemistry and habitat data (1,617 of the 3,737, 70.3 % of which were probabilistically selected). Prior to any analyses, the dataset was randomly divided into calibration (CAL; 70 % of all sites) and validation sets (VAL; remaining 30 % of all sites).

The development of a biological assessment MMI uses a reference condition approach (Hughes 1995, US EPA 1996; Barbour et al. 1999; Stoddard et al. 2006). The sensitivity of indicator metrics was evaluated by comparing the range of values among all reference sites in a class to sites known to be stressed by chemical or physical factors. The WVDEP relies on a combination of quantitative and qualitative physical and chemical attributes and narrative criteria to identify reference quality streams (Table 1). Additionally, we selected candidate reference sites by examining historic data (if available) and by consulting with regional professionals that have knowledge of local streams. To be classified as reference, a site must meet all of the listed conditions in Table 1. However, in areas where high quality reference sites are scarce, a site could be listed as reference even if it failed one of the criteria but best professional judgment provided evidence that the site could be included without degrading the reference condition classification. Thus, we would characterize most WV reference sites as "minimally disturbed" but retained sites considered to be "least disturbed" (after Stoddard et al. 2006).

Following EPA guidance (US EPA 1996; Barbour et al. 1999) the data were first divided into reference (REF), stressed (STRESS), and other, or non-reference (non-REF) sites. We defined STRESS sites as meeting at least one of the abiotic criteria (physical and chemical) shown in Table 1. These criteria are similar to the original WVSCI stress site criteria (Gerritsen et al. 2000) and cover a broad range of potential stressor variables across WV. Non-REF sites included all other sites that were not classified as either REF or STRESS and were used in combination with REF and STRESS sites for purposes of deriving disturbance gradients and metric correlations, metric standard values, and in precision analyses. Overall (combined CAL and VAL), the data represented 391 REF, 2,384 non-REF, and 962 STRESS sample sites (see Table 1). Figure 1 depicts the distribution of REF and STRESS sites among WV ecoregions and major catchments.

Data analysis

Community classification

Macroinvertebrate assemblages were grouped into natural classes based on inferences from community similarity analysis and ordinations (Barbour et al. 1999). We evaluated several combinations of seasonal and geographic factors to explain natural variability in macroinvertebrates found at REF sites. Seasonal categories (e.g., spring and summer) were biologically-relevant (i.e., they relate to known life histories of native taxa; e.g., Sweeney 1984). Geographical stratification followed Level III ecoregions after Woods et al. (1996). We evaluated the effect of stream size (e.g., mean stream width) on community structure and final GLIMPSS scores with correlation analysis. Classification schemes were evaluated with mean similarity analysis (MEANSIM) (Van Sickle 1997) and non-metric multidimensional scaling (NMDS) (McCune and Grace 2002). The following combinations of strata were evaluated: Level III ecoregion (67, 69, and 70); season (spring (March–through May), summer (June–early October)); bioregion (combined level III mountain ecoregions (67/69), plateau (70)); level III ecoregion × season; bioregion × season.

We calculated classification strength using MEAN-SIM with a Bray–Curtis similarity matrix of reference site communities (Hawkins and Norris 2000) by comparing the average within-class similarity (W) to the average between-class similarity (B) among strata combinations. The classification strength (CS) is simply W minus B. NMDS of REF sites was performed

Table 1 Screening criteria used to select reference and stressed sites for metric testing

	Reference	Non-reference	Stressed
Dissolved oxygen (mg/l)	>5	Not meeting either sets of criteria	<4
pH	>6 and <9		<4 or >9
Specific Conductance (μS/cm)	<500		>1,000
Fecal Coliform (col/100 ml)	<800		>4,000
Epifaunal substrate score	≥11		≤6 (and <120)
Channel alteration score	≥11		≤6 (and <120)
Sediment deposition score	≥11		≤6 (and <120)
Bank vegetation (each bank) score	≥6		≤6 (and <120)
Riparian zone (each bank) score	≥6		≤3 (and <120)
Total habitat score	≥130		
Point sources upstream	0		
Obvious non-point sources	0		
Violations of WQS	0		
No. of calibration sites	273	1,669	674
No. of validation sites	118	715	288
Total sites (n=3,737)	391	2,384	962

Reference sites must meet all criteria and stressed sites must meet only one criterion; however, stressed habitat criteria must meet a combination of both triggers of individual habitat metric and a <120 total habitat score. Number of calibration and validation sites per category is also provided
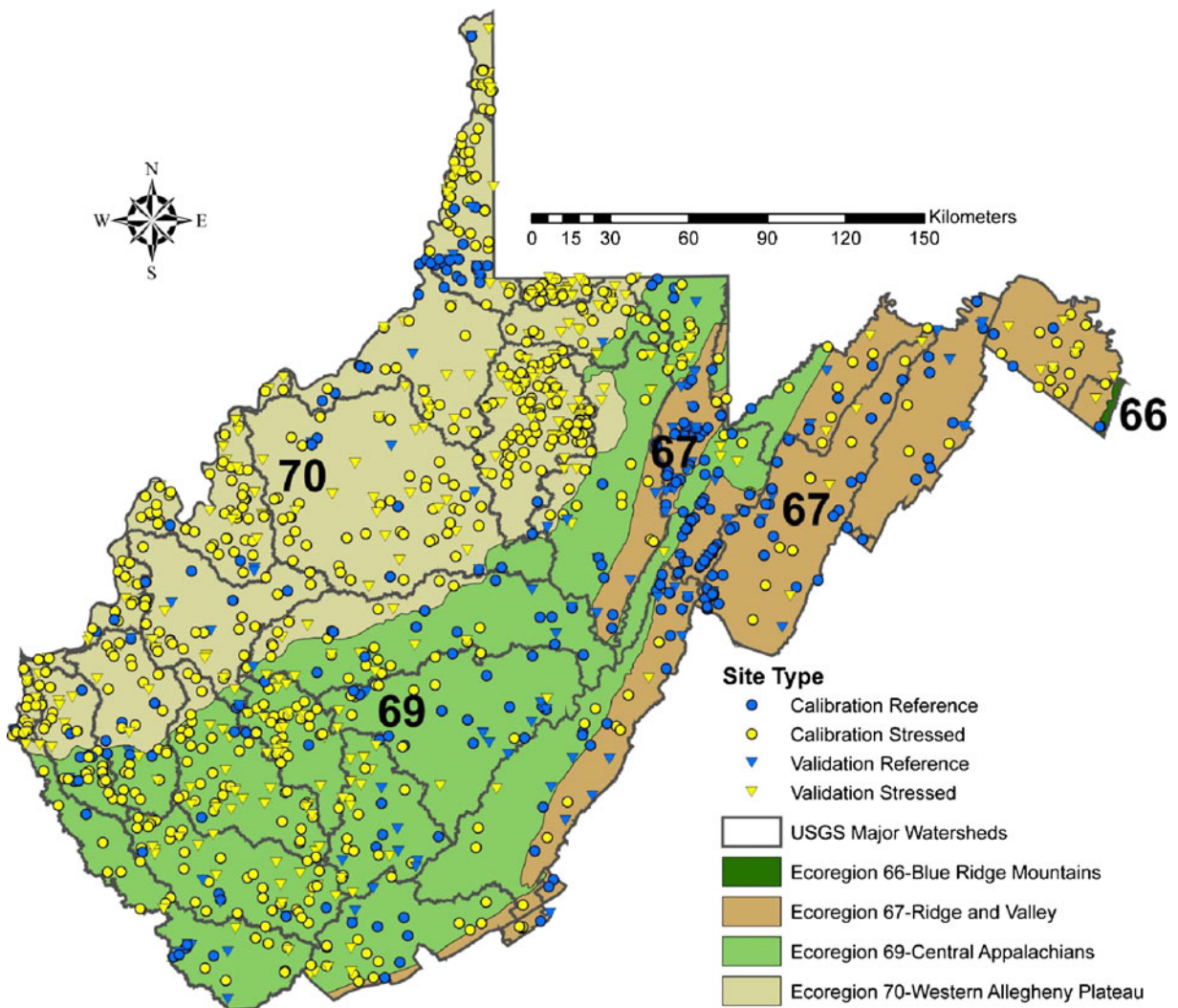
Fig. 1 Location of WV sites distributed among major watersheds and ecoregions for calibration and validation reference and stressed sites used for index development. *Numbers* represent level III ecoregion codes

with PC-ORD™ (v. 6, Gleneden Beach, OR) and although this ordination method is non-parametric, genus-level invertebrate abundances were log $(x+1)$ transformed to reduce any effect of skewed abundance distributions on the outcome of the ordination. NMDS was performed using the Bray–Curtis coefficient on six dimensions with 250 real runs and 250 randomized runs. Taxa observed at less than 2.5 % of all REF sites were omitted from the ordination (NMDS based on 158 taxa out of 322 total taxa). By excluding these infrequent taxa, multivariate analyses are more robust and patterns are more evident (McCune and Grace 2002). These omitted taxa were added back into the calculations for all metrics and the final GLIMPSS.

Metric selection and index construction

Forty-one biological metrics were calculated from queries programmed in the WVDEP database (see Appendix 1 for complete list). Ecological attributes included six categories of richness, composition, dominance, tolerance, trophic or functional feeding groups, and habit (Barbour et al. 1999). Taxon traits (e.g., pollution tolerance values, habit, and feeding) are coded in the WVDEP database and derived from published literature (e.g., Merritt et al. 2008) or other neighboring state attribute tables (e.g., PA and KY); ttrait metrics were based on WVDEP-specific designations. Evaluation of metrics followed published

techniques (e.g., Blocksom 2003; Hering et al. 2006; Whittier et al. 2007; Stoddard et al. 2008; Blocksom and Johnson 2009). After classification, we evaluated metrics within each stratum by running them through a step-wise performance process. This process included sensitivity (i.e., ability to distinguish between REF and STRESS), response to a human disturbance gradient, sufficient range and scope for detecting impairment, and lack of redundancy with other metrics. Representation of the six ecological categories was considered during the metric selection process.

First in the step-wise process, we evaluated metric sensitivity using discrimination efficiency (DE) (Gerritsen et al. 2000). Percent DE was calculated as the number of STRESS site metric values that fell below the reference set 25th centile (or >75th centile for negative response metrics) divided by the total number of STRESS sites, and multiplied by 100. We removed metrics that had DE values of less than 65 % unless the metric offered additional ecological information desired for the index (e.g., habit, trophic, and tolerance) and the metrics passed other selection criteria (see below). Those metrics that had the highest DE were considered for inclusion in the multimetric index.

Secondly, metrics significantly ($p < 0.01$) and moderately correlated ($|r| >= 0.25$) to the disturbance gradient were retained for further evaluation. We modeled a human disturbance gradient from linear combinations of abiotic factors at 1,617 sites using principal components analysis (PCA) (McCune and Grace 2002). We evaluated individual metric response along the disturbance gradient (PC1 scores) using Pearson correlation. The variables used in the PCA included same-time measures of: pH, and log transformations of temperature, D.O., fecal coliform, conductivity, sulfate, chloride, total phosphorus, nitrite–nitrate, total suspended solids, total aluminum, total iron, total manganese, and 7 of the 10 EPA RBP habitat metrics (channel flow status, velocity regime, and frequency of riffles were excluded).

Thirdly, we evaluated redundancy between metric pairs (highly correlated r values of >0.75) using Pearson correlation and further compared pairs using scatterplots. For metrics that were >0.75, both metrics were judged as adding information and were retained for further analysis if nonlinear relationships were apparent or if there was sufficient dispersion (i.e., scatter) of the paired metric data (Barbour et al. 1999).

The fourth step removed metrics with limited range. This involved richness metrics consisting of <5 taxa and abundance metrics with a range of <10 % (Klemm et al. 2003; Blocksom and Johnson 2009). The final step in metric selection evaluated a form of variability, referred to as "scope of impairment" (SOI) (modified after Klemm et al. 2003; Blocksom and Johnson 2009). We examined boxplots of the REF site metrics within individual strata. Metrics were rejected if the interquartile range was greater than the range between zero and the lower quartile, that is, an interquartile coefficient of >1.

Metric scoring and aggregation into an index

The set of most sensitive and responsive, and least redundant and variable metrics were aggregated for each stratum so that indicators could best contribute to the final GLIMPSS. These metrics included "positive" (i.e., increase with improving water or habitat quality) and "negative" (decrease with improving quality) scoring metrics. For scoring purposes, metrics were normalized by first calculating the 5th and 95th centiles of each metric based on all sites within each stratum. Using these upper and lower centiles (ceilings and floors), we scored metrics following Blocksom (2003), Hering et al. (2006), Whittier et al. (2007) Blocksom and Johnson (2009). Normalization (scoring) of a positive-direction metric was based on the following equation: (metric value−floor)/(ceiling−floor)*100; negative-direction metrics were scored as: (ceiling−metric value)/(ceiling−floor)*100. This resulted in dimensionless and equally-weighted metric scores on a scale of 0−100 %. If a positive-responding metric scored over 100 (e.g., a value above the 95th centile), then it was corrected to the maximum score of 100. Alternatively, if metric values fell outside the 95th or 5th centile (depending on metric direction), it received a score of zero. All metric scores were then averaged to produce the final index score (after Gerritsen et al. 2000). Four interpretive condition categories (very good, good, degraded, and severely degraded) were established based on REF distributions. GLIMPSS impairment thresholds were based upon REF 5th centiles within each stratum (designating "good" from "degraded"). Very good conditions were met when scores were more than the 25th centile of REF scores; scores that were <50 % of each impairment threshold represented severely degraded conditions.

Index performance: sensitivity and responsiveness

GLIMPSS impairment thresholds were based upon the 5th centile of the reference distribution within each stratum. The CAL GLIMPSS was evaluated for DE and response (i.e., correlation) to the human disturbance gradient (PC1 from the 1,617-site symmetric dataset) in each stratum. We then used the independent VAL dataset to calculate the classification efficiency (CE), or ability of GLIMPSS to correctly assign sites to either reference or stress categories (Southerland et al. 2007). This is different from DE in that CE was calculated as the sum of the number of VAL REF sites scoring above the 5th centile, plus the number of VAL STRESS sites scoring below the 5th centile of the development REF distribution, divided by the total number of sites. We also compared VAL sites to the PC1 disturbance gradient with correlation analysis.

Reproducibility

An estimate of GLIMPSS measurement error and associated precision (a type of performance measure) from same-day duplicates collected at 90 sites was estimated within individual strata according to Stribling et al. (2008). Firstly, the GLIMPSS mean square error (MSE or variance) was calculated; the square root of the MSE (the RMSE) provides the standard deviation, which is an estimate of measurement error. We used the standard deviation to calculate 90 % confidence intervals (90 % CI) around a single observation using the standardized z-score. The coefficient of variation (%CV) and relative percent difference (RPD) followed standard calculations. Estimates of precision are essential to the interpretation of the index because it helps identify variability from sampling method and laboratory processing and can be used to compare precision among field teams or between entities. Estimates of measurement error also provide a means to compare GLIMPSS scores between single site observations. To estimate natural temporal and method variability, we repeated this analysis with GLIMPSS scores for 30 REF sites revisited yearly within the same index period to evaluate natural annual variability at REF sites. These estimates can be used to detect long-term trends in GLIMPSS scores at individual sites through time.

Relationship between WVSCI and GLIMPSS

In order to compare the GLIMPSS with the WVSCI, the WVSCI impairment threshold (5th centile) was first updated with all current REF sites used in GLIMPSS (391 vs. original 109 used by Gerritsen et al. 2000); this increased the original threshold from 68.0 to 71.6. We then calculated and compared each index impairment rate (i.e., number of sites impaired based on all sites and within individual strata).

GLIMPSS with family-level Chironomidae

Within each stratum, we developed a second index differing only in the use of family-level vs. genus-level taxonomic identification of a group of dipterans known as Chironomidae (non-biting midges), termed GLIMPSS (CF). Although it is well known that some chironomid genera are responsive and can indicate certain stressors (Rosenberg 1992), chironomid taxonomy increases sample processing time and requires additional expertise that may be lacking in some laboratories. We used the identical datasets to calibrate and validate the GLIMPSS (i.e., CAL and VAL; REF, non-REF, and STRESS) to re-analyze GLIMPSS (CF) metrics in which all chironomid genera in the benthic enumeration table were first converted (collapsed) to family level. Stratum-specific GLIMPSS metrics that involve chironomid genera include: no. of total genera, no. of intolerant genera (pollution tolerance value (PTV)<4), no. of intolerant genera (PTV<3) no. of clinger genera, HBI, % tolerant (PTV>6), % 5 dominant genera, no. of scraper genera, and no. of shredder genera. Second, these metrics were then recalculated with chironomids collapsed to the family level. The % Orthocladiinae metric, a chironomid sub-family metric, was automatically omitted from the GLIMPSS (CF). In the case of % Orthocladiinae, two highly comparable analogs (% Chironomidae and % Chironomidae+Annelida) were tested as a potential replacement metric. All other metrics that comprise the GLIMPSS were retained and unmodified (e.g., no. of Ephemeroptera genera, % EPT (minus *Cheumatopsyche*)).

Metric re-testing and confirmation involved checking the DE, PCA correlation, redundancy magnitude, range, and SOI of the newly modified GLIMPSS metrics from the CAL dataset. The upper (95th) and lower (5th) centiles for metrics were recalculated from the full dataset and GLIMPSS scores were calculated

for all CAL and VAL sites. As in the development of the full GLIMPSS, the GLIMPSS (CF) was similarly evaluated for sensitivity and responsiveness. Finally, we compared the full GLIMPSS and the modified GLIMPSS (CF) models using scatterplots and Pearson correlation coefficients.

Results

Community classification

MEANSIM showed that bioregion–season (CS= 18 %) and ecoregion–season (CS=17 %) had the highest classification strengths at REF sites (Table 2). Ecoregion and bioregion alone produced the lowest CSs (10 and 12 %, respectively) indicating the stronger influence that seasonality had on structuring REF benthic communities in these regions. In the NMDS analysis, a three-dimensional solution was fitted after 110 iterations. Axes (i.e., dimensions) 1 and 2 were 100 % orthogonal and explained the most variance in the distance between the original distance matrix and final configuration (28 and 23 %, respectively) and stress was relatively low (18.9 %). Axis 3 explained only 11 % variance and was not plotted. Visual inspection of the NMDS ordinations (Fig. 2 a–d) showed comparable results to MEANSIM analysis (i.e., classification strength can be visualized in the ordinations). The best ordination was bioregion-season (Fig. 2d) which had the least scatter across strata and good clustering within strata. Although ecoregion-season had high CS, there was considerable overlap of ecoregion 67 and 69 sites in both seasons (Fig. 2c).

Some environmental factors were more highly correlated with NMDS axis scores (e.g., elevation and Julian day) than others (e.g., longitude and latitude). Intuitively, bioregion-season can be largely explained by elevation (i.e., mountain areas of ecoregion 67–69 vs. lower elevations in 70) and Julian day (spring vs. summer index period). The strong significant influence of Julian day and temperature on axis 1 (r=0.70 and 0.72, respectively) confirms that benthic communities changed substantially along a seasonal (spring to summer) continuum. Stream width, longitude, latitude, elevation, pH, or specific conductance were not significantly correlated (p>0.05) to axis 1 scores (all $|r|$<0.28). Only elevation was significantly correlated with axis 2 scores (r=0.49, p<0.05), as characterized by the separation of mountain and plateau regions along axis 2.

The ordination results indicated that splitting the data into common seasons (i.e., spring and summer) in addition to bioregions provided good separation. Thus, we chose four classification strata for index development purposes: Mountain Spring (MT Sp; CAL n=86 sites), Mountain Summer (MT Su; CAL n=126 sites), Plateau Spring (PL Sp; CAL n=31), and Plateau Summer (PL Su; CAL n=28 sites).

The stressor gradient

A synthetic human disturbance gradient was modeled with PCA. Table 3 lists the eigenvalues, percent variance explained by each axis, and factor coefficients of each environmental variable on the respective axes. PC axis 1 explained roughly 26 % of the variance in the dataset (eigenvalue=5.1) and was judged suitable for use as a human disturbance gradient.

Table 2 Similarity analysis (MEANSIM) results for strata combinations using Bray–Curtis coefficients from REF calibration sites based on log-transformed (x+1) abundance data showing within (W) and between (B) similarity

	n groups	W	B	$W-B$ (CS; %)	p
Ecoregion	3	0.56	0.46	10	<0.0001
Season	2	0.57	0.42	15	<0.0001
Ecoregion–Season	6	0.64	0.47	17	<0.0001
Bioregion	2	0.56	0.44	12	<0.0001
Bioregion–Season	4	0.62	0.44	18	<0.0001

Significance tests based on 1,000 permutations

CS classification strength (%)
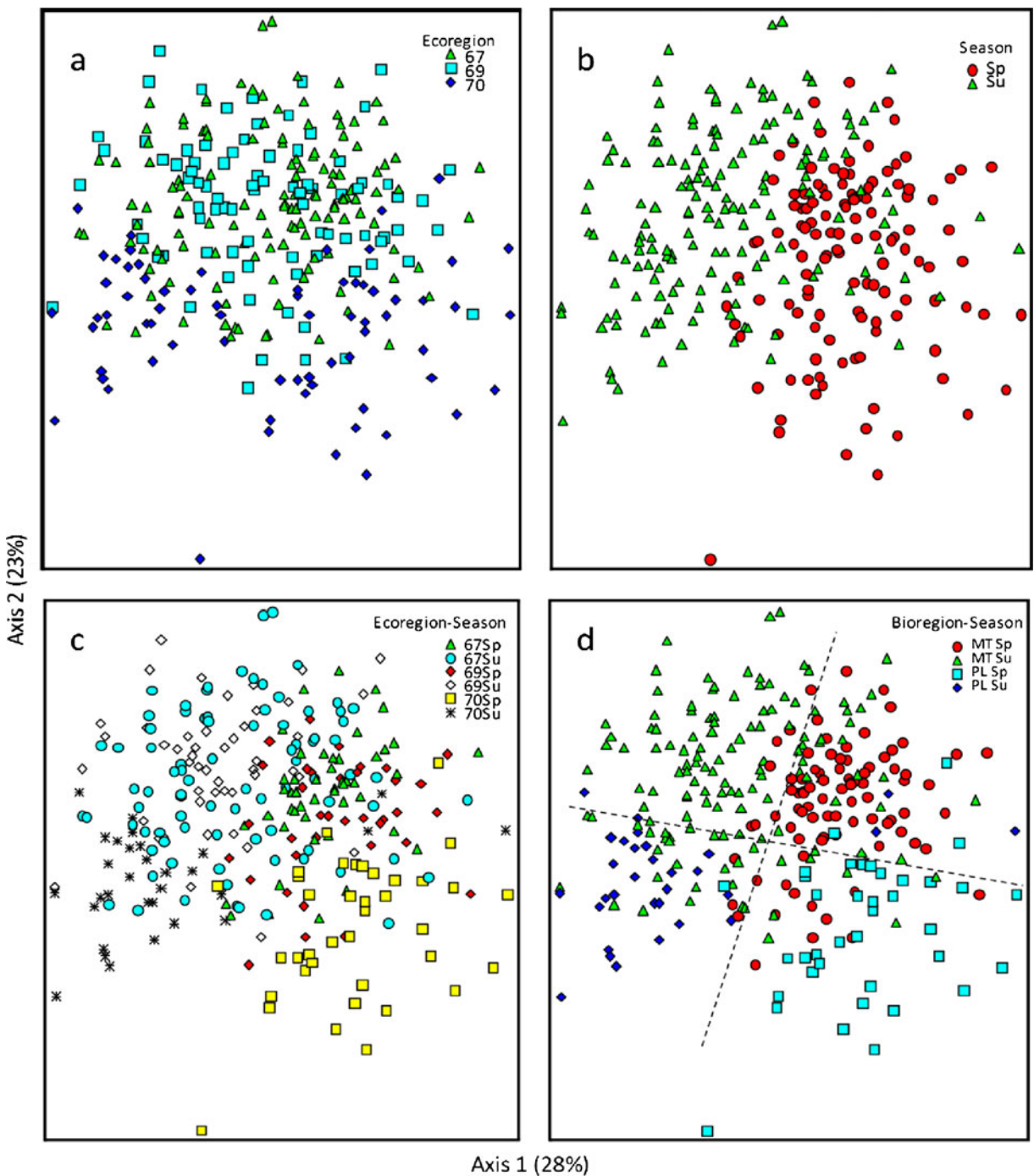
Fig. 2 NMDS ordination of CAL reference assemblages for ecoregion (**a**), season (**b**), ecoregion–season (**c**), and bioregion (**d**). *Data in parentheses* report the percent variance of axis contribution to the ordination. Bioregions include mountains (combined ecoregions 66, 67, and 69) and plateau (ecoregion 70). *MT Sp* Mountain Spring, *MT Su* Mountain Summer, *PL Sp* Plateau Spring, *PL Su* Plateau Summer. *Dashed lines* in (**d**) drawn by eye for emphasis

Specific conductance had the highest correlation of chemical variables to axis 1 (+0.47) followed closely by temperature (+0.46). But daily temperatures fluctuate widely; the correlation using instantaneous readings

Table 3 PCA results (eigenvalues and percent variance accounted for) and factor correlation coefficients of the first three components based on 1,617 sites

	Axis 1	Axis 2	Axis 3
Eigenvalues	5.1	2.6	2.1
Percentage	25.9	13.0	10.2
Cumulative percentage	25.9	38.5	48.7
Temperature	0.46	0.31	−0.26
pH	0.42	0.66	−0.14
Dissolved oxygen	−0.12	−0.01	0.25
Specific conductance	0.47	0.12	−0.38
Fecal coliform	0.19	0.07	−0.09
Sulfate	0.36	0.01	−0.33
Chloride	0.28	0.10	−0.21
Total suspended solids	0.23	−0.25	−0.17
P total	0.06	−0.10	−0.22
NO_2–NO_3	0.23	0.11	−0.34
Aluminum (total)	0.20	−0.64	−0.22
Iron (total)	0.20	−0.53	−0.18
Manganese (total)	0.25	−0.59	−0.27
Epifaunal substrate score	−0.63	0.15	−0.37
Embeddedness score	−0.62	0.27	−0.26
Channel alteration score	−0.58	−0.16	−0.30
Sediment deposition score	−0.60	0.23	−0.30
Bank stability score	−0.58	0.00	−0.42
Bank vegetation score	−0.69	−0.12	−0.45
Riparian zone score	−0.68	−0.19	−0.34

might simply relate to regional and seasonal patterns which were likely controlled by our selected stratification. Habitat metrics were also strongly correlated to axis 1 (+0.58–0.69). Measures of pH and total metals were mostly correlated with axis 2 (>+0.50), but this axis only accounted for 13 % of the total variance (eigenvalue=2.6). Appendix 1 summarizes general stratum-specific abiotic data for REF and STRESS sites.

Metric evaluation and selection

The stepwise selection process yielded GLIMPSS models that incorporate ten metrics in both MT strata (spring and summer), 8 metrics in the PL Spring, and 9 metrics in PL Summer stratum. We selected one or more metrics from each of the six ecological categories in all strata, except PL Sp where no FFG or dominance metric met the metric test criteria. Table 4 summarizes metric performance (i.e., sensitivity, responsiveness, and redundancy) by bioregion-season

for the selected metric sets. See Appendix 2 for stratum-specific results from all metric testing.

While there was consistent performance observed for many metrics, other metrics showed wide ranges of sensitivity across strata, indicating differences in biological potential from seasonal and geographic factors or sensitivity to region-specific stressors. For example, the % 5 dominant taxa metric was very sensitive in the PL Su (DE=74.8) but was rejected from further consideration in PL Sp (DE=47.1). By comparison, %Orthocladiinae was very sensitive in the MT Sp (DE=87.3) but had much less discrimination ability in the PL Su (DE=60.2). Although no. of EPT genera had high DE in all strata, we chose to use independent measures of Ephemeroptera, Plecoptera, and Trichoptera genera where possible, in order to benefit from the known diagnostic capability of these individual metrics (Karr and Chu 1999). Moreover, the no. of Trichoptera genera metric was not sensitive in many strata (except MT Sp). While %EPT is a commonly

Table 4 List of selected GLIMPSS metrics by stratum with %DE, correlation to PCA axis 1 ($p<0.05$), and maximum correlation value for selected metric pairs as a measure of redundancy (Redun; absolute value)

Group		DE	PCA	Redun
MT Sp				
Tol	No. of intolerant taxa (PTV<4)	92	−0.61	0.74
Comp	% Orthocladiinae individuals	89	0.49	0.45
Tol	Modified HBI	84	0.62	0.45
Rich	No. of Trichoptera genera	81	−0.39	0.59
Rich	No. of Ephemeroptera genera	79	−0.36	0.45
Rich	No. of Plecoptera genera	78	−0.53	0.74
Habit	No. of Clinger genera	75	−0.47	0.66
Comp	% Ephemeroptera individuals	75	−0.31	0.34
Dom	% 5 dominant genera	73	0.35	0.64
FFG	No. of scraper genera	68	−0.42	0.62
PL Sp				
Tol	Modified HBI	87	0.38	0.57
Comp	% EPT (minus *Cheumatopsyche*)	83	−0.36	0.70
Tol	No. of intolerant taxa (PTV<4)	82	−0.49	0.67
Rich	No. Plecoptera genera	79	−0.36	0.65
Tol	% tolerant (PTV>6)	77	0.24	0.45
Comp	% Chironomidae+Annelida	75	0.31	0.70
Rich	No. of Ephemeroptera genera	65	−0.35	0.59
Habit	No. of clinger genera	61	−0.38	0.67
MT Su				
Tol	Modified HBI	96	0.66	0.74
Tol	No. of intolerant taxa (PTV<4)	95	−0.68	0.74
Rich	No. of Plecoptera genera	91	−0.62	0.74
Comp	% EPT (minus *Cheumatopsyche*)	90	−0.59	0.74
Habit	No. of clinger genera	79	−0.39	0.69
Rich	No. of Ephemeroptera genera	78	−0.43	0.39
Dom	% 5 dominant genera	72	0.26	0.67
FFG	No. shredder genera	71	−0.45	0.47
Rich	No. total genera	68	−0.31	0.69
Comp	% Orthocladiinae individuals	67	0.34	0.41
PL Su				
Tol	No. of intolerant taxa (PTV<3)	98	−0.53	0.70
Tol	Modified HBI	94	0.55	0.75
Comp	% EPT (minus *Cheumatopsyche*)	86	−0.54	0.75
Habit	No. of clinger genera	83	−0.54	0.64
Dom	% 5 dominant genera	75	0.32	0.62
Rich	No. of total genera	74	−0.37	0.64
Rich	No. of Ephemeroptera genera	69	−0.52	0.46
Comp	% Chironomidae	65	0.52	0.53
FFG	No. of scraper genera	63	−0.43	0.44

Metrics are sorted by DE (in decreasing order). Some metric names are abbreviated (see Appendix 2 for full names). Metric groups listed as richness (Rich), composition (Comp), tolerance (Tol), habit, dominance (Dom), and feeding group (FFG)

MT Sp Mountain Spring, *MT Su* Mountain Summer, *PL Sp* Plateau Spring, *PL Su* Plateau Summer

used metric in many MMIs, it was apparent that component orders were not always discriminatory (e.g., % Trichoptera and % Plecoptera components failed DE and SOI in many strata). However, excluding the

tolerant and abundant caddisfly *Cheumatopsyche* (after Pond et al. 2003) often increased %EPT performance. Like in the DE analysis, we observed differences in responsiveness (PC 1 correlation) between metrics across strata. The most notable of these was no. of scraper genera which was significantly related to increasing disturbance in the MT Sp ($r=-0.45$) but was much weaker in the MT Su ($r=-0.24$). This could be due to the fact that scraper richness naturally declines (i.e., reduced range) in summer when streams are more fully canopied, rather than depicting a loss of sensitivity to stress. Several metrics (e.g., HBI, no. of intolerant genera, and no. of clinger genera) consistently showed high correlation to PC 1 in all strata. Overall, there were few redundant pairs and most metrics appeared to offer somewhat different information as denoted by having correlations well below 0.75. Metrics that had high DE and stressor responsiveness rarely showed redundancy. For example, in the PL Su, no. of Plecoptera genera was highly redundant with no. of intolerant genera (PTV<3) (0.93); however, since no. of intolerant (PTV<3) had higher range and DE, it was selected as the preferred metric. For the selected metric sets, maximum redundancy magnitudes (within each stratum) are reported in Table 4. Appendix 2 tabulates those metrics that passed DE and PCA tests but failed the SOI test. For instance % scrapers had good DE (79 %) and moderate PC 1 correlation ($r=-0.34$) in the MT Sp, but had an unacceptable SOI of 1.42. In the MT Su stratum, % Ephemeroptera (minus *Baetis*) had a similarly good DE (77 %), but a SOI of 1.37. In these cases, we selected comparably sensitive and responsive metrics with acceptable SOIs. Out of all metrics, only % Annelida failed the range test. Within each stratum, we reported CAL metric upper and lower 5th centiles (Table 5) and scored metrics for each site in the CAL dataset.

Index performance

To examine the sensitivity of GLIMPSS, we plotted score distributions using box plots and then calculated DE within individual strata. GLIMPSS discriminated between REF and STRESS sites with a high degree of efficiency (>75 %) (Fig. 3). Non-REF sites fell into an intermediate position between REF and STRESS, with respect to score distribution. DE was greatest in the PL Su stratum where 89 % of the STRESS sites fell below the 5th centile of the REF distribution. The

independent VAL ($n=118$ REF sites, 288 STRESS sites across all strata) CE (i.e., % of correctly classified REF and STRESS sites) was excellent for all strata (Fig. 4), indicating successful validation (CE ranged from 89 to 95 %). GLIMPSS was also responsive to increasing stress (as indicated by PC 1); stratum-specific correlation of all CAL and VAL GLIMPSS scores were significantly correlated ($p<0.001$) to the disturbance gradient. Correlations with PC 1 ranged from −0.60 to −0.70 for both CAL and VAL MT Sp and MT Su sets; correlations in PL Su were −0.66 and −0.62 for CAL and VAL, respectively. PL Sp had the lowest correlation to disturbance at CAL and VAL sites (−0.48 and −0.42, respectively).

REF GLIMPSS relationships with natural gradients

Natural abiotic variable correlations (Julian day, latitude, longitude, stream width, and elevation) with REF site GLIMPSS were weak ($|r|<0.22$) and non-significant ($p>0.05$) indicating that seasonal and geographical stratification effectively accounted for most natural variation. However, scores in the PL Su were significantly negatively correlated (albeit only moderately) with stream width ($r=-0.44$) and conductivity ($r=-0.45$). Although specific conductance can increase with watershed disturbance, it might also represent a gradient of geologic or edaphic variability within this ecoregion.

Final metric scoring

Data from the CAL and VAL sets were combined and upper and lower 5th centiles were recalculated to produce final GLIMPSS scoring benchmarks for the four strata. Appendix 3 (Table 11) provides final GLIMPSS upper and lower 5th centiles for all metrics in each stratum. Overall, there were only minor adjustments made from the CAL values using the larger dataset. A few metrics showed substantial variation across seasons (e.g., % EPT (minus *Cheumatopsyche*) and no. of intolerant genera in PL Sp vs. PL Su) or bioregions (e.g., no. of intolerant genera (PTV<4) in MT Sp vs. PL Sp); however, other metrics were not considerably different across strata.

Reproducibility

Precision of GLIMPSS was satisfactory based on same-day replicate sampling (Table 6). PL Su had the highest variability of all strata with respect to %CV and RPD

Table 5 Calibration standard values (ceilings=upper 95th, and floors=lower 5th centiles) for metrics (by stratum) used for CAL and VAL scoring purposes

	Metric	MT Sp (n=488)		MT Su (n=1,071)		PL Sp (n=457)		PL Su (n=600)	
	Direction	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor
No. of total genera	+			38	14			34	14
No. of intolerant genera (PTV<4)	+	19	1	15	0	15	1		
No. of intolerant genera (PTV<3)	+							7	0
No. of Ephemeroptera genera	+	8	1	9	0	10	1	7	0
No.of Plecoptera genera	+	8	1	7	0	7	0		
No. of Trichoptera genera	+	7	1						
No. of clinger genera	+	20	4	19	5	17	3	15	4
Modified HBI (genus level)	−	6.14	2.26	6.27	2.89	6.54	2.52	6.24	3.73
% dominant 5 genera	−	91.8	48.4	91.7	51			91.5	53.3
% tolerant (PTV>6)	−					65	0.0		
% Ephemeroptera	+	58.1	0.5						
% EPT (minus *Cheumatopsyche*)	+			85.4	5.3	91.2	2.6	67.1	1.4
% Orthocladiinae	−	50.9	0.5	36.8	0.4				
% Chironomidae	−							69.1	4.0
% Chironomidae+Annelida	−					83.5	1.8		
No. of scraper genera	+	8	0					7	1
No. of shredder genera	+			5	0				

See Appendix 3 for final upper and lower metric calculation bounds

MT Sp Mountain Spring, *MT Su* Mountain Summer, *PL Sp* Plateau Spring, *PL Su* Plateau Summer, "+" positive responding metrics (to increasing water quality), "−" negative responding metrics (to increasing water quality)

(but these were highly dependent on the low population mean). Overall, estimates of measurement error (SD) were fairly stable across strata (Table 6). We also evaluated GLIMPSS scores at REF sites for precision, by comparing annual revisits to 30 sites within the same index period (Table 6). These sample events occurred between 1 and 8 years apart, and we assumed that REF site environmental condition was unchanged, so changes in the MMI should not result from anthropogenic disturbance. Differences in scores, however, do incorporate many factors: timing within an index period, preceding weather patterns, natural disturbance, and any unknown human impacts that may have occurred within the catchment between sample years. Despite these factors, SD, CV (%), and RPD were very good and considerably less than all same-day duplicates.

Comparison of WVSCI and GLIMPSS

The indices differed in assessing impairment 15 % of the time; the majority of agreements were at the high-and low-end scoring sites. Based on the full 3,737-site dataset, GLIMPSS rated 2,218 as impaired vs. 2,088 for WVSCI (130 less sites); however, there were stratum-specific differences in impairment rates. For example, in MT Sp (total n=697) and MT Su (total n=1530), GLIMPSS assessed 10 and 9 % more impaired sites than WVSCI, respectively. In contrast, the PL Sp (total n=653) GLIMPSS rated only 1 % more impaired sites, while in PL Su (total n=857), the family-level WVSCI rated 9 % more sites as impaired than the GLIMPSS. Using <50 % of threshold to demarcate the worst condition category, the GLIMPSS rated 626 sites (out of 3,737) as "severely degraded" while WVSCI rated 258.

Comparison of GLIMPSS vs. GLIMPSS (CF) performance

Using identical metric evaluation techniques, the process yielded sets of modified metrics that were as robust as in the full model. The final GLIMPSS (CF)
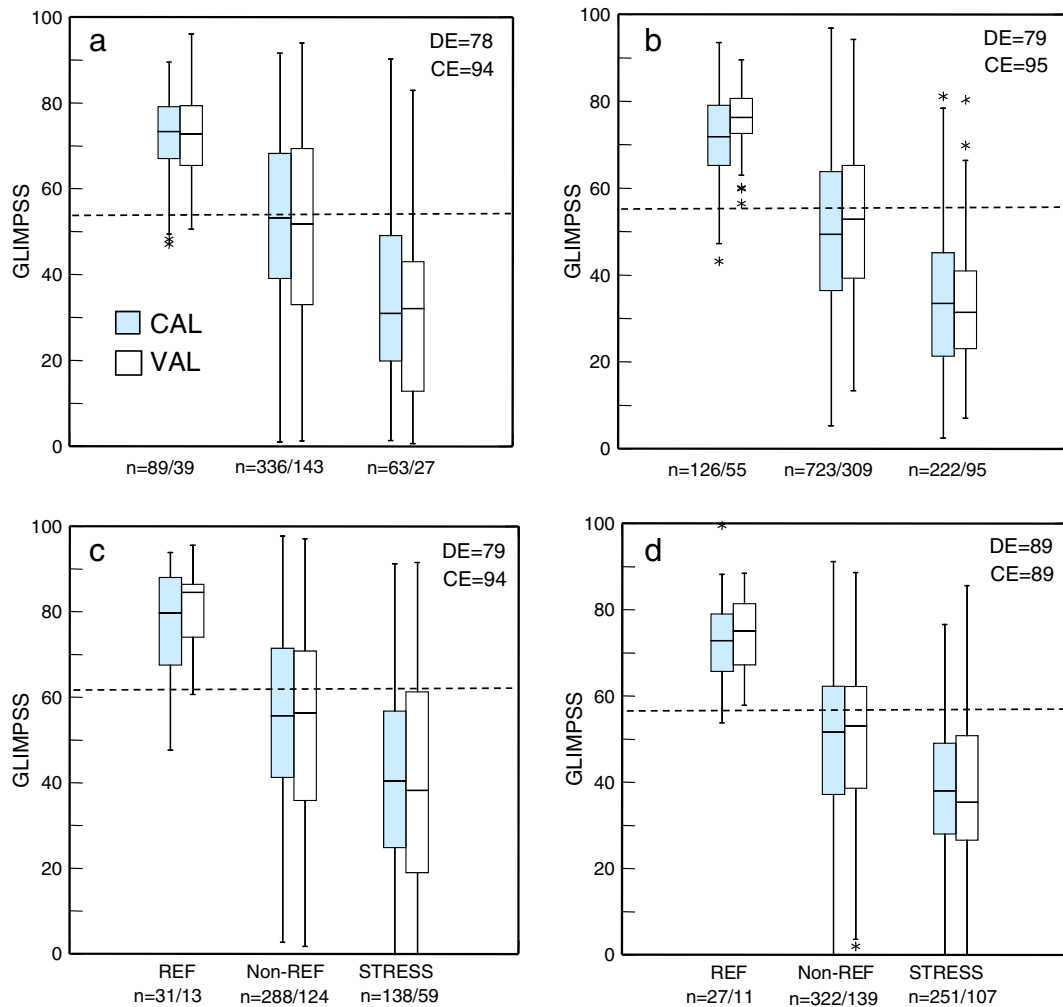
Fig. 3 Boxplots of calibration (*CAL*) and validation (*VAL*) GLIMPSS scores between REF, non-REF, and STRESS categories in MT Sp (**a**), MT Su (**b**), PL Sp (**c**), and PL Su (**d**). Number of sites in each stratum shown as *n*=CAL/VAL. Percent discrimination efficiency (*DE*) for CAL and percent classification efficiency (*CE*) for VAL also provided. *Dashed line* represents approximate 5th centile of REF, indicating the impairment threshold. See Fig. 2 and text for stratum definitions

incorporates 10 metrics in the MT Sp, 9 metrics in the MT Su, 7 metrics in the PL Sp, and 9 metrics in the PL Su (see Appendix 3, Table 12). Using mean DE, mean correlation with PCA, and mean redundancy magnitude across metrics within each stratum (Table 7), we conclude that on average, the GLIMPSS (CF) metrics performed similarly to, or slightly better than, the full GLIMPSS metric sets, indicating that the selected GLIMPSS (CF) metrics are highly suitable for assessment purposes. As with the full GLIMPSS, the 95th and 5th centile of the CAL and VAL datasets were recalculated and all metrics were scored for each site within individual strata. Final upper and lower 5th centiles for each GLIMPSS (CF) metric within strata are shown in Appendix 3 (Table 12). Similar to the full GLIMPSS, box plots (Fig. 4) indicate excellent sensitivity of GLIMPSS (CF). The simplified index could discriminate between REF and STRESS sites with a high degree of efficiency (≥75 % of STRESS sites score below 5th centile of REF distribution). DE was greatest in the PL Su stratum where ~90 % of the STRESS sites fell below the 5th centile of the REF distribution.

As expected, the two models were strongly correlated (Fig. 5). There was slightly more scatter in MT Su and
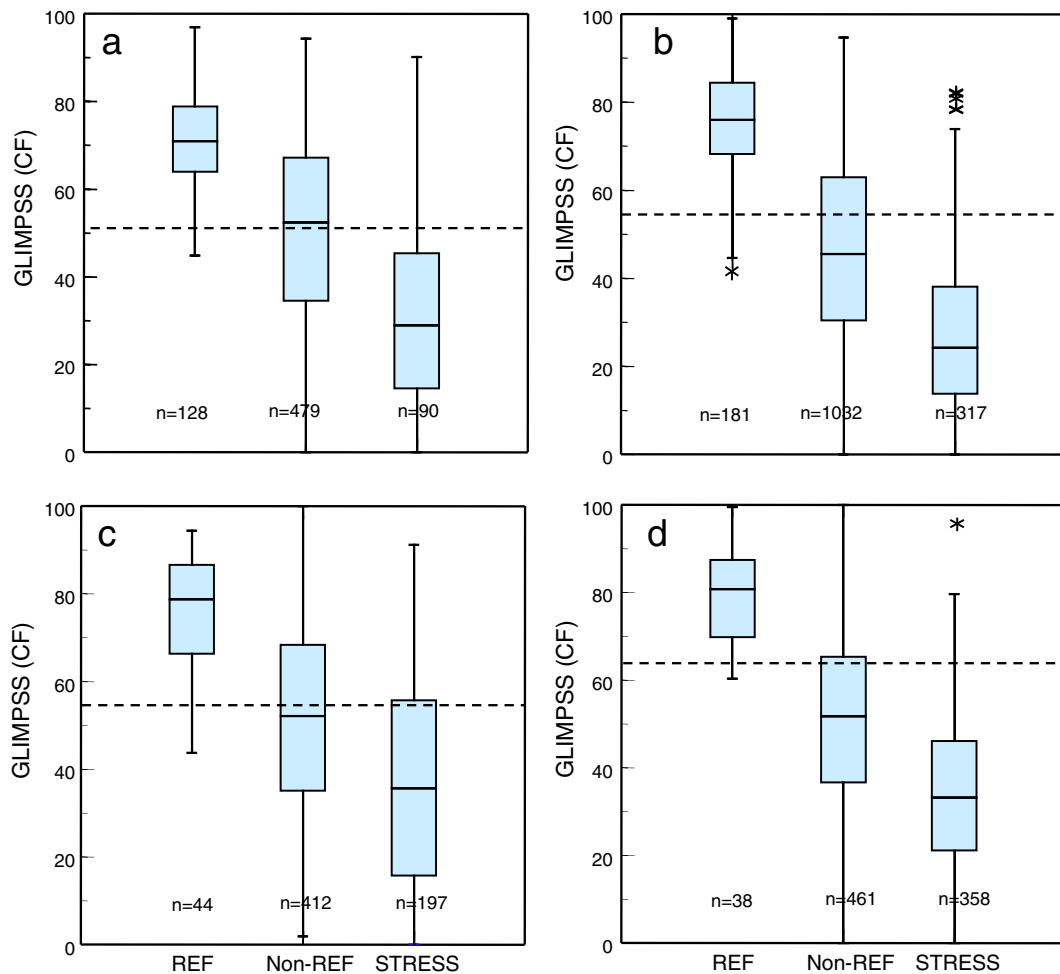
Fig. 4 Box plots of REF vs. non-REF and STRESS site classes using GLIMPSS (*CF*) scores from full dataset (total number of sites per stratum and class provided) among MT Sp (**a**), MT Su (**b**), PL Sp (**c**), and PL Su (**d**) strata. *Dashed line* represents approximate 5th centile of REF distribution (impairment threshold). See Fig. 2 and text for stratum definitions

PL Su compared with other strata. In each stratum, the full GLIMPSS tended to score higher than GLIMPSS (CF) as depicted by points lying above the 1:1 line. Adjustments to the impairment thresholds (based on the 5th centile of REF) only ranged from 2–4 points; however, in the lower left quadrants of the plots (impaired sites), larger differences were noted indicating that the GLIMPSS (CF) rated impaired sites as being in worse condition than with the full GLIMPSS.

Final benchmarks based on the reference distribution

We elected to retain the 5th centile of reference scores as an impairment threshold as with the original

WVSCI (after Gerritsen et al. 2000) for GLIMPSS and GLIMPSS (CF). Although other states often use the 10th or 25th centile of the reference distribution as impairment thresholds, our rationale to use the 5th centile is based on our confidence in the quality of the reference condition (see abiotic data in Appendix 1). Because of this high confidence in reference conditions, we used 25th centile of the reference distribution as a lower threshold to identify and protect high quality biological assemblages found in WV. Values below the impairment threshold were partitioned to provide categories that reflect increased stress to biological communities (e.g., degraded, severely degraded). Table 8 presents stratum-specific GLIMPSS

Table 6 GLIMPSS precision estimates and statistics from same-day duplicates (*n*=90) and annual revisits at REF sites (*n*=30 sites) among all strata and within individual strata

GLIMPSS (same-day duplicates)	MSE	SD	Population mean	One-tailed 90 % C.I.	CV (%)	RPD
All strata (*n*=90)	45.8	6.8	47.4	8.7	14.3	19.3
MT Sp (*n*=16)	38.9	6.2	54.4	8.0	11.5	16.1
MT Su (*n*=47)	42.4	6.5	46.8	8.3	13.9	17.1
PL Sp (*n*=10)	27.9	5.3	49.2	6.8	10.8	16.9
PL Su (*n*=17)	48.9	7.0	39.9	9.0	17.5	27.6
REF GLIMPSS (annual re-visits)						
All strata (*n*=30)	42.6	6.5	80.5	8.4	8.1	9.4
MT Sp (*n*=8)	27.8	5.3	76.6	6.8	6.9	8.0
MT Su (*n*=12)	43.9	6.6	82.9	8.5	8.0	9.2
PL Sp (*n*=3)	22.0	4.7	78.8	6.0	6.0	6.1
PL Su (*n*=7)	70.1	8.4	81.6	10.7	10.3	13.2

MT Sp Mountain Spring, *MT Su* Mountain Summer, *PL Sp* Plateau Spring, *PL Su* Plateau Summer

and GLIMPSS (CG) scores indicating the 5th and 25th centile values and equal bisection of the impairment range.

Discussion

Development of MMIs for WV wadeable streams followed common standardized techniques for selecting metrics (Barbour et al. 1999; Klemm et al. 2003; Stoddard et al. 2008) so that the final index was sensitive, responsive, and precise. Moreover, by selecting metrics from key ecological attribute categories (e.g., richness, composition, tolerance, dominance, trophic groups, and habit), a more comprehensive view of stream condition is assessed as the index encompasses and integrates multiple features of a stream's biological system (Barbour et al 1999). Many state water agencies

have incorporated MMIs into their routine CWA monitoring and assessment programs; most having calibrated their own MMI using state-specific data. Regionally-based MMIs also exist (e.g., Klemm et al. 2003; Stoddard et al. 2005) and large-scale, nationally based bioassessment models have been developed by US EPA (e.g., Stoddard et al. 2008); European countries and the European Union also assess streams with national-based indices. Interestingly, variations and combinations of many metrics selected for state or regional MMIs reveal that similar macroinvertebrate groups (e.g., Ephemeroptera richness and chironomid abundance) and traits (e.g., intolerants, scrapers, and shredders) are frequently shared.

The GLIMPSS and GLIMPSS (CF) are robust yet practical tools for evaluating impacts to water quality, instream and riparian habitat, and aquatic wildlife in wadeable riffle-run streams based on sensitivity,

Table 7 Comparison of metric performance values for calibration data: discrimination efficiency (%DE), responsiveness (*r* value for PCA), and redundancy magnitude (*r* value for selected metric pairs) between stratum-specific GLIMPSS metrics and GLIMPSS (CF) metrics

	DE	DE (CF)	PCA	PCA (CF)	Redun	Redun (CF)
MT Sp	79.4	79.7	0.45	0.47	0.57	0.54
MT Su	80.7	85.9	0.47	0.53	0.63	0.68
PL Sp	75.3	75.4	0.37	0.39	0.63	0.65
PL Su	78.5	81.1	0.49	0.54	0.61	0.67

Values are means across all metric sets within each stratum

MT Sp Mountain Spring, *MT Su* Mountain Summer, *PL Sp* Plateau Spring, *PL Su* Plateau Summer
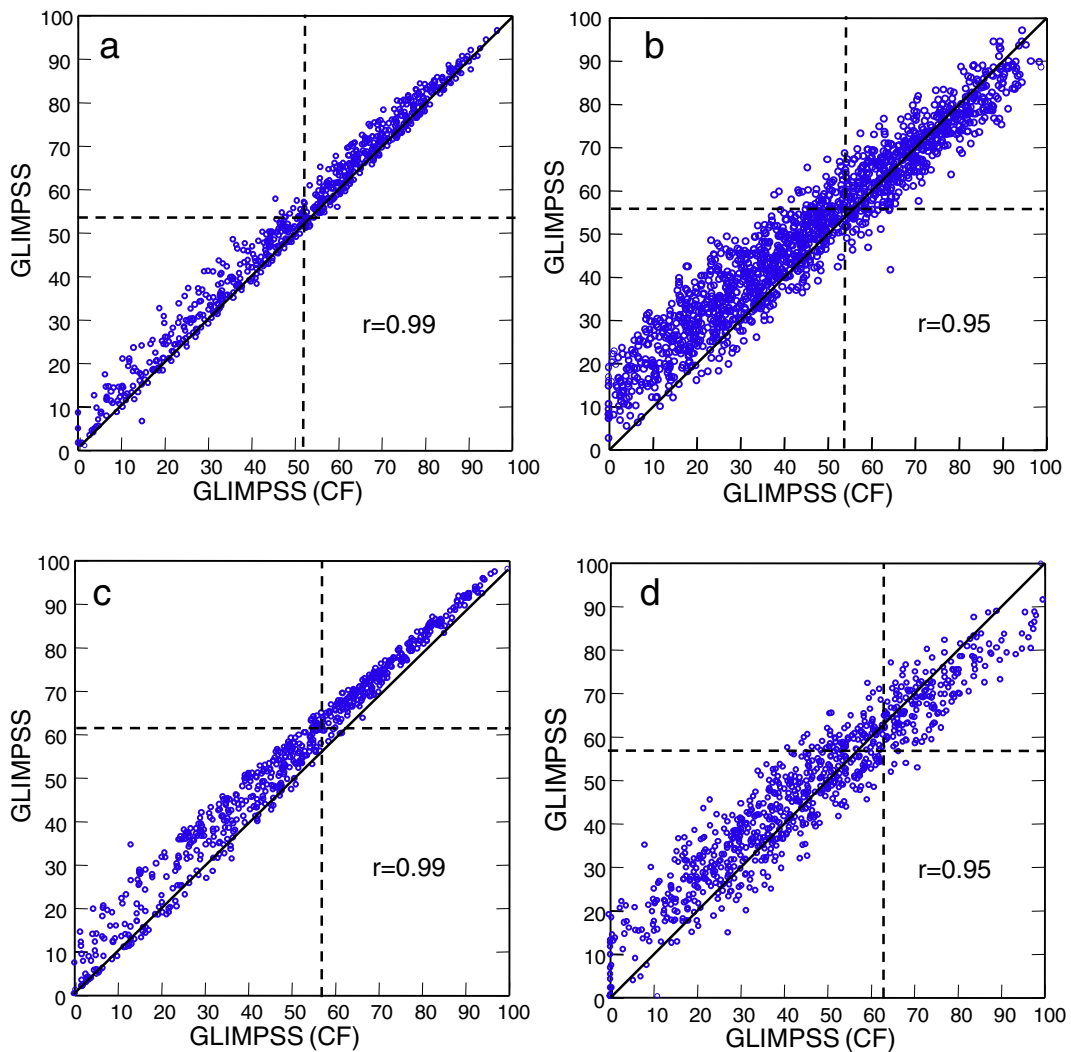
Fig. 5 Relationships between full GLIMPSS and modified GLIMPSS (*CF*) for MT Sp (**a**), MT Su (**b**), PL Sp (**c**), and PL Su (**d**) strata. *Solid lines* represent 1:1 relationship while dashed lines approximate the 5th centile thresholds of REF sites. See Fig. 2 and text for stratum definitions

responsiveness, precision and independent validation. Confounding of macroinvertebrate responses by seasonal (Julian day), and geographical (latitude, longitude, and elevation) influences was effectively removed by stratification. Stream size was also not significantly correlated to GLIMPSS scores (except in PL Su). Barbour et al. (1999) and Karr and Chu (1999) recommended that site classification (e.g., ecoregions, seasonal index periods) were critical to developing refined biocriteria using MMIs. Overall, the GLIMPSS models can be used for a broad spectrum of water resource management programs including (1) assessing streams for ALU attainability,

including characterizing the existence and extent of point and nonpoint source stressors, (2) targeting and prioritizing watersheds for remedial or preventive programs, (3) evaluating the effectiveness of nonpoint source best management programs and mitigation projects, and (4) identifying exceptional quality streams for enhanced protection and conservation. Future analyses should focus on using GLIMPSS models to compare relationships with different stressor and land use types, and to identify stressor thresholds for maintaining and protecting ALUs.

Genus-level taxonomy offered obvious improvements for bioassessments compared with family-level

Table 8 Final GLIMPSS and GLIMPSS (CF) scoring criteria for all strata

		MT Sp	MT Su	PL Sp	PL Su
GLIMPSS					
Very good	25th centile	≥66	≥67	≥66	≥65
Good	5th centile	53	55	61	57
Degraded	Upper bisection	26–52	27–54	30–60	28–56
Severely degraded	Lower bisection	<26	<27	<30	<28
GLIMPSS (CF)					
Very good	25th centile	≥64	≥69	≥66	≥70
Good	5th centile	51	54	57	62
Degraded	Upper bisection	26–50	27–53	29–56	31–61
Severely degraded	Lower bisection	<26	<27	<29	<31

Scores are rounded to nearest whole number. Impairment threshold is at the 5th centile value where bisection of this value demarcates degraded and severely degraded condition. See Fig. 2 and text for stratum definitions

MT Sp Mountain Spring, *MT Su* Mountain Summer, *PL Sp* Plateau Spring, *PL Su* Plateau Summer

WVSCI. First, genus-level taxonomy identified distinct seasonal and geographical classification strata, helping to refine ecological expectations across WV. Second, it allowed for the selection of several additional indicator metrics with larger response ranges, and could better track stressors in different seasons and within particular bioregions. Since macroinvertebrate communities were well-clustered within individual strata based on NMDS, we chose to construct individual MMIs for each bioregion and season. Vlek et al. (2004) argued that developing separate MMIs for each stratum avoided selection of metrics based on natural differences in streams (or seasons), rather than on the ability to detect degradation. In our dataset, several metrics differed seasonally and regionally in their ability to distinguish stressed sites and their measured response to a stressor gradient. Our results confirmed the need to test and select metrics in the context of a classification scheme such as bioregion and season which led to stratum-specific differences in impairment rates, possibly indicating regional differences in stressor intensity and the quality of the REF assemblages (Vlek et al. 2004).

While more sites were rated as impaired with GLIMPSS as compared with WVSCI overall, PL Su was an exception. This suggests that, in general, WVSCI over-identified degradation at sites in this stratum because WVSCI metric scoring and the impairment benchmark were based on statewide expectation criteria which were weighted heavily by REF sites in mountain ecoregions (Gerritsen et al. 2000). However, PL Su GLIMPSS had the highest CE and

assessed nearly 3x more "severely degraded" sites than WVSCI. Overall, GLIMPSS rated more than twice the number of sites as "severely degraded" compared with WVSCI across all strata.

The use of genus-level PTVs and the associated increase in range of the modified genus-level HBI, and the increase in range of taxa richness increases the sensitivity and range of response of the index (Lenat and Resh 2001). Some of the more obvious improvements in metrics using genus-level data were the increases in range within specific indicator groups (e.g., Ephemeroptera, Plecoptera, clingers, scrapers) and finer resolution of macroinvertebrate PTVs. For instance, common mayfly families (Heptageniidae and Baetidae) have family-level PTVs of 4. In WV, as many as 10 heptageniid genera and 13 baetid genera are known to occur with PTVs ranging from 0 to 4 and 2 to 10, respectively. Similarly, the family-level PTV for craneflies (Tipulidae) is 3, but 19 genera are reported in WV streams with genus-level PTVs ranging from 2 to 7. Comparatively, Pond et al. (2008) reported that genus HBI was more sensitive than family HBI, had wider range of response, and stronger correlations to stressors among WV unmined and coal mined catchments. For those metrics that count taxa richness, the gain in information can be substantial (e.g., loss of several heptageniid mayfly genera at a site would occur before the family was extirpated). We found this to be the case with other families, further confirming the benefit of using genus-level macroinvertebrate data for more accurate

assessments. Interestingly, the commonly employed total number of taxa metric (no. of total genera) was redundant with, and not nearly as sensitive or responsive as other higher performing metrics in some strata.

Since metric scoring and the GLIMPSS reference distributions differ across strata (seasons or regions), it is difficult to directly compare index scores between the strata without further standardization. Therefore, a "percent of threshold" value for each sample, calculated as actual stratum-specific MMI score divided by impairment threshold value, multiplied by 100, accomplishes this standardization. Thus, when compared with the 5th percentile of the reference distribution within a particular stratum, a percent of threshold value >100 % is unimpaired, while a score <100 % is impaired. Other applications of this method to interpret relative site ratings could be done by calculating the percent of other benchmarks found in Table 8 (e.g., the 25th percentile of REF; scores of <50 % of thresholds have severely degraded condition).

Conclusions and recommendations

A refined MMI developed for WV wadeable streams incorporates genus-level taxonomy and regional and seasonal stratification. The GLIMPSS is applicable for use in most moderate to high gradient streams throughout the state, including streams that do not flow year-round (based on the characteristics of the dataset used to develop the index). However, future sampling and analysis is required to develop assessments for true limestone springs, non-riffle/run low gradient streams, and larger rivers in WV. To correctly apply GLIMPSS, it is critical that WVDEP standardized field methodology (e.g., target habitat, sampling gear, and sample area) and laboratory processing (e.g., subsampling and taxonomic identification) are closely followed. Although the delineation of seasons is relatively straightforward, professional judgment should be applied when sample dates fall close to season cutoffs. For example, after a cooler than normal spring, sites at higher elevations may exhibit spring-like communities well into June. We recommend a 2- to 3-week buffer between seasons to remove seasonal uncertainties and improve assessment performance.

For streams that transcend plateau and mountain regions, the dominant contributing catchment area above the site should determine which suite of metrics and final assessment criteria should be applied. Autoecological designations (e.g., tolerance values for HBI and intolerant genera, FFGs for scraper or shredder genera, or habits for clinger genera) must be identical to those used by WVDEP as the GLIMPSS is specifically calibrated to those designations. Whether to use the full GLIMPSS or the GLIMPSS (CF) depends on chironomid taxonomic expertise and laboratory effort applied to the benthic sample. It was clear that the two indices performed comparably. However, the slight increase in sensitivity with the CF model suggests that although chironomid genera are known to be good indicators of certain stressors, it is permissible to exclude genus-level chironomid taxonomy for broader condition assessments with MMIs in WV streams.

The fundamental improvement of genus- over family-level data is representativeness. Compared with family-level taxonomic data, genus-level assessments more accurately represent the composition and diversity of the aquatic community in WV's flowing streams. Using macroinvertebrate data (one assemblage) to represent ALU (which can also include mussels, salamanders, fish, and a host of other aquatic life) is valid, robust, and cost-effective for identifying degraded streams. At a minimum, using genus-level macroinvertebrate data is critical in order to better detect impacts to aquatic resources, instead of relying on assessments based on coarser taxonomic resolution.

Appendix 1

Table 9 Summary statistics (minimum, median, and maximum) for selected environmental variables at reference (REF) and stressed (STRESS) sites among strata

	MT Sp REF				MT Sp STRESS				PL Sp REF				PL Sp STRESS			
	n	Min	Median	Max	n	Min	Median	Max	n	Min	Median	Max	n	Min	Median	Max
Stream Width (m)	125	0.7	2.8	15.8	89	0.6	2.6	13.7	43	0.9	2.7	17.5	193	0.6	3.0	19.3
Temperature (°C)	126	1.4	11.0	18.4	87	4.7	14.2	28.1	43	2.7	13.9	19.8	195	5.8	16.1	27.5
pH	126	5.3	6.9	8.2	87	2.8	7.3	8.8	43	6.5	7.3	8.4	195	6.1	7.6	10.1
Dissolved Oxygen (mg/l)	126	7.3	10.3	15.9	87	7.3	9.6	13.6	43	6.3	10.0	14.5	192	4.6	10.0	15.5
Specific Conductance (μS/cm)	126	15.4	42	265	87	24	196	2,767	43	50	154	502	195	55.5	214	4,953
Fecal Coliform (col/100 ml)	112	0	5	240	79	2	86	55,000	37	2	30	2,050	188	0.59	200	180,000
Sulfate (mg/l)	113	2	8	78	68	5	22.5	560	34	8	20.2	65.7	109	12	27	2,165
Chloride (mg/l)	111	1	1	16.8	56	1.0	5.6	629	34	1.8	4.7	14.7	108	1	6	1,180
TSS (mg/l)	106	1	3	38	75	1	3	59.2	29	2	3	10	152	1	3.3	52
Total P (mg/l)	111	0.01	0.02	0.39	65	0.01	0.02	0.65	36	0.01	0.02	0.1	113	0.01	0.02	0.2
NO_2–NO_3	111	0.01	0.38	1.1	65	0.01	0.3	9.81	35	0.01	0.12	1.72	110	0.01	0.1	3.3
Total Al (mg/l)	112	0.02	0.09	0.92	66	0.02	0.15	14.8	34	0.05	0.145	2.44	102	0.02	0.155	3.97
Total Fe (mg/l)	112	0.02	0.08	1.26	69	0.04	0.27	16.1	34	0.02	0.19	1.63	147	0.02	0.36	19
Total Mn (mg/l)	112	0.003	0.01	0.075	66	0	0.05	5.39	33	0.003	0.02	0.06	103	0.005	0.044	0.72
Total habitat score	126	132	168	197	88	77	111	185	41	134	151	169	196	55	104	159

	MT Su REF				MT Su STRESS				PL S u REF				PL S u STRESS			
	n	Min	Median	Max	n	Min	Median	Max	n	Min	Median	Max	n	Min	Median	Max
Stream width (m)	181	0.9	3.1	28.3	317	0.46	3.3	20.3	37	0.7	3.2	11.7	353	0.26	3.2	30.6
Temperature (°C)	181	9.9	16.2	24.8	306	10.4	19.4	31.5	37	11.9	19.4	23.4	354	12.9	21.6	31.9
pH	181	6	7.09	8.27	306	2.55	7.7	9.25	37	6.65	7.72	8.23	354	2.4	7.7	9.93
Dissolved oxygen (mg/l)	181	6	8.9	12.7	305	2.1	8.9	17.1	34	5.0	8.8	12.2	351	1.0	8.4	18.4
Specific conductance (μS/cm)	181	17	50	247	307	22	597	3,341	37	61	198	508	355	89	690	11,227
Fecal coliform (col/100 ml)	166	0	14	760	271	0	290	91,000	31	0	102	780	343	2	420	80,000
Sulfate (mg/l)	150	1	6.21	50.4	170	0.94	178.5	2,097	21	6	29	58	241	5	410	6,000
Chloride (mg/l)	147	1	2	37.5	113	0.3	7.0	125.0	21	1.4	3.9	14	134	1.1	18.4	1,673
TSS (mg/l)	144	1	3.143	20	176	1	4	58	23	1	5	26	271	1	4	144
Total P (mg/l)	150	0.01	0.02	0.35	108	0.01	0.027	1.31	21	0.01	0.02	0.06	134	0.01	0.02	0.26

Table 9 (continued)

	MT Su REF				MT Su STRESS				PL S u REF				PL S u STRESS			
	n	Min	Median	Max	n	Min	Median	Max	n	Min	Median	Max	n	Min	Median	Max
NO$_2$–NO$_3$	151	0.02	0.30	1.33	107	0.01	0.30	10.40	21	0.04	0.30	0.95	135	0.0	0.1	30.0
Total Al (mg/l)	151	0.008	0.094	0.790	169	0.02	0.10	62.70	21	0.02	0.10	0.87	239	0.02	0.11	55.20
Total Fe (mg/l)	151	0.009	0.130	1.31	170	0.02	0.23	99.8	23	0.02	0.11	1.06	272	0.01	0.41	200.0
Total Mn (mg/l)	151	0.002	0.014	0.550	169	0.00	0.06	7.81	21	0.01	0.02	0.05	237	0.00	0.11	18.20
Total RBP habitat score	181	136	164	195	317	53	113	184	37	135	150	172	356	49	111	164

MT Sp Mountain Spring, *MT Su* Mountain Summer, *PL Sp* Plateau Spring, *PL Su* Plateau Summer

Appendix 2

Table 10 List of all tested metrics, percent discrimination efficiency (DE), PCA correlation, maximum redundancy magnitude (Max Redun; within core selected metrics), and results of scope of impairment (SOI) tests among strata

Group	Metric	MT Sp				MT Su				PL Sp				PL Su			
		DE	PCA	Max Redun	SOI >1	DE	PCA	Max Redun	SOI >1	DE	PCA	Max Redun	SOI >1	DE	PCA	Max Redun	SOI >1
Rich	No. of total genera	71	-0.32			68	-0.31	0.69		40	-0.27			74	-0.37	0.64	
Rich	No. of EPT genera	89	-0.46			95	-0.55			74	-0.46			90	0.53		
Rich	No. of Ephemeroptera genera	79	-0.36	0.45		78	-0.43	0.39		66	-0.35	0.59		69	-0.52	0.46	
Rich	No. of Plecoptera genera	78	-0.53	0.74		91	-0.62	0.74		79	-0.38	0.65		89	-0.49		
Rich	No. of Trichoptera genera	81	-0.39	0.59		51	0.08		X	30	-0.28			44	-0.17		
Comp	% EPT	73	-0.46			75	-0.52			82	-0.34			68	-0.49		
Comp	% EPT (minus *Cheumatopsyche*)	76	-0.49			90	-0.59	0.74		83	-0.36	0.70		86	-0.54	0.75	
Comp	% EPT (minus *Cheumatopsyche*+ *Baetis*)	73	-0.55			86	-0.48			75	-0.36			86	-0.53		
Comp	% Ephemeroptera	75	-0.31	0.34		75	-0.30			68	-0.23		X	73	-0.47		X
Comp	% Ephemeroptera (minus *Baetis*)	71	-0.18			77	-0.27		X	64	-0.23			75	-0.47		X
Comp	% Plecoptera	62	-0.30		X	90	-0.41		X	80	-0.28		X	94	-0.51		X
Comp	% Trichoptera	63	-0.10		X	38	0.08		X	47	0.03		X	18	0.06		X
Comp	% Hydropsychidae	44	0.11			41	0.23			33	0.10			51	0.10		X
Comp	% Chironomidae	68	0.39			61	0.38			74	0.30			65	0.52	0.53	
Comp	% Annelida	44	0.22		X	44	0.11		X	37	0.15		X	24	0.06		X
Comp	% Chironomidae+ Annelida	73	0.49			62	0.39			75	0.31	0.70		61	0.52		
Comp	% Chironomini	28	0.23			28	0.26			50	0.14		X	54	0.37		
Comp	% Orthocladiinae	89	0.49	0.45		67	0.34	0.41		76	0.19			60	0.30		
Comp	% Tanytarsini	41	-0.33			39	0.02			41	0.17		X	37	0.08		
Comp	% Non-Insect	38	0.07			34	0.13			28	0.02			42	0.11		
Tol	No. of intolerant genera (PTV<3)	89	-0.58			95	-0.68			82	-0.45			98	-0.53	0.70	
Tol	No. of intolerant genera (PTV<4)	92	-0.61	0.74		95	-0.68	0.74	X	82	-0.49	0.67		96	-0.54		
Tol	Modified Hilsenhoff Biotic Index (HBI)	84	0.62	0.45		96	0.66	0.74		80	0.41	0.57		95	0.55	0.75	
Tol	% Tolerant (PTV>6)	70	0.34			64	0.29			77	0.24	0.45		58	0.44		
Dom	% 5 dominant genera	73	0.35	0.64		72	0.26	0.67		47	0.25			75	0.32	0.62	X

Table 10 (continued)

Group	Metric	MT Sp DE	PCA	Max Redun	SOI >1	MT Su DE	PCA	Max Redun	SOI >1	PL Sp DE	PCA	Max Redun	SOI >1	PL Su DE	PCA	Max Redun	SOI >1
FFG	No. of shredder genera	66	−0.36			71	−0.45	0.41		49	−0.20			37	−0.31		X
FFG	No. of scraper genera	68	−0.42	0.62		61	−0.24			46	−0.35			63	−0.43	0.44	
FFG	No. of collector genera	41	0.00			49	−0.15			21	−0.11			43	−0.21		
FFG	No. of filterer genera	38	−0.20		X	21	0.09			7	−0.03		X	21	−0.10		
FFG	No. of predator genera	61	−0.24			49	−0.26			47	−0.24			56	−0.17		
FFG	% shredders	51	−0.17		X	65	−0.30		X	67	−0.18		X	54	−0.14		X
FFG	% scrapers	79	−0.34		X	47	−0.03		X	44	−0.06		X	43	−0.33		X
FFG	% collectors	56	0.31			31	0.04			14	−0.04			30	0.13		
FFG	% filterers	22	−0.05			45	0.18			42	0.13			60	0.08		
FFG	% predators	48	−0.10			50	−0.11		X	52	0.02		X	52	0.09		
Habit	No. of clinger genera	75	−0.47	0.66		79	−0.39	0.69		61	−0.38	0.67		83	−0.54	0.64	
Habit	% burrowers	29	−0.01			33	0.15			23	0.13			29	0.43		
Habit	% climbers	22	0.06			22	0.13			52	0.22			37	0.11		
Habit	% clingers	64	−0.28			36	−0.03			44	−0.10			38	−0.27		
Habit	% sprawlers	40	−0.07			23	−0.12			12	−0.14			13	−0.02		
Habit	% swimmers	22	0.07			15	−0.14			4	−0.12			16	−0.33		

MT Sp Mountain Spring, *MT Su* Mountain Summer, *PL Sp* Plateau Spring, *PL Su* Plateau Summer

Appendix 3

Table 11 Upper and lower bounds (ceilings and floors) for metrics (by stratum) used for final scoring of GLIMPSS

GLIMPSS	MT Sp (n=693)		MT Su (n=1,530)		PL Sp (n=653)		PL Su (n=858)	
	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor
No. of total genera			38	14			34	14
No. of intolerant genera (PTV<4)	19	1	15	0	15	1		
No. of intolerant genera (PTV<3)							7	0
No. of Ephemeroptera genera	10	1	9	0	10	1	7	0
No. of Plecoptera genera	8	0	7	0	7	0		
No. of Trichoptera genera	7	1						
No. of clinger genera	20	4	19	5	17	3	15	4
Modified HBI (genus level)	6.19	2.25	6.2	2.79	6.65	2.51	6.32	3.82
% 5 dominant genera	92	48	91.4	51			91.5	52.8
% tolerant (PTV>6)					71.4	0		
% Ephemeroptera	59.9	0.48						
% EPT (minus *Cheumatopsyche*)			86	5.2	90.8	2.4	67.1	1.3
% Orthocladiinae	52.8	0.49	37.1	0				
% Chironomidae							68.8	3.3
% Chironomidae+Annelida					85.1	1.8		
No. of scraper genera	8	0					7	1
No. of shredder genera			5	1				

MT Sp=Mountain Spring, MT Su=Mountain Summer, PL Sp=Plateau Spring, PL Su=Plateau Summer

Table 12 Upper and lower bounds for metrics (by stratum) used for final scoring of GLIMPSS (CF)

GLIMPSS (CF)	MT Sp (n=693)		MT Su (n=1,530)		PL Sp (n=653)		PL Su (n=858)	
	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor
No. of total genera			30	8			25	8
No. of intolerant genera (PTV<4)	18	1	15	0	14	1		
No. of intolerant genera (PTV<3)							7	0
No. of Ephemeroptera genera	10	1	9	0	10	1	7	0
No. of Plecoptera genera	8	0	7	0	7	0		
No. of Trichoptera genera	7	1						
No. of clinger genera	19	3	18	4	16	3	14	3
Modified HBI (genus level)	5.87	2.2	5.9	2.8	5.94	2.46	5.98	3.84
% 5 dominant genera	96.7	56	96.7	57.7			97.5	64.4
% Ephemeroptera	60.2	0.48						
% EPT (minus *Cheumatopsyche*)			86	5.2	90.8	2.4	67.1	1.3
% Chironomidae							68.8	3.3
% Chironomidae+Annelida	76.1	2.8			85.1	1.8		
No. of scraper genera	8	0					7	1
No. of shredder genera			4	0				

MT Sp Mountain Spring, *MT Su* Mountain Summer, *PL Sp* Plateau Spring, *PL Su* Plateau Summer

References

Arscott, D. B., Jackson, J. K., & Kratzer, E. B. (2006). Role of rarity and taxonomic resolution in a regional and spatial analysis of stream macroinvertebrates. *Journal of the North American Benthological Society, 25*, 977–997.

Bailey, R. C., Norris, R. H., & Reynoldson, T. B. (2001). Taxonomic resolution of benthic macroinvertebrate communities in bioassessments. *Journal of the North American Benthological Society, 20*, 280–286.

Baptista, D. F., Buss, D. F., Engler, M., Giovanelli, A., Silveira, M. P., & Nessimian, J. L. (2007). A multimetric index based on benthic macroinvertebrates for evaluation of Atlantic Forest streams at Rio de Janeiro State, Brazil. *Hydrobiologia, 575*, 83–94.

Barbour, M.T., Gerritsen, J., Snyder, B.D., Stribling, J.B. (1999). Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates, and fish, second edition. EPA 841-B-99-002. US Environmental Protection Agency, Office of Water, Washington, DC.

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

Blocksom, K. A., & Johnson, B. R. (2009). Development of a regional macroinvertebrate index for large river bioassessment. *Ecological Indicators, 9*, 313–328.

Carter, J. L., & Resh, V. H. (2001). After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. *Journal of the North American Benthological Society, 20*, 658–682.

Gerritsen, J., Burton, J., & Barbour, M. T. (2000). *A stream condition index for West Virginia wadeable streams*. Owing Mills: Tetra Tech, Inc.

Hawkins, C. P., & Norris, R. H. (2000). Performance of different landscape classifications for aquatic bioassessments: introduction to the series. *Journal of the North American Benthological Society, 19*, 367–369.

Hering, D., Feld, C. K., Moog, O., & Ofenbock, T. (2006). Cook book for the development of a multimetric index for biological condition of aquatic ecosystems: experiences from the European AQEM and STAR projects and related initiatives. *Hydrobiologia, 566*, 311–324.

Hughes, R. M. (1995). Defining acceptable biological status by comparing with reference conditions. In W. S. Davis & T. P. Simon (Eds.), *Biological assessment and criteria: Tools for water resource planning and decision making* (pp. 31–47). Boca Raton: Lewis Publishers.

Jones, F. C. (2008). Taxonomic sufficiency: the influence of taxonomic resolution on freshwater bioassessments using benthic macroinvertebrates. *Environmental Reviews, 16*, 45–69.

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

Klemm, D. J., Blocksom, K. A., Fulk, F. A., Herlihy, A. T., Hughes, R. M., Kaufmann, P. R., Peck, D. V., Stoddard, J. L., Thoeny, W. T., Griffith, M. B., & Davis, W. S. (2003). Development and evaluation of a macroinvertebrate Biotic Integrity Index (MIBI) for regional assessing Mid-Atlantic Highland streams. *Environmental Management, 31*, 656–669.

Lenat, D. R., & Resh, V. H. (2001). Taxonomy and stream ecology—the benefits of genus- and species-level identifications. *Journal of the North American Benthological Society, 20*, 287–298.

McCune, B., & Grace, J. B. (2002). *Analysis of ecological communities*. Gleneden Beach: MjM Software Design.

Merritt, R. W., Cummins, K. W., & Berg, M. B. (Eds.). (2008). *An introduction to the aquatic insects of North America* (4th ed.). Dubuque: Kendall/Hunt.

Pond, G.J., Call, S.C., Brumley, J.F., & Compton, M.C. (2003). The Kentucky macroinvertebrate bioassessment index: derivation of regional narrative criteria for headwater and wadeable streams. Kentucky Department for Environmental Protection, Division of Water, Frankfort, Ky. Available at: http://water.ky.gov/Documents/QA/MBI/Statewide_MBI.pdf

Pond, G. J., Passmore, M. E., Borsuk, F. A., Reynolds, L., & Rose, C. J. (2008). Downstream effects of mountaintop coal mining: comparing biological conditions using genus- and family-level bioassessment tools. *Journal of the North American Benthological Society, 27*, 717–737.

Resh, V. H., & Jackson, J. K. (1993). Rapid assessment approaches to biomonitoring using benthic macroinvertebrates. In D. M. Rosenberg & V. H. Resh (Eds.), *Freshwater biomonitoring and benthic macroinvertebrates* (pp. 195–233). Norwell: Kluwer Academic Publishers.

Reynoldson, T., Rosenberg, D. M., & Resh, V. H. (2001). Comparison of models predicting invertebrate assemblages for biomonitoring in the Fraser River catchment, British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences, 58*, 1395–1410.

Rosenberg, D. M. (1992). Freshwater biomonitoring and Chironomidae. *Aquatic Ecology, 26*, 101–122.

Rosenberg, D. M., & Resh, V. H. (1993). Introduction to freshwater biomonitoring and benthic macroinvertebrates. In D. M. Rosenberg & V. H. Resh (Eds.), *Freshwater biomonitoring and benthic macroinvertebrates* (pp. 1–9). Norwell: Kluwer Academic Publishers.

Schmidt-Kloiber, A., & Nijboer, R. C. (2004). The effect of taxonomic resolution on the assessment of ecological water quality classes. *Hydrobiologia, 516*, 269–283.

Southerland, M. T., Rogers, G. M., Kline, R. J., Morgan, R. P., Boward, D. M., Kazyak, P. F., Klauda, R. J., & Stranko, S. A. (2007). Improving biological indicators to better assess the condition of streams. *Ecological Indicators, 7*, 751–767.

Stoddard, J.L., Peck, DV., Paulsen, S.G., Van Sickle, J.C., Hawkins, P., Herlihy, A.T., et al. (2005). *An Ecological Assessment of Western Streams and Rivers*. EPA 620/R-05/005, US Environmental Protection Agency, Washington, DC.

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

Stoddard, J. L., Herlihy, A. T., Peck, D. V., Hughes, R. M., Whittier, T. R., & Tarquinio, E. (2008). A process for creating multimetric indices for large-scale aquatic surveys.

Journal of the North American Benthological Society, 27, 878–891.

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

Sweeney, B. W. (1984). Factors influencing life-history patterns of aquatic insects. In V. H. Resh & D. M. Rosenberg (Eds.), *The ecology of aquatic insects* (pp. 56–100). New York: Praeger.

US Environmental Protection Agency (US EPA) (1996). Biological criteria: Technical guidance for steams and small rivers, revised edition. Report number EPA-440-5-91-003. Office of Water. Washington DC

Van Sickle, J. (1997). Using mean similarity dendrograms to evaluate classifications. *Journal of Agricultural, Biological, and Environmental Statistics, 2*, 370–388.

Verandes, S. G., & Cortes, R. M. V. (2010). Evaluating macroinvertebrate metrics for ecological assessment of streams in northern Portugal. *Environmental Monitoring and Assessment, 166*, 201–221.

Vlek, H. E., Verdonschot, P. F. M., & Nijboer, R. C. (2004). Towards a multimetric index for the assessment of Dutch streams using benthic macroinvertebrates. *Hydrobiologia, 516*, 173–189.

West Virginia Department of Environmental Protection (WVDEP) (2011). Watershed Branch 2011 Standard Operating Procedures. West Virginia Department of Environmental Protection, Division of Water and Waste Management, Watershed Assessment Branch. Charleston, WV. Available at: http://www.dep.wv.gov/WWE/watershed/Pages/WBSOPs.aspx. Accessed 27 December 2011.

Whittier, T. R., Hughes, R. M., Stoddard, J. L., Limnicky, G. A., Peck, D. V., & Herlihy, A. T. (2007). A structured approach for developing indices of biotic integrity: three examples from streams and rivers of the western USA. *Transactions of the American Fisheries Society, 136*, 718–735.

Woods, A.J., Omernik, J.M., Brown, D.D., & Kiilsgaard, C.W. (1996). Level III and IV Ecoregions of Pennsylvania and the Blue Ridge Mountains, the Ridge and Valley, and the Central Appalachians of Virginia, West Virginia, and Maryland. EPA/600R-96/077. USEPA, ORD, Corvallis, OR.