Ecological Indicators 40 (2014) 117–126

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Evaluation of suitability and comparability of stream assessment indices using macroinvertebrate data sets from the Northern Lakes and Forests Ecoregion

Stephanie A. Ogren[a,*], Casey J Huckins[b]

[a] Little River Band of Ottawa Indians, Natural Resources Department, 375 River Street, Manistee, MI 49660, USA
[b] Michigan Technological University, Department of Biological Sciences, 1400 Townsend Drive, Houghton, MI 49931, USA

ARTICLE INFO

Article history:
Received 25 July 2013
Received in revised form 7 January 2014
Accepted 14 January 2014

Keywords:
Stream assessment
Macroinvertebrates
Bioassessment
Index of biotic integrity
Biological condition
Scale

ABSTRACT

Researchers and managers within the Upper Midwest currently use a variety of sampling methodologies and biological indices to assess ecological condition of stream systems. With multiple entities collecting bioassessment data it is important that we determine the comparability of data and the indices derived from these data for effective assessment of natural systems. In this study we assessed the similarity of data collected by different agencies and we focused on data from one watershed to examine the outputs of different indices for stream assessment, and the temporal variation of index score within sites. We compared duplicate macroinvertebrate community data collected by the Little River Band of Ottawa Indians and the Michigan Department of Environmental Quality for overall community composition and index scores derived from these data. Duplicate samples were similar in composition index scores. Taxonomic resolution was addressed and indicated that genus level resolution gives a more favorable score when using indices. We also evaluated the utility of currently available macroinvertebrate indices of biotic integrity to assess data from the Big Manistee River watershed. The indices evaluated were the Hilsenhoff biotic index, the benthic community index for the Northern Lakes and Forests (NLFBCI), the Great Lakes Environmental Assessment Survey (GLEAS) procedure 51 for macroinvertebrates and a biological condition gradient model for the Upper Midwest. Outputs from the indices were moderately correlated (Spearman rank order correlation, $r = 0.35–0.698$) though they indicated different assessments of overall site integrity. Compared with larger scale regional indices, locally calibrated indices generally classified sites as having better biological condition. Replicate samples collected within sites indicated the GLEAS had higher levels of variability (0–265%CV) within sites than the other indices (<10%CV). Data from long-term (10 year) monitoring stations were used to evaluate seasonal and long-term index performance. There were differences in index score classifications from spring and fall samples indicating that standardization of sampling time is necessary for comparative analysis. Temporal trends over 10 years reveal natural variation and set the baseline for evaluating the influence of anthropogenic effects. Overall, results indicate that choice of index can alter assessment of site condition. For bioassessment in the Big Manistee River watershed the NLFBCI performs well and accurately reflects site condition.

1. Introduction

Whether the goal is to protect a relatively pristine ecosystem, manage an actively used system, or restore a degraded one, the approach and success relies on our knowledge and assessment of the physical and biological condition of ecosystems. Aquatic biological monitoring has been recognized as the first step in protecting biological integrity (Karr and Chu, 1999). Assessing the ecological condition of a site may be approached through multiple methods, often with the estimated biological condition dependent on many factors, including the organisms selected for use in the interpretation (Carter and Resh, 2001), how data are interpreted (Cao et al., 2005), and methods used to collect the data (Hughes and Peck, 2008).

Numerous national, regional and local organizations have independently developed aquatic assessment programs producing many innovative technical approaches for data acquisition and interpretation (Davies and Jackson, 2006) but with little standardization; therefore, determining the comparability of data collected

* Corresponding author. Tel.: +1 231 398 2192; fax: +1 231 723 8873.
E-mail address: sogren@lrboi.com (S.A. Ogren).

http://dx.doi.org/10.1016/j.ecolind.2014.01.018
1470-160X/© 2014 Elsevier Ltd. All rights reserved.

and resulting assessments is needed (Cao and Hawkins, 2011). The ability to utilize multiple sources of data could benefit programs by allowing for validation of assessments if they are shown to be comparable (Herbst and Silldorff, 2006; Gerth and Herlihy, 2006; Rehn et al., 2007).

Biotic indices have been developed to aid in the interpretation of biological assessment data. The product of a biotic index is a single site- and time-specific numeric score that can be interpreted within a regional gradient of condition (Karr and Chu, 1999). Assessment of the utility and applicability of these indices over spatial (Ode et al., 2008) and temporal scales is also necessary (Mazor et al., 2009). Determining comparability of this numeric score and inferences derived from these endpoints has become necessary to improve regulatory credibility, reduce redundancy, increase efficiency, improve long-term monitoring programs, expand assessments to a broader scale and generally increase sample size, which would improve assessment (Cao and Hawkins, 2011). In the Upper Midwest of the United States there are numerous indices available; however, determining the appropriate index and when to apply it is problematic. One biological data set can be interpreted in different ways and subsequently indicate different courses of action based on which index is applied.

Agencies within the Upper Midwest currently use disparate sampling methodologies and biological indices to assess stream systems. Often, management agencies use indices that are not directly comparable, having varying scales and different classification schemes. One of our goals for this study was to determine if indices developed for use at different spatial scales in the Upper Midwest (Fig. 1) would produce concordant index scores within and across sites. We used a nested approach to evaluate sites based on scores from indices developed with increasing geographic scope. By nested approach we mean that the data set from the Big Manistee River watershed is within the state of Michigan, which is within the Northern Lakes and Forests Ecoregion within the Upper Midwest. Scoring of sites is in comparison to reference condition or theoretical natural state utilized in the original development of the index. The natural variation across a larger region may limit discrimination of site specific differences in a regionally derived index. A locally derived index may be necessary for discrimination of smaller changes in biotic integrity (Ode et al., 2008). A nested approach to data interpretation may lead to better understanding of variation in ecological condition and the geographic scope appropriate for interpretation.

Evaluation of stream condition is also dependent on the temporal stability of a system (Milner et al., 2006). Temporal variation in community assemblage occurs both seasonally and annually. Seasonal variability has been shown by others to be dependent on the system evaluated (Linke et al., 1999; Morais et al., 2004; Maloney and Feminella, 2006; Callanan et al., 2008; Kappes et al., 2010). Annual variation has been less well studied (Jackson and Fureder, 2006) however; it has been shown that understanding annual variation is necessary to improve bioassessment when disturbance is subtle (Huttunen et al., 2012).

We evaluated the utility of currently available macroinvertebrate indices of biotic integrity to assess macroinvertebrate community data from the Big Manistee River watershed data set from the northwest Lower Peninsula of Michigan, USA. The five indices evaluated include the Hilsenhoff (HBI) (family and genus level) biotic indices (Hilsenhoff, 1987, 1988), the benthic community index for the Northern Lakes and Forests (NLFBCI) (Butcher et al., 2003), the Great Lakes Environmental Assessment Survey (GLEAS) procedure 51 for macroinvertebrates (Creel et al., 1998) and a Biological Condition Gradient model (BCG) for the Upper Midwest (Gerritesn and Stamp, 2012). The HBI was developed to evaluate organic stream pollution based on genus or family level tolerance values (G-HBI, F-HBI, respectively) for Wisconsin macroinvertebrates. Community-based indices are used to assess biological integrity using a combination of metrics such as native composition and relative sensitivity to environmental conditions. For example, the NLFBCI is a genus level assessment useful for delineating impaired sites from non-impaired sites in the Northern Lakes and Forests Ecoregion. The GLEAS was developed for use in Michigan with separate family level scoring for each ecoregion in the state resulting in a narrative classification of site scores as excellent, acceptable, or poor. The BCG, originally described by Davies and Jackson (2006), was calibrated for use in the Upper Midwest (Gerritesn and Stamp, 2012) and is based on the relationship between stressors in the environment and corresponding ecological response of the aquatic community indicated with a numeric value from one to six. In this study, macroinvertebrate community data collected through the Little River Band of Ottawa Indians (LRBOI) baseline monitoring and assessment program as well as State of Michigan Department of Environmental Quality (MI-DEQ) macroinvertebrate community data from the trend monitoring program were compiled and analyzed with available indices.

The objectives of this study were to (1) determine if data from multiple agencies could be effectively combined and integrated into a larger watershed dataset and (2) assess concordance of regional indices.

2. Methods

2.1. Study area

The Big Manistee River watershed (Fig. 1) is in the northern Lower Peninsula of Michigan, has an area of approximately 490,000 ha, spans 11 counties and includes the 1836 Reservation of the Little River Band of Ottawa Indians (LRBOI). The watershed is primarily forested (56%), with scrub/shrub and grassland covering 16% and wetlands comprising an additional 13%. There is some agricultural use in the form of grazing and row crops (9%) with developed land covering 6% of the watershed (NLCD, 2006). There are 3191 km of stream within the Big Manistee River watershed (NLCD, 2006). The lower portion of the Big Manistee River is federally recognized as a wild and scenic river with upper portions of the mainstem and sections of tributaries designated by the State of Michigan as Natural Rivers and Blue Ribbon Trout Streams.

2.2. Data acquisition

The LRBOI Natural Resources Department sampled benthic macroinvertebrates annually, beginning in 2002, using a multihabitat rapid bioassessment protocol (Barbour et al., 1999) to provide data for biological assessment of the watershed. Sampling occurred seasonally in the spring and fall of each year (2002–2011) at four long-term, fixed monitoring sites with reach lengths 40 × stream width. Habitat types (e.g., riffles and pools) were sampled in approximate proportion to their representation of surface area. Macroinvertebrates were preserved and identified in a laboratory. Additionally, three simultaneous replicate samples were collected from nine independent stream reaches in 2009. Three reaches were located on Sickle Creek, Bear Creek, and Pine Creek respectively ($n=9$), and were separated by a distance of 40 × stream width. Macroinvertebrate data was also acquired from State of Michigan assessments. In 2009 the State of Michigan MI-DEQ conducted an assessment of 23 sites in the Big Manistee River Watershed as part of the state monitoring program, which is on a 5 year watershed rotation (Lipsey, 2010). Macroinvertebrate assessments conducted through this effort followed the Great Lakes and Environmental Assessment Section (GLEAS) Procedure 51 (Creel et al., 1998) protocols. This protocol is used by the State of Michigan for

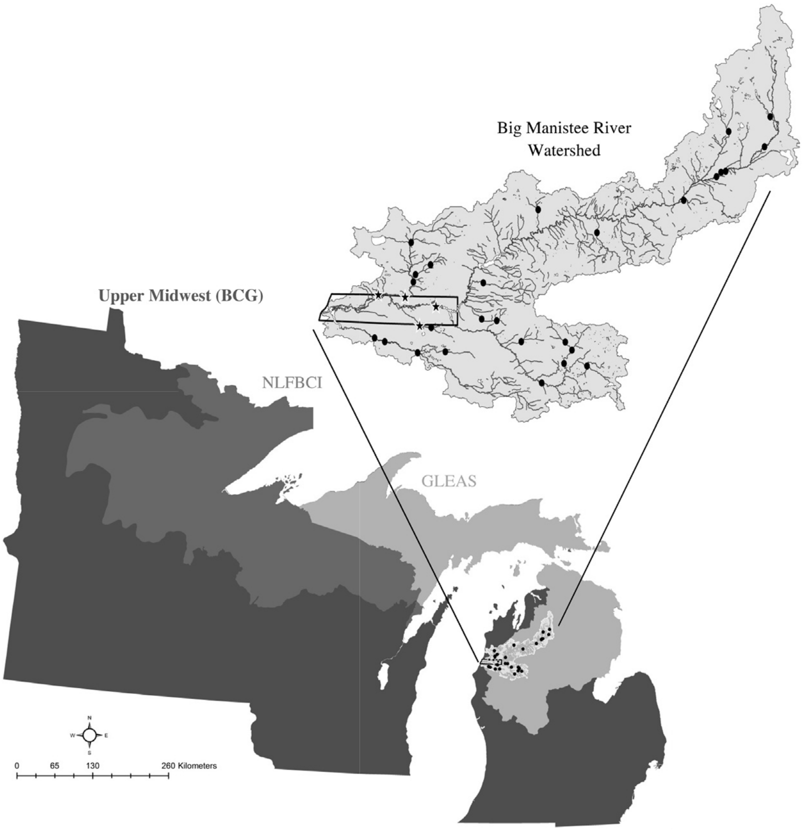


Fig. 1. Location of watershed sampling sites within the Big Manistee River watershed located in Michigan, USA. Area depicted includes Upper Midwest (dark gray) where the Biological Condition Gradient (BCG) model was calibrated, Northern Lakes and Forests Ecoregion (gray) where a biotic condition index (NLFBCI) was calibrated, the Michigan portion of the ecoregion (light gray) where the Great Lakes Environmental Assessments Section (GLEAS) index was calibrated and the Big Manistee River watershed where data was collected. The dark rectangle in the watershed is the Little River Band of Ottawa Indians 1836 exterior reservation boundaries where long term data was collected. All dots in the watershed are sampling locations with long term stations identified with a star.

biological assessments throughout the State and is very similar to the LRBOI protocol with proportional habitat being sampled for macroinvertebrate community composition. The GLEAS samples were subsampled to 100 organisms and processed to family in the field while the LRBOI samples were subsampled to 300 organisms and processed to family in a laboratory from 2002 to 2007 and to genus from 2008 to 2011.

2.3. Index applicability

Regional macroinvertebrate indices derived from three spatial scales were evaluated for interpretation of bioassessment data. Macroinvertebrate indices were considered if they were developed for use in the Upper Midwest, the Northern Lakes and Forests Ecoregion (Omernik, 1987) or the State of Michigan (Fig. 1). Sampling protocols for each index including collection method, timeframe and thermal regime were included in the comparison. The applicability matrix highlights differences in the requirements for each index that met the above criteria (Table 1). A total of five index scores were calculated: the NLFBCI, the G-HBI and F-HBI, the GLEAS and a BCG model for the Upper Midwest (Table 2). All indices were generally similar in approach, though each had slight variations in sampling protocols and data requirements. All indices were developed for use in cold water systems with multiple habitats in the stream reach sampled during field collections. However, there were some discrepancies in taxonomic resolution requirements (family or genus) and spatial scales of index development. Scale of index development ranged from the entire Upper Midwest to a specific ecoregion within the state of Michigan.

2.4. Data resolution and source

The feasibility of integrating datasets was evaluated through comparison of taxonomic resolution and precision of duplicate assessment. To evaluate the effect of taxonomic resolution on index sensitivity, index scores from each of four long-term monitoring sites were calculated from data with genus and family level resolution. Comparisons of truncated data (i.e., family level) to original genus level data were completed for eight paired samples (Spring and Fall for each year, 2008–2011) for each site utilizing a Wilcoxon Signed Rank Test in SigmaPlot version 12.2 (Systat Software Inc., 2012).

Consistency of macroinvertebrate sampling between agencies was verified using data from three duplicate sites sampled by both MI-DEQ and LRBOI in 2009. Both agencies sampled the same reaches independently using their respective protocols. Duplicate sites were of average quality and similar in size to other sites in the watershed assessment. LRBOI data was converted from genus level to family level resolution to match the lowest taxonomic unit available for MI-DEQ data. Index scores generated from these samples were evaluated with a Mantel test (Mantel, 1967) to determine if the two sampling efforts produced similar data (Mazor et al., 2010). Multivariate analysis (Mantel test) was conducted in PC-ORD version 6.0 (McCune and Grace, 2002; McCune and Mefford, 2006) for MI-DEQ and LRBOI macroinvertebrate community data. Mantel's R was used to determine correlation between community compositions of samples. Sorensen distance was used as a dissimilarity measure for the paired LRBOI and MI-DEQ matrices.

Table 1
Index applicability matrix describing core attributes for each of five indices used for analysis: Hilsenhoff biotic index (F-HBI, G-HBI), Northern Lakes and Forests benthic community index (NLFBCI), Great Lakes Environmental Assessment Section (GLEAS) index and the biological condition gradient (BCG). The BCG has different models for cool and cold water streams based on mean July temperatures. We used the cold water model.

	F-HBI	G-HBI	NLFBCI	GLEAS	BCG
Development region	WI	WI	Ecoregion	MI	MN,WI,MI
Taxonomic resolution	Family	Genus	Genus	Family	Genus
Sampling protocol*	Multihabitat[a]	Multihabitat[a]	Multihabitat[b]	GLEAS 51[c]	RBP[d]
Temperature regime	Regional	Regional	Regional	Ecoregion	(<17.5 °C)

* References for sampling protocols: (a) Hilsenhoff (1987), (b) Chirhart (1998), (c) Creel et al. (1998), and (d) Gerritesn and Stamp (2012).

2.5. Index precision

Replicate samples were collected to determine the effect of within site variability on index scores. To assess index score repeatability, three simultaneous replicate samples were collected by LRBOI in 2009 at each of nine independent site locations. Coefficient of variation (CV) and standard deviation (SD) were calculated for each site. The comparability of index scores was evaluated using numeric scales, thresholds, and classification systems among indices. To determine whether the different indices resulted in the correlated assessment of the macroinvertebrate communities, Spearman rank order correlation, were completed in SigmaPlot version 12.2 (Systat Software Inc., 2012), among index scores from 30 independent site assessments conducted throughout the watershed in 2009. Data was aggregated from the 23 MI-DEQ assessments, the four long-term monitoring LRBOI sites and an additional three sites from the replicate sampling (one site from each stream) for a total of 30 sites. For comparisons of NLFBCI, GLEAS and HBI indices, scores were calculated at the family level as that was taxonomic resolution available for MI-DEQ data. To compare the BCG scores, genus level data had to be used, which was available from the four long-term monitoring sites collected seasonally over four years ($N = 8$ for each site). All comparisons with the BCG were made with indices calculated from genus level resolution. The intent of the analysis was to look at scores generated from samples and the relationship of those scores.

2.6. Site assessments

Index scores from 30 sites throughout the Big Manistee River watershed were assessed to determine if sites scored similarly across indices. Data was aggregated from the 23 MI-DEQ assessments, the four long-term monitoring LRBOI sites and an additional three sites from the replicate sampling (one site from each stream) for a total of 30 sites. Indices were calculated based on family level resolution as that was the lowest taxonomic level available for MI-DEQ data. Proportion of sites in various numeric and categorical rankings of indices were calculated. Site assessments based on the nested indices (GLEAS being locally calibrated and NLFBCI being regionally calibrated) were compared by assessing proportionate divergence of scores away from a specific threshold in the index. A Wilcoxon Signed Rank test was completed in SigmaPlot version 12.2 (Systat Software Inc., 2012) to test the pairs of index scores (GLEAS and NLFBCI) generated from each site and the difference in the proportionate divergence from the threshold.

2.7. Temporal trends

Seasonal index scores from 2002 to 2011 from the four long-term monitoring locations were analyzed using a Wilcoxon Signed Rank test in SigmaPlot version 12.2 (Systat Software Inc., 2012) to determine if there were seasonal affects discernible by index score. Spring and fall samples were paired for the analysis and if a season was missing data, that pair was omitted in the statistical analysis. Index scores were also plotted against time to examine seasonal trends by year at each site and evaluate variability (CVs). Each of the index scores were calculated for four long-term monitoring stations in the watershed over ten years to track index output over time.

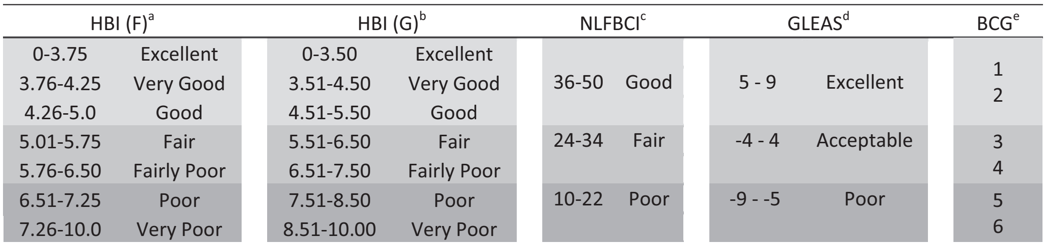
3. Results

3.1. Data resolution and source

The index scores calculated from family and genus level taxonomic data were not significantly different for the HBI (family and genus) at any of the sites ($P > 0.37$, Wilcoxon Signed Rank test) (Fig. 2a). The index scores from the GLEAS and the NLFBCI (Fig. 2b and c) genus and family level pairs were significantly different at all sites tested ($P < 0.01$, Wilcoxon Signed Rank test). The output scores from the GLEAS and NLFBCI were greater (suggesting better condition) when calculated from genus level data.

Family level community composition data from the three sites with both MI-DEQ and LRBOI data were analyzed with a Mantel test in PC-ORD (Table 3). The Mantel test compared matrices based on species composition and indicated paired matrices were significantly correlated ($P < 0.01$). Sampling completed by MI-DEQ and

Table 2
Numeric index scores and associated classification levels for the Hilsenhoff biotic index (HBI-F, HBI-G), the Northern Lakes and Forests benthic community index (NLFBCI), the Great Lakes and Environmental Assessment Section (GLEAS) index and the numeric levels for the biological condition gradient (BCG) model for the Upper Midwest. Color gradations indicate groupings based on similarities in classification levels.



HBI (F)[a]		HBI (G)[b]		NLFBCI[c]		GLEAS[d]		BCG[e]
0-3.75	Excellent	0-3.50	Excellent	36-50	Good	5 - 9	Excellent	1 2
3.76-4.25	Very Good	3.51-4.50	Very Good					
4.26-5.0	Good	4.51-5.50	Good					
5.01-5.75	Fair	5.51-6.50	Fair	24-34	Fair	-4 - 4	Acceptable	3 4
5.76-6.50	Fairly Poor	6.51-7.50	Fairly Poor					
6.51-7.25	Poor	7.51-8.50	Poor	10-22	Poor	-9 - -5	Poor	5 6
7.26-10.0	Very Poor	8.51-10.00	Very Poor					

*References for scoring: (a) Hilsenhoff (1987), (b) Hilsenhoff (1988), (c) Butcher et al. (2003), (d) Creel et al. (1998), and (e) Davies and Jackson (2006).

Table 3
Sampling of same sites (PLD, PLU and BCR) independently completed by agency staff (LRBOI, MI-DEQ) and evaluated for index scores and Mantel's r based on family level community composition data. Index scores include the family level Hilsenhoff biotic index (F-HBI), the Northern Lakes and Forests Ecoregion benthic community index (NLFBCI) and the Great Lakes and Environmental Assessment Section (GLEAS) index.

	PLD		PLU		BCR	
	LRBOI	MI-DEQ	LRBOI	MI-DEQ	LRBOI	MI-DEQ
F-HBI	4.32	4.36	3.99	4.23	4.01	3.63
NLFBCI	36	36	32	34	36	34
GLEAS	4	4	1	2	5	4
Mantel's r	0.151		0.353	0.109		

LRBOI at three sites indicated index scores were similar. Data collected by these two agencies resulted in identical scores for both the NLFBCI and the GLEAS at one site (PLD). The other two sites only varied by one point for the GLEAS scores and two points (one classification level) for the NLFBCI. The F-HBI scores fell within the same scoring level for PLD (Good) and PLR (Very Good) while BCR was scored as an Excellent site based on MI-DEQ sampling and a Very Good site based on LRBOI sampling.

3.2. Index precision

Coefficient of variation and standard deviation based on three simultaneous replicate samples were variable in precision depending on the index used (Fig. 3). The F-HBI and G-HBI ranged from 0 to 9% CV and 0.01 to 0.40 SD and 0 to 8% CV and 0.01 to 0.30 SD respectively. Sites scored with the NLFBCI ranged from 0 to 8% CV and 0 to 2.31 SD. When scored with the BCG, samples ranged from 0 to 25% CV and 0 to 0.58 SD. The GLEAS scores had the most variability and ranged from 0 to 265% CV and 0 to 2.65 SD for specific site replicates. The average coefficient of variation (CV) across sites was under 10% for all metrics except for the GLEAS, which was 68%. Results were similar for the average standard deviation (SD) with the F-HBI, G-HBI and the BCG all under 0.25 SD. Both the NLFCBI and the GLEAS had higher average standard deviations above 1.0 SD.

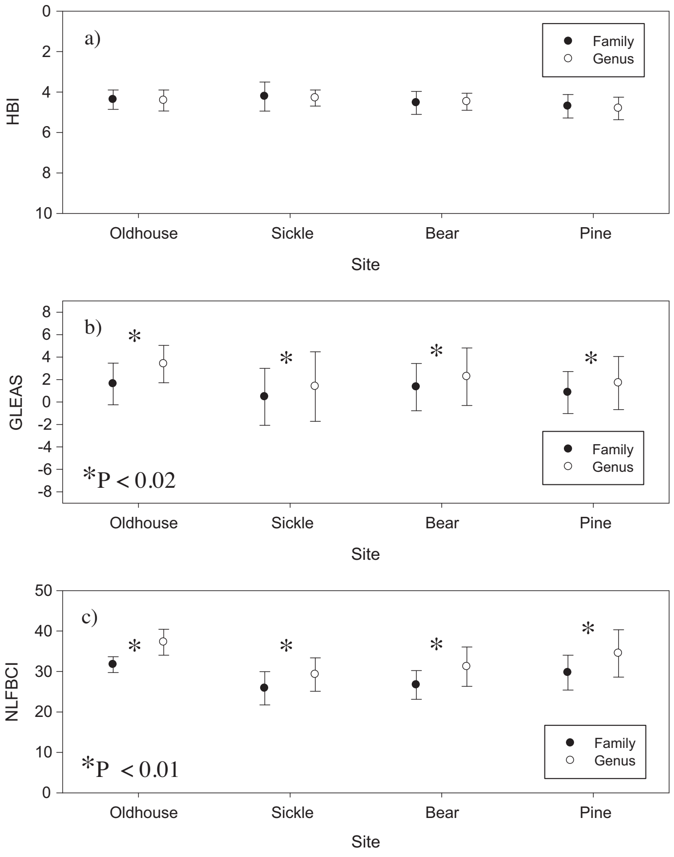


Fig. 2. Comparison of family and genus level scores for three indices: (a) Hilsenhoff biotic index, (b) Great Lakes Environmental Assessment Section index and (c) Northern Lakes and Forests benthic community index. Paired scores (genus and family outputs) at each site were tested using 8 samples at each site collected from 2008 to 2011 (Wilcoxon Signed Rank test).

Correlations among indices were varied (Fig. 4). Spearman rank order correlation analysis showed output scores from the G-HBI were not significantly correlated with the scores from the BCG ($P = 0.60$, $\rho = 0.204$) however, the F-HBI was correlated with both the NLFBCI ($P < 0.01$, $\rho = 0.553$) and the GLEAS ($P < 0.01$, $\rho = 0.629$). The NLFBCI was significantly correlated with the GLEAS ($P < 0.010$, $\rho = 0.698$) and had the highest correlation coefficient. The BCG was weakly correlated with both the NLFBCI ($P < 0.01$, $\rho = 0.376$) and the GLEAS ($P < 0.01$, $\rho = 0.350$). The NLFBCI, GLEAS and F-HBI/G-HBI all increased in relation to each other as did the BCG and the NLFBCI and the GLEAS. The indices with no significant correlation were the G-HBI and the BCG.

3.3. Site assessments

When using indices to assess sites throughout the watershed, the F-HBI, which has seven classifications, generally indicated 26 of the 30 sites ranked above the good threshold while the remaining four sites were in the fair category. Overall, 86% of sites were scored as BCG Tier 3 sites, which correspond to the narrative of the BCG model that states that there is loss of some rare native taxa and some shifts in relative abundance (Davies and Jackson, 2006). The GLEAS index scored 50% the sites as "excellent" and 50% as "acceptable", the top two of the three tiers of classification in the GLEAS and while the NLFBCI also showed a similar trend in ranking based

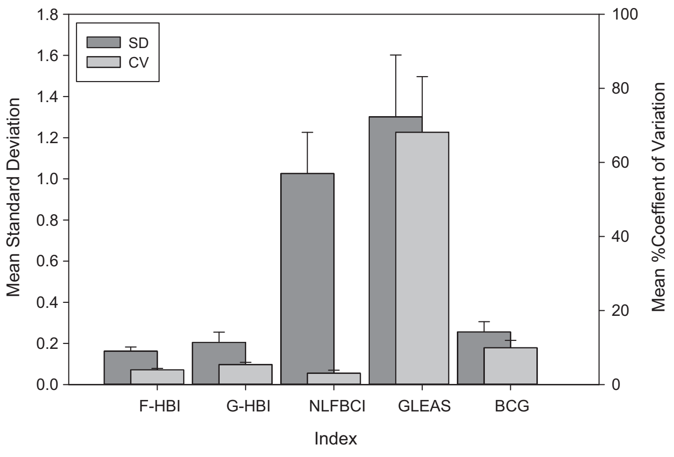


Fig. 3. Mean (+1 S.E.) standard deviation and coefficient of variation among replicate samples (3) collected at nine locations in 2009. Black bars represent the mean standard deviation (SD) of the replicates for each index across the nine sites. The gray bars represent the mean percent coefficient of variation (CV).

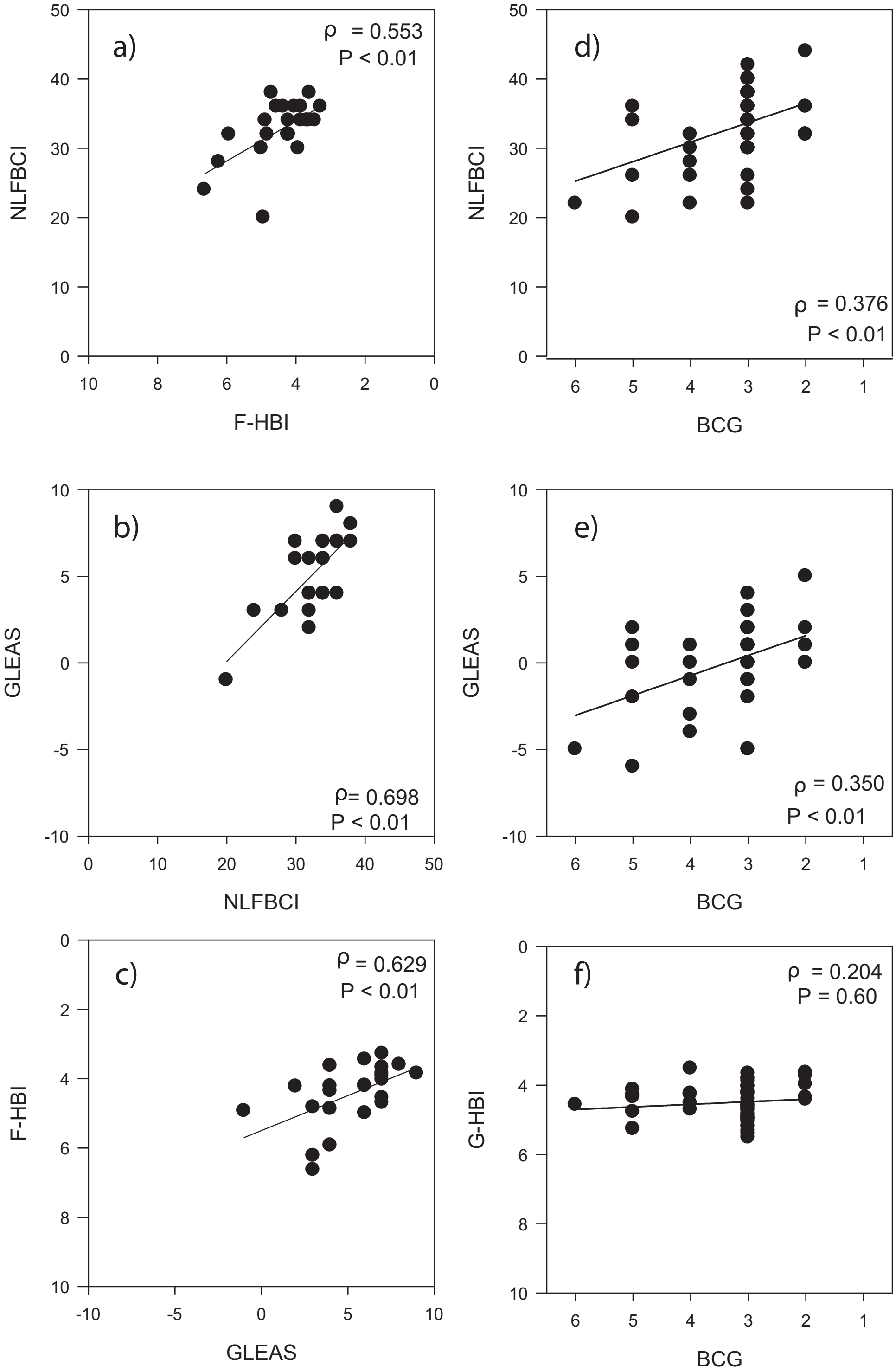


Fig. 4. Spearman's rank order correlation (ρ) between index scores. Figure labels a–c utilized family level data from the 30 sites taken throughout the watershed in 2009. To analyze the BCG genus level data were required and therefor data ($n = 32$) used were from seasonal sampling that occurred at four sites from 2008 to 2011 (figure d–f).

on category (43% in the top and 53% in the fair category); however, discrepancies in scoring appear when actual score values are analyzed (Fig. 5). When evaluated based on a proportional measure away from the good or acceptable threshold, the two indices scored the sites differently ($P < 0.01$, Wilcoxon Signed Rank test). The GLEAS often scored sites higher, and gave a more favorable view of the watershed as a whole, than the NLFBCI which often produced scores very close to the threshold between fair and good.

3.4. Temporal trends

Four sites were assessed with the BCG (genus level resolution) seasonally from 2008 to 2011. Only one site (Sickle) had different scores for fall and spring and the divergence increased through time (Fig. 6c). All other sites were similar to the Oldhouse site (Fig. 6g) and did not show differences in spring and fall BCG scores though statistical analysis was not completed due low sample size. The HBI,

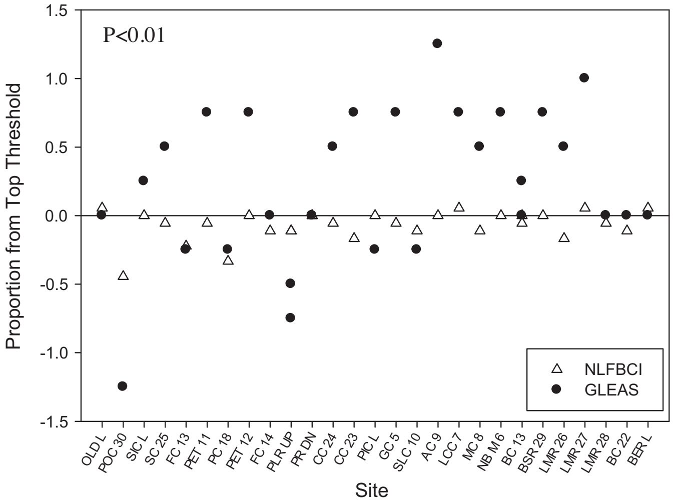


Fig. 5. Representation of the proportion away from the good or acceptable threshold each site scored with the Northern Lakes and Forest benthic community index (NLFBCI) and the Great Lakes Environmental Assessment Section (GLEAS) index. This allows for a comparison of the relative scores for the two community indices and how assessment scores rate the condition of a given site. When paired site scores were analyzed with the Wilcoxon Signed Rank test there was a significant difference ($P < 0.01$). The horizontal line indicates the good or acceptable threshold for the indices.

NLFBCI and GLEAS site assessments from 2002 to 2011 (Fig. 6a, b, d–f, h) indicated that none of the output scores showed significant seasonality ($P > 0.01$, Wilcoxon Signed Rank test). However, when sites were assessed on an annual basis by season they occasionally indicate different categories of classification (Fig. 6). For example, in 2003 the GLEAS produced different categories for the analysis of the sickle site (Fig. 6a). Fig. 6 indicates that the seasonal sampling, though not significantly different in scores, can lead to different categories of classification with the HBI, GLEAS and the NLFBCI. Site show temporal variation (CVs over year and season) in output scores with NLFBCI CVs ranging from 14% (Sickle) to 9% (Oldhouse); GLEAS score CVs from 80% (Sickle) to 61% (Oldhouse); BCG score CVs from 33% (Sickle) to 0% (Oldhouse) and HBI score CVs from 29% (Sickle) to 11% (Oldhouse).

4. Discussion

Important decisions about the management and the use of natural resources are often influenced by the estimated condition of a site (USEPA, 2011), which in recent times tends to be based on metric calculations or an index such as an IBI (Karr and Chu, 1999). Results of this study reveal that estimation of site quality can be influenced by the choice of the index and taxonomic resolution of data. We have shown that assessments of environmental condition are generally concordant among different indices; however, vary in magnitude (fair, good, excellent). Thus, which index is used has management implications, and awareness of biases and strengths of each index improves assessment and interpretation of site scores and results within a regional context.

4.1. Data resolution and source

The nature (e.g., resolution) and the source of the data (i.e., by whom and how collected) that is used to develop site scores can influence our characterization of different systems and our ability to merge data sets for broader spatial and temporal coverage of system assessment. For example, while there is argument for fine resolution in taxonomy for discriminating subtle ecological signals (Waite et al., 2004; Feio et al., 2006; Hawkins, 2006) there is also indication that biotic index scores may not always be sensitive to taxonomic resolution and for some applications more coarse taxonomic resolution (e.g., family) may be acceptable. The F-HBI

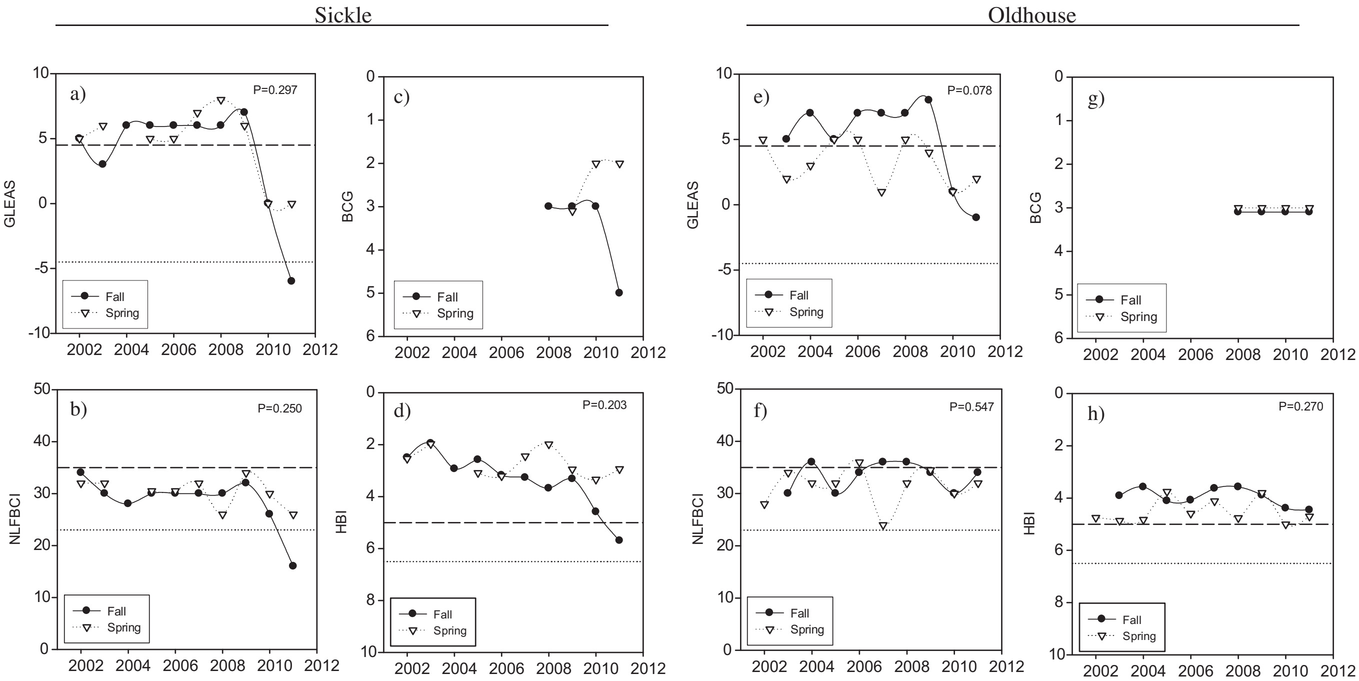


Fig. 6. Spring and fall index scores for two long-term monitoring sites in the Big Manistee River watershed from 2002 to 2011. The Great Lakes environmental and assessment section index (GLEAS) scores (a and e), the Northern Lakes and Forests Ecoregion benthic community index (NLFBCI) (b and f) and the Hilsenhoff biotic index (HBI) (d and h) spring and fall output pairs were analyzed using a Wilcoxon Signed Rank test which did not indicated differences in the seasonal index scores. The biological condition gradient (BCG) data was only available for four years and did not provide enough data to run the statistical analysis (c and g). Only the most variable site (sickle) and the least variable site (Oldhouse) are shown for reference as they represent the extremes of the data. Long dashed horizontal lines in each panel represent the top threshold for each index while the fine dotted line represents the bottom threshold for each index.

has been described as less accurate than the G-HBI (Hilsenhoff, 1988) but there was no significant difference based on family or genus level scoring for the range of values exhibited at the long-term monitoring sites in this study. For both the GLEAS and the NLFBCI the score derived from genus level data was higher than family level data scores. This can be partially explained by the use of individual metrics based on richness values and the inherent increase in richness values as taxonomic resolution increases. Both Bailey et al. (2001) and Chessman et al. (2007) found small differences in sensitivity between family and genus but determined no appreciable information was gained by the added effort of lower taxonomic resolution for bioassessment. In the Big Manistee watershed, where sites are of generally good quality, greater taxonomic resolution is necessary if the goal is to distil small differences in the relatively high quality sites.

Agencies have historically built bioassessment programs to suit specific monitoring and regulatory needs. These long-term monitoring programs provide consistency that is necessary for tracking trends over time (Herbst and Silldorff, 2006). If methods, data, and results were comparable there would be benefit to collaboration and sharing of data for greater regional determination of environmental conditions. In a survey of methods used by state agencies Carter and Resh (2001) found a large range of field and laboratory methods that could limit effective integration of data sets. Though field and laboratory methods between MI-DEQ and LRBOI varied slightly, comparisons between community composition and index scores derived from this data show similar results. Species composition of samples collected by the two agencies was not significantly different at the three replicate sites; however, index score classification derived from this data did vary at one of the three duplicate sites (e.g., LRBOI data scored the site as “excellent” while the MI-DEQ data scored it as “acceptable” using the GLEAS index. Differences in classification levels could be an issue if these indices were utilized for listing sites as impaired. If management recommendations were based on bioassessment, effects of variation in classification could be rectified by conducting multiple assessments for a specific site.

4.2. Index precision

Stream habitat and the associated macroinvertebrate assemblages are spatially variable (Palmer et al., 1997; Lake et al., 2000), yet bioassessment may be based on a single sample or multiple samples from a small area to represent the integrity of a stream reach. In a summary of state agencies that use macroinvertebrates for biomonitoring it was found that 56.1% of programs surveyed (48 States and District of Columbia) conducted replicate sampling for site characterization (Carter and Resh, 2001). Our study found that index scores based on concurrent replicate samples from a site differed in variability depending on index used. The GLEAS index had a much higher average variability (65% CV) and replicates ranged over 5 points for a single site (265% CV) with three replicates, whereas the average CV was below 10% for the NLFBCI, the BCG and the HBI. Mazor et al. (2009) found that average CVs for replicates ranged from 22 to 27% for IBI scores. Herbst and Silldorff (2006) used CVs of 15–20% as their data quality objective for aggregate multimetric IBI scores at reference sites. Nichols et al. (2006) concluded that a single macroinvertebrate collection would be acceptable if the habitat was not variable and was in good condition, but if there was a higher level of habitat heterogeneity then multiple collections were necessary. Depending on the index used there is evidence that replicate samples are necessary for a more accurate assessment of condition. Specifically, with higher variability in scores generated by the GLEAS samples we would advocate using multiple samples for assessments that lead to management decisions if using the GLEAS.

Indices were concordant except when comparing the BCG with the G-HBI. The range of condition determined for our sites was limited in scale for both the G-HBI and the BCG. Also, the HBI was developed to indicate issues from organic pollution and this tolerance-based index may not be comparable to scoring using community composition and comparison to reference conditions. Spearman correlation coefficients for significant relationships among indices (NLFBCI, GLEAS, F-HBI, G-HBI and BCG) were low ($r = 0.35–0.698$) compared with other previous research. Herbst and Silldorff (2006) found moderate correlations (Spearman’s $r = 0.70–0.86$) among indices from sites in the eastern Sierra Nevada of California. Hawkins et al. (2010), also found moderate correlations (Pearson’s $r = 0.63–0.92$) among index scores at sites within the Columbia River basin. When evaluations were completed spanning seven countries throughout Europe, Birk and Hering (2006) found more variable correlations (Pearson’s $r = 0.20–0.077$) among indices. Five streams with 11 sites tested in Australia showed moderate correlations ($r = 0.66–0.89$) between bioindicators (Nichols et al., 2010). Because results from the indices were concordant, sites scored with the NLFBCI, GLEAS, or the BCG will generally reflect similar patterns of biotic condition if tracked over time.

4.3. Site assessments

Based on the region for which they were calibrated, all indices examined in this study were appropriate for use with Big Manistee River watershed dataset. Meador et al. (2008) found in a study of the western US that regional IBIs can work at multiple spatial scales and corroborate those developed at more local geographic scale. Over three geographically separate regions in Oregon and California, models have been developed that contain metrics that function well across ecoregions (Waite et al., 2010). However, locally calibrated indices have also been found to outperform regional indices for site specific assessments (Mykrä et al., 2008; Ode et al., 2008). Overall, the GLEAS (locally calibrated index) and the NLFBCI (regional index) assessment scores provide a favorable view of the watershed where approximately half of the sites were in the top level classification for both indices. However, when assessed using the proportional divergence of the site score away from the highest threshold of condition, the NLFBCI index generally scored sites lower than the GLEAS. This may be an artifact of the taxonomic resolution of the dataset and an indication that genus level resolution is needed for valid assessment using the NLFBCI. Considering the nested nature of the indices, the local calibration of the GLEAS may indicate that, of the Michigan NLF ecoregion, sites in the Big Manistee River watershed sites rank well comparatively. The NLFBCI may give a better overall evaluation of how sites rank in relation to the rest of the ecoregion. Expanding further, the BCG scored six of the seven watershed sites as Tier three and the remaining site as Tier four. This model may give better insight as to the condition of sites relative to a larger regional picture including the state, ecoregion and Upper Midwest. These results exemplify that care must be taken in choosing indices as well as interpreting results from the scores.

4.4. Temporal trends

Over 10 years, paired seasonal index scores were not significantly different; however, on an annual basis there were differences in seasonal scores that could lead to variation in classification of stream sites. We detected no significant trend where one season produced a consistently higher index score. This is in contrast to other studies that found consistent differences in seasonal index scores (Linke et al., 1999; Callanan et al., 2008; Kappes et al., 2010). Others have found multimetric scores to be insensitive to season and showed no differences in scores based on season (Morais et al., 2004; Maloney and Feminella, 2006). Index scores for long-term monitoring sites in the Big Manistee River watershed showed seasonal differences in classification on an annual basis sufficient to alter interpretation of system condition. Ensuring consistency in sampling season is important for accurate assessment and comparisons to reference sites especially if scores are to be used for designating impairment or management action.

Jackson and Fureder (2006), in a review of bioassessment papers (1987–2004), found only 46 papers with long-term (>5 year) data sets. They stressed the need for long-term consistent research to accurately describe the variation, type, magnitude and direction of response signals. Long-term assessments are necessary to begin to understand the natural fluctuations and track anthropogenic effects. There have been conflicting reports in the literature as to the stability of assemblage composition over time. Temporal changes have been observed in Mediterranean systems (Feio et al., 2010) and in pristine Alaskan streams (Milner et al., 2006). In reference sites Nichols et al. (2010) found persistent communities and no significant change in bioindicators over 15 years. Mazor et al. (2009) determined through a twenty-year assessment of four sites that a snapshot approach to bioassessment could lead to incorrect conclusions if natural fluctuations are not taken into account. Huttunen et al. (2012) also found that even with low annual variation there were discrepancies in index scores describing ecological status of sites and that use of one year of data would be problematic for making informed management decisions.

By tracking index scores over multiple years, variation over time is revealed. Conclusions based on a one year assessment would likely be very different than conclusions based on data from 2008 to 2011 (BCG data). This pattern is even more pronounced when specific watershed sites were evaluated over 10 years utilizing the HBI, NLFBCI and the GLEAS. Observing similar patterns in multiple indices through time can distinguish long-term natural variability as compared to anthropogenic effects. Long-term variability has not been well studied in stream systems, though its suggested importance is well-documented (Jackson and Fureder, 2006). Documenting long-term variability will improve assessment of biological quality specifically where disturbance is subtle (Huttunen et al., 2012).

The results found here highlight the benefits and difficulty of utilizing multiple indices developed at different scales with geographically small data sets. Aggregating data from multiple agencies, assessing comparability issues and ensuring index scores are comparable is necessary for an expanded scope of site characterization. The benefit of being able to assess local sites at multiple scales and broadening the scope of assessments leads to a better understanding of ecological condition. The goal of bioassessment is to evaluate the ecological condition of a site, reach, watershed or region. Using multiple lines of evidence in the form of multiple indices will help assess the condition of a site and put it into a larger regional perspective. If indices and thresholds are to be used in management decision making process, replicates or multiple samples over time should be used due to variance in index scores. Long-term assessments are necessary to evaluate site condition and assess natural fluctuation. Each index used in this study was originally developed for different geographic scales and we found that use of the NLFBCI provides an effective assessment of the Big Manistee River watershed.

Acknowledgements

We thank the staff of the Little River Band of Ottawa Indians (LRBOI) Natural Resources Department who collected the field data. Funding for macroinvertebrate collection was provided by the U.S. EPA Clean Water Act Section 106 grant program funding to the Little River Band of Ottawa Indians (I00E56502). Academic, financial, and infrastructural support was provided by Michigan Technological University Department of Biological Sciences. We thank Bob Newport (U.S. EPA) for the years of discussion on comparative bioassessment that helped to formulate the basis for this work, Allison Smart (LRBOI) for GIS mapping and Marty Holtgren (LRBOI) for manuscript review and assistance with graphics.

References

Bailey, R., Norris, R.H., Reynoldson, T.B., 2001. Taxonomic resolution of benthic macroinvertebrate communities in bioassessments. J. N. Am. Benthol. Soc. 20, 280–286.

Barbour, M.T., Gerritsen, J., Snyder, B.D., Stribling, J.B., 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, 2nd ed. U.S. Environmental Protection Agency, Office of Water, EPA Report 841-B-99-002, Washington, DC.

Birk, S., Hering, D., 2006. Direct comparison of assessment methods using benthic macroinvertebrates: a contribution to the EU Water Framework Directive intercalibration exercise. Hydrobiologia 566, 401–415.

Butcher, J.T., Stewart, P.M., Simon, T.P., 2003. A benthic community index for streams in the Northern Lakes and Forests Ecoregion. Ecol. Indic. 3, 181–193.

Cao, Y., Hawkins, C.P., Storey, A.W., 2005. A method for measuring the comparability of different sampling methods used in biological surveys: implications for data integration and synthesis. Fresh. Biol. 50, 1105–1115.

Cao, Y., Hawkins, C.P., 2011. The comparability of bioassessments: a review of conceptual and methodological issues. J. N. Am. Benthol. Soc. 30, 680–701.

Callanan, M., Baars, J.R., Kelly-Quinn, M., 2008. Critical influence of seasonal sampling on the ecological quality assessment of small headwater streams. Hydrobiologia 610, 245–255.

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. J. N. Am. Benthol. Soc. 20, 658–682.

Creel, W.S., Hanshue, S., Kosek, S., Oemke, M., Walterhouse, M., 1998. GLEAS Procedure Metric Scoring and Interpretation. In: Schneider, J.C. (Ed.), Manual of Fisheries Survey Methods II: With Periodic Updates. Michigan Department of Natural Resources, Fisheries Special Report 25, Ann Arbor, MI (Chapter 25).

Chessman, B., Williams, S., Besley, C., 2007. Bioassessment of streams with macroinvertebrates: effect of sampled habitat and taxonomic resolution. J. N. Am. Benthol. Soc. 26, 546–565.

Chirhart, J., 1998. Invertebrate sampling procedures for Northern Lakes and Forests streams. In: Simon, T.P., Stewart, P.M. (Eds.), Standard Operating Procedures for Development of Watershed Indicators in REMAP: Northern Lakes and Forests Streams. U.S. Environmental Protection Agency, Region 5 Water Division, Watershed and Non-Point Source Branch, Chicago, IL.

Davies, S.P., Jackson, S.K., 2006. The biological condition gradient: a descriptive model for interpreting change in aquatic ecosystems. Ecol. Appl. 16, 1251–1266.

Feio, M.J., Reynoldson, T.B., Craca, M.A.S., 2006. The influence of taxonomic level on the performance of a predictive model for water quality assessment. Can. J. Fish. Aquat. Sci. 63, 367–376.

Feio, M.J., Coimbra, C.N., Craca, M.A.S., Nichols, S.J., Norris, R.H., 2010. The influence of extreme climatic events and human disturbance on macroinvertebrate community patterns of a Mediterranean stream over 15 years. J. N. Am. Benthol. Soc. 29, 1397–1409.

Gerth, W.J., Herlihy, A.T., 2006. Effect of sampling different habitat types in regional macroinvertebrate bioassessment surveys. J. N. Am. Benthol. Soc. 25, 501–512.

Gerritesn, J., Stamp, J., 2012. Calibration of the Biological Condition Gradient (BCG) in Cold and Cool Waters of the Upper Midwest: BCG-based Indexes (BCG-I) for Fish and Benthic Macroinvertebrate Assesmblages. Tetra Tech, Inc., Owings Mills, MD.

Hawkins, C.P., Cao, Y., Roper, R., 2010. Method of predicting reference conditions affects the performance and interpretation of ecological indices. Fresh. Biol. 55, 1066–1085.

Hawkins, C.P., 2006. Maintaining and restoring the ecological integrity of freshwater ecosystems: refining biological assessments. Ecol. Appl. 16, 1249–1311.

Herbst, D.B., Silldorff, E.L., 2006. Comparison of the performance of different bioassessment methods: similar evaluations of biotic integrity from separate programs and procedures. J. N. Am. Benthol. Soc. 25, 513–530.

Hilsenhoff, W.L., 1987. An improved biotic index of organic stream pollution. Great Lakes Entomol. 20, 31–39.

Hilsenhoff, W.L., 1988. A modification of the biotic index of organic stream pollution to remedy problems and permit its use throughout the year. Great Lakes Entomol. 31, 1–12.

Hughes, R.M., Peck, D.V., 2008. Acquiring data for large aquatic resource surveys: the art of compromise among science, logistics and reality. J. N. Am. Benthol. Soc. 27, 837–859.

Huttunen, K.L., Mykrä, H., Muotka, T., 2012. Temporal variability in taxonomic completeness of stream macroinvertebrate assemblages. Fresh. Sci. 31, 423–441.

Jackson, J.K., Fureder, L., 2006. Long-term studies of freshwater macroinvertebrates: a review of the frequency, duration and ecological significance. Fresh. Biol. 51, 591–603.

Kappes, H., Sundermann, A., Haase, P., 2010. High spatial variability biases the space-for-time approach in environmental monitoring. Ecol. Indic. 10, 1202–1205.

Karr, J.R., Chu, E.W., 1999. Restoring Life in Running Waters: Better Biological Monitoring. Island Press, Washington, DC.

Lake, P.S., Palmer, M.A., Biro, P., Cole, J., Covich, A.P., Dahm, C., Gibert, J., Goedkoop, W., Martens, K., Verhoeven, J., 2000. Global change and the biodiversity of freshwater ecosystems: impacts on linkages between above-sediment and sediment biota. BioScience 50, 1099–1107.

Linke, S., Bailey, R.C., Schwindt, J., 1999. Temporal variability of stream bioassessment using benthic macroinvertebrates. Fresh. Biol. 42, 575–584.

Lipsey, T., 2010. MI-DEQ Staff Report. Biological and Water Chemistry Surveys of Selected Stations in the Manistee, Little Manistee, and Big Sable river watersheds Grand Traverse, Kalkaska, Lake, Manistee, Mason, Osceola, and Wexford Counties, Michigan. MI/DNRE/WB-10/016.

Maloney, K.O., Feminella, J.W., 2006. Evaluation of single- and multi-metric benthic macroinvertebrate indicators of catchment disturbance over time at the Fort Benning Military Installation, Georgia, USA. Ecol. Indic. 63, 469–484.

Mantel, N.A., 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27, 209–220.

Mazor, R.D., Schiff, K., Ritter, K., Rehn, A., Ode, P., 2010. Bioassessment tools in novel habitats: an evaluation of indices and sampling methods in low-gradient streams in California. Environ. Monit. Assess. 167, 91–104.

Mazor, R.D., Purcell, A.H., Resh, V.H., 2009. Long-term variability in bioassessments: a twenty-year study from two northern California streams. Environ. Manage. 43, 1269–1286.

McCune, B., Grace, J.B., 2002. Analysis of Ecological Communities. Mjm Software Design, Gleneden Beach, OR, US.

McCune, B., Mefford, M.J., 2006. PC-ORD: Multivariate Analysis of Ecological Data. MJM Software. Version 5.14, Gleneden Beach, Oregon.

Meador, M.R., Whittier, T.R., Goldstein, R.M., 2008. Evaluation of an index of biotic integrity approach used to assess biological condition in western US streams and rivers at varying spatial scales. Trans. Am. Fish. Soc. 137, 13–22.

Milner, A.M., Conn, S.C., Brown, L.E., 2006. Persistence and stability of macroinvertebrate communities in streams of Denali National Park, Alaska: implications for biological monitoring. Fresh. Biol. 51, 373–387.

Morais, M., Pinto, P., Guilherme, P., Rosado, J., Antunes, I., 2004. Assessment of temporary streams: the robustness of metric and multimetric indices under different hydrological conditions. Hydrobiologia 516, 229–249.

Mykrä, H., Aroviita, J., Kotanen, J., Hämäläinen, H., Muotka, T., 2008. Predicting the stream macroinvertebrate fauna across regional scales: influence of geographical extent on model performance. J. N. Am. Benthol. Soc. 27, 705–716.

Nichols, S.J., Robinson, W.A., Norris, R.H., 2006. Sample variability influences on the precision of predictive bioassessment. Hydrobiologia 572, 215–233.

Nichols, S.J., Robinson, W.A., Norris, R.H., 2010. Using the reference condition maintains the integrity of a bioassessment program in a changing climate. J. N. Am. Benthol. Soc. 29, 1459–1471.

NLcd, 2006. National Oceanic and Atmospheric Administration, Coastal Services Center. 1995-Present. The Coastal Change Analysis Program (C-CAP) Regional Land Cover. NOAA Coastal Services Center, Charleston, SC, Accessed at: www.csc.noaa.gov/digitalcoast/data/ccapregional

Omernik, J.M., 1987. Ecoregions of the conterminous United States map (scale 1:7,500,000). Ann. Assoc. Am. Geogr. 77, 118–125.

Ode, P.R., Hawkins, C.P., Mazor, R.D., 2008. Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. J. N. Am. Benthol. Soc. 27, 967–985.

Palmer, M.A., Covich, A.P., Findlay, B.J., Gilbert, J., Hyde, K.D., Johnson, R.K., Kairesalo, T., Lake, S., Lovell, C.R., Naiman, R.J., Ricci, C., Sabater, F., 1997. Biodiversity and ecosystem processes in freshwater sediments. AMBIO 26, 571–577.

Rehn, A.C., Ode, P.R., Hawkins, C.P., 2007. Comparisons of targeted-riffle and reach-wide benthic macroinvertebrate samples: implications for data sharing in stream-condition assessments. J. N. Am. Benthol. Soc. 26, 332–348.

Systat Software Inc., 2012. Systat Software SigmaPlot Users Guide 12.2. San Jose, CA.

USEPA (US Environmental Protection Agency), 2011. A Primer on Using Biological Assessments to Support Water Quality Management, EPA-810-R-11-01. Office of Science and Technology, Office of Water, Washington, DC.

Waite, I.R., Herlihy, A.T., Larsen, D.P., Urquhart, N.S., Klemm, D.J., 2004. The effects of macroinvertebrate taxonomic resolution in large landscape bioassessments: an example from the Mid-Atlantic Highlands, USA. Fresh. Biol. 49, 474–489.

Waite, I.R., Brown, L.R., Kennen, J.G., May, J.T., Cuffney, T.F., Orlando, J.L., Jones, K.A., 2010. Comparison of watershed disturbance predictive models for stream benthic macroinvertebrates for three distinct ecoregions in western US. Ecol. Indic. 10, 1125–1136.