Development of the New Hampshire Benthic Index of Biotic Integrity

Prepared by Karen Blocksom

U.S. Environmental Protection Agency

Ecological Exposure Research Division

National Exposure Research Laboratory

26 W. Martin Luther King Drive

Cincinnati, OH 45268

The New Hampshire Department of Environmental Services (NHDES) Biomonitoring Program has been assessing the biological health of aquatic ecosystems throughout the state since 1995, focusing on wadeable streams. These assessments are used to establish reference locations for "least disturbed" conditions in the state and to identify areas that are biologically impaired. Eventually, such information will aide in prioritizing those areas needing management, restoration, or preservation efforts.   The biological data collected from streams have been used primarily in a qualitative manner to make assessments for the 305(b) report on the condition of aquatic resources and the 303(d) list of impaired waters. However, the U.S. Environmental Protection Agency (USEPA) is encouraging states to implement plans for developing numeric biological standards. The NHDES Biomonitoring Program is working towards calibrated metrics that will be incorporated into the State of New Hampshire Surface Water Quality Regulations, supporting the current narrative biological standards. In an effort to reach that goal, USEPA's National Exposure Research Laboratory (NERL) in Cincinnati, Ohio, has assisted the NHDES Biomonitoring staff in developing a draft Benthic Index of Biotic Integrity (B-IBI). In this report, we describe in detail the stepwise process (Figure 1) used to develop a macroinvertebrate multimetric index used to assess the biotic integrity of New Hampshire's wadeable streams.

Data Set

The sites used in the development process were those sampled between 1997 and 2002 from wadeable streams across New Hampshire. This restriction of data was set because from 1997 to the present, rock baskets have been used to collect macroinvertebrates, but prior to 1997, macroinvertebrates were collected using a kick

net method. Macroinvertebrate data were randomly divided into a calibration (125 sites) and a validation (31 sites) data set. For most steps in index development, only the calibration data were used, although the two datasets were combined to provide sufficient data for some analyses.

Development of Stressor Gradient

An index to measure the types and level of human disturbance at each site was developed by NHDES personnel. All sites were combined for this step. The factors included in the index were based on GIS coverages of land use and point sources, as well as a qualitative habitat index that was applied during macroinvertebrate sample collection. At the watershed scale, percent land cover as developed (residential, commercial, and industrial) and percent agriculture were calculated from the New Hampshire Land Cover (NHLC) data. The percent of water impounded was also calculated at the watershed scale. Local scale variables were calculated for a 300-foot buffer on either side of all hydrologic features within the watershed AND within a 1-mile radius of the site. The variables calculated at the local scale were densities of Ground Water Hazard Inventory (GWHI) sites, Resource Conservation and Recovery Act (RCRA) sites, junkyards, dams, water withdrawals, National Pollutant Discharge Elimination System (NPDES) sites, and roads. Finally, the qualitative habitat score was considered to be an immediate in-stream measure of human influence. Each variable was scored on a 0 to 3 scale, with 0 signifying minimal disturbance for that variable, based upon overall distributions across all sites sampled between 1997 and 2001 (Table 1). Scores were summed for watershed and local scale disturbance scores, as well as across all variables for a total disturbance score, the human disturbance

gradient (HDG) score. The total score was divided into four disturbance levels with equal scoring ranges: 0-4 (Best), 5-9 (Good), 10-14 (Fair), and >14 (Worst). The entire range observed across all sites was 0 to 19.

Identification Reference and Impaired Sites

Sites with disturbance scores falling into the "Best" range (0-4) were considered representative of reference conditions. Forty-six reference sites were identified based on this criterion across the state. There were nine validation reference sites, and 37 calibration reference sites.

Stream Classification

Methods

After reference sites were identified, we were able to evaluate various classification schemes for New Hampshire streams. We used only reference sites in this evaluation to limit the variability to that due to classes of streams, and calibration and validation reference sites were combined. Using data at the lowest taxonomic level, we used only taxa with more than 2 occurrences in the reference data set. We performed nonmetric multi-dimensional scaling (NMDS) using Sorensen distances between sites based on taxonomic presence-absence data because of inconsistency in combining rock baskets, affecting abundance data in unknown ways. The NMDS was performed in PC-ORD 4.0 (MjM Software, Gleneden Beach, Oregon) using the autopilot mode, which tests solutions for 1 through 6 axes and chooses the solution with the lowest stress value.

Results

A two-axis solution was recommended these data and had a final stress of 20.883. Axis 1 accounted for about 32% of the variability and axis 2 accounted for approximately 53% of the variability. Classification schemes based on U.S. Forest Service bioregions, basins, stream order, Omernik level III ecoregions, and regions developed by The Nature Conservancy (TNC) were evaluated by examining ordination plots coded by class. The TNC regions provided the most promising classification scheme, separating into two groups along the second NMDS axis (Figure 2). The four TNC regions were grouped into Northern (regions 3 and 4) and Southern (regions 1 and 2) regions. The second axis was correlated with both latitude and elevation, supporting grouping into Northern and Southern regions (Figure 3).

Metric Evaluation

Metric Calculation

Metrics representing taxa richness, taxonomic composition, pollution tolerance, functional feeding groups, and behavioral habits were calculated (Table 2). Because of varied numbers of organisms identified, all richness metrics were rarefacted to a 100-organism subsample size (Hurlbert, 1971). To prepare the data for metric calculations, taxonomic resolution was adjusted to a consistent level for each taxonomic group across all sites. This taxonomic level was genus for most groups but family level for Chironomidae. In addition, the taxa in each sample were marked for exclusion from richness metric calculations if any individuals in that taxonomic group were identified to a more specific taxonomic level. For example, if an individual was only identified as Plecoptera and other Plecoptera individuals in that sample were identified to the genus

level, the order-level taxon Plecoptera was excluded from richness calculations while genus-level Plecoptera taxa were included.

Adjustment for natural factors

<u>Methods</u>

Each metric was evaluated for its relationship with three variables presumed to have a possible effect on values under reference conditions: watershed size, elevation, and gradient (slope). We examined plots of each metric against each variable, paying particular attention to the upper boundary for positive metrics (those that have larger values under better conditions), the lower boundary for negative metrics (those having smaller values under better conditions), and both boundaries for variable metrics. To avoid making adjustments for non-natural factors, we used only reference sites (across validation and calibration data sets) in plots. In addition, we combined bioregions and included both calibration and validation data sets to increase sample size and our ability to detect relationships. For metrics having an apparent relationship with a variable, we calculated a 95th percentile regression for trends with the upper boundary and a 5th percentile regression for trends with the lower boundary. These regressions were based on methods described in Blackburn et al. (1992). If the regression was significant ($p < 0.05$), the adjusted metric value for each site was calculated as the predicted metric value based on the value of the natural factor for that site subtracted from the actual observed metric value at that site. If more than one natural factor showed a relationship to a given metric, a multiple regression was performed in the same manner described above.

<u>Results</u>

Four metrics were related to either elevation or watershed size or both. No metrics showed a relationship with stream gradient. The Ephemeroptera taxa and intolerant taxa metrics were related to elevation, percent collector-filterers was related to watershed area, and percent Trichoptera was related to both watershed size and elevation. Equations for adjusted values for each metric are provided in Table 3. These adjusted metric values were used in all further analyses except the range test.

Range test

The range of each metric was evaluated based on the calibration data and for both regions combined. Richness metrics with a range of less than 4 (maximum – minimum value) and percentage metrics with a range of less than 10% were eliminated from further consideration. Burrower taxa richness, swimmer taxa richness, climber taxa richness, and percent climbers were all eliminated due to insufficient range.

Relationships with stressors

Methods

Each metric was evaluated for its relationship to potential stressor variables in two ways. First, Spearman rank correlations were calculated between metrics and raw and scored HDG component variables, as well as the local, watershed, and overall HDG scores. Correlations with a p-value of 0.001 or less were considered significant. The second evaluation involved the creation of box plots of metric values for reference and impaired sites (HDG scores of 15 or more). Each plot was scored on a scale of 0 (complete overlap of boxes) to 3 (no overlap) using the system described in Barbour et al. (1996). Metrics with scores of 2 or 3 for all sites and within at least the Southern region were considered adequately and consistently responsive. Correlation analysis

was performed on all calibration sites combined, but box plots were created for all calibration data and for the Northern and Southern regions separately.

<u>Results</u>

There were 17 metrics correlated with HDG scores and 15 correlated with raw HDG component values (Table 4). Metrics that were correlated with a particular score were often also correlated with the raw variable, but there were a few cases where a metric was only correlated with one of the two types of variables. The metrics correlated with multiple abiotic variables tended to be those showing responsiveness in box plots (Table 5). However, the range of abiotic conditions was much smaller in Northern sites than in Southern sites, limiting the usefulness of box plots for Northern sites. Only metrics with at least one correlation or one responsive box plot were considered further for inclusion in an index (Table 6).

Redundancy

Metrics still being considered for an index were evaluated for redundancy with one another using Pearson correlations. Pearson correlations with a magnitude of 0.8 or larger indicated that two metrics provided similar enough information to include only one of the pair in an index. Only eight of the remaining metrics showed redundancy (Table 7).

Index Assembly and Evaluation

Metric Selection

Twelve alternative sets of metrics were selected from the remaining metrics (Table 7). Pairs of redundant metrics were not both included in any single set of metrics. Sets of metrics were chosen to reflect as many assemblage attributes as

8

possible, and the robustness of relationships with abiotic variables was considered in selecting metrics.

Scoring of metrics

Two scoring schemes were evaluated for each set of metrics. In one method, termed the *reference site* method, the 75th percentile of reference sites (regions combined) was used to set an upper threshold for positive metrics, and the 25th percentile of reference was used to set a threshold for negative metrics. The second method, termed the *all sites* method, used the 95th percentile across all sites as a threshold for positive metrics and the 5th percentile across all sites for negative metrics. Scoring for each metric was on a continuous scale from 0 to 100. For positive metrics, scores were calculated as:

$$\text{Score} = \frac{observed}{threshold} * 100$$

For negative metrics, scores were calculated as:

$$\text{Score} = \frac{(\max - observed)}{(\max - threshold)} * 100$$

For each of the two remaining metrics adjusted for watershed size or elevation, a constant was added to the equation for residuals before thresholds were calculated. Addition of the constant simply allowed all values to remain positive but did not change other properties of the metric. Scoring thresholds for each metric are provided in Table 8. All scores were truncated to a range of 0 to 100. The index score was calculated as the average of its component metric scores.

Evaluation of alternative indices

Methods

9

We used both the calibration and validation data sets to test alternative indices. For each data set, we calculated the discrimination efficiency (DE) of each alternative index as the proportion of sites with HDG scores greater than 14 (N = 9 for calibration, N = 3 for validation) or greater than 9 (N=35 for calibration, N=13 for validation) having index scores below the 25th percentile of reference sites. We also compared the standard deviation of scores across reference sites for each alternative index. We then calculated the Spearman rank correlation of each index with the HDG total score. Finally, we calculated correlations for a subset of data between index alternatives and pH, dissolved oxygen (D.O.), and conductivity.

Results

Using calibration data, we found that DEs were similar between the two scoring schemes but varied among the index alternatives (Table 9). However, the standard deviation of scores among reference sites was consistently higher for the *reference site* method of scoring than for *all sites* scoring. Spearman correlations with the HDG scores were similar between scoring methods, but did vary slightly among alternative indices. Using validation data, the two scoring schemes were again quite similar, although the standard deviation of reference was slightly lower for the *reference site* method. We selected the *all sites* scoring method because the variability associated with this method was generally lower and was lower with a larger data set.

Alternative index 11 performed the best overall with the calibration data, having the highest DE when the worst two tiers of HDG scores were considered and one of the highest correlations with HDG scores. This index also had the smallest standard deviation among reference for calibration sites. Using the validation data, alternative 11

had one of the higher standard deviations for *all sites* scoring and a moderate value for *reference site* scoring (Table 10). For validation sites, alternative index 6 performed well in all respects. Alternative 11 also had the highest correlation with D.O. and among the highest correlations with conductivity and pH (Table 11).

Laboratory methods comparison

During 1997 and 1998, NHDES used an Imhoff cone for processing samples in the laboratory using the Maine DEP subsampling method (NHDES, 2002). Beginning in 1999, samples were processing using a Caton subsampler (NHDES, 2002). Because there was concern that these two methods might result in differences in the size and type of organisms that tended to be sorted, we compared index alternatives 6 and 11 between the two methods using calibration and validation data sets combined. In the Northern region, there were only three observations overall that were based on the Caton method, so comparisons were limited to Southern sites, where the distribution between the two methods was more even (Table 12).

There were obvious differences between the two laboratory methods, with Imhoff samples tending to achieve higher scores than Caton samples. These differences were most pronounced for the Best and Worst HDG groups. However, there was greater overlap between the two methods for alternative index 11 than for alternative index 6. These results indicate that, although there are differences between methods, using all of the data to set metric thresholds will provide a more conservative estimate of condition because the current method will tend to score lower. This means that index scores based on the Caton method may tend to indicate that sites are in poorer condition than they actually are. However, until sufficient data can be collected using

the Caton laboratory method, the thresholds based on all data should be used to ensure a sufficient level of protection.

Selection of index and setting biocriteria

Based on the results of the evaluation of index alternatives, as well as the comparison of laboratory sampling methods, we selected index alternative 11 using the *all sites* scoring method. The scoring for the seven metrics in this index, the Benthic Index of Biotic Integrity (B-IBI) for New Hampshire streams, is provided in Table 13. We used the reference distribution within each bioregion to set biocriteria. Because reference sites identification was based on imperfect information about each site, we did not regard reference sites as a true representation of reference conditions. Therefore, we used the 25th percentile of the reference distribution in each region as the threshold for attainment of aquatic life use standards. For the Northern region, this threshold was 77.0, and for the Southern region, it was 66.4.

Future work

The B-IBI thresholds are meant to be temporary and would likely change once enough Caton subsampling data are available. Under the current thresholds, more sites will likely be assessed as non-attaining than would be the case if metric scoring thresholds were based solely on Caton data. In addition, it is possible that some metrics may no longer appear responsive to stressors when only Caton subsampling data are used. Thus, at least a partial re-analysis of metrics is required in order to incorporate additional Caton data and replace Imhoff cone data into the development process.

Currently, no repeat visits to sites are available to estimate temporal variability associated with the B-IBI. However, this is an important index feature which should be evaluated in the future. A specific effort to visit a random subset of sites multiple times within a year or over multiple years is necessary to address this issue.

Literature Cited

Barbour, M.T., J. Gerritsen, G.E. Griffith, R. Frydenborg, E. McCarron, J.S. White, and M.L. Bastian. A framework for biological criteria for Florida streams using benthic macroinvertebrates. Journal of the North American Benthological Society 15:185-211.

Blackburn, T.M., J.H. Lawton, and J.N. Perry. 1992. A method of estimating the slope of upper bounds of plots of body size and abundance in natural animal assemblages. *OIKOS* 65:107-112.

Hurlbert, S.H. 1971. The nonconcept of species diversity: A critique and alternative parameters. *Ecology* 52:577-586.

New Hampshire Department of Environmental Services (NHDES). 2002. *Biomonitoring Program Protocols*. New Hampshire Department of Environmental Services, Concord, New Hampshire.

Table 1. Human influence gradient scoring scheme.

Abbreviation	Variable	Scores			
		0	1	2	3
LC_DEV	% Developed land (watershed)	<2	<4	<8	≥8
LC_AG	% Agriculture (watershed)	<2	<4	<8	≥8
IMPD	% Impounded (watershed)	0	<0.02	<0.05	≥0.05
GWHI	GWHI sites (#/mi^2)	0	<2	<5	≥5
RCRA	RCRA sites (#/mi^2)	0	<2	<5	≥5
JKYD	Junkyards (#/mi^2)	0	<1	<3	≥3
DAMS	Dams (#/mi^2)	0	<1	<2	≥2
WATUSE	Water withdrawals (#/mi^2)	0	<1	<3	≥3
NPDES	NPDES sites (#/mi^2)	0	<1	<2	≥2
RDDEN	Road density (units?)	<1	<2	<3	≥3
HABT	Habitat score				

Table 2. Macroinvertebrate metrics calculated for evaluation.

Metric	Expected response to disturbance
Total taxa	Decrease
EPT (Ephemeroptera + Plecoptera + Trichoptera) taxa	Decrease
Ephemeroptera taxa	Decrease
Plecoptera taxa	Decrease
Trichoptera taxa	Decrease
% EPT	Decrease
% Ephemeroptera	Decrease
% Plecoptera	Decrease
% Trichoptera	Decrease
% EPT/(EPT + Chironomidae)	Decrease
% Chironomidae	Increase
% Non-insects	Increase
% Dominant taxon	Increase
Collector-gatherer taxa richness	Variable
Collector-filterer taxa richness	Variable
Scraper taxa richness	Variable
Shredder taxa richness	Variable
% Collector-gatherers	Variable
% Collector-filterers	Variable
% Scrapers	Decrease

Metric	Expected response to disturbance
% Shredders	Variable
% Predators	Variable
% Scrapers/(Scrapers + Collector-filterers)	Decrease
Clinger taxa richness	Decrease
Burrower taxa richness	Variable
Swimmer taxa richness	Variable
Sprawler taxa richness	Variable
Climber taxa richness	Variable
% Clingers	Decrease
% Burrowers	Variable
% Swimmers	Variable
% Sprawlers	Variable
% Climbers	Variable
Intolerant taxa richness	Decrease
Tolerant taxa richness	Increase
% Intolerant	Decrease
% Tolerant	Increase
Hilsenhoff Biotic Index (HBI)	Increase

Table 3. Equations for adjustment of metrics related to elevation and watershed area.

Metric	Adjusted value	Adjusted R^2
Ephemeroptera taxa	Observed value – (4.9 + 0.001*elevation)	0.91
Intolerant taxa	Obs. value – (8.9 + 0.0034*elevation)	0.67
% Collector-filterers	Obs. value – (26.3 + 28.4*log10(watershed size))	0.68
% Trichoptera	Obs. value – (11.6 + 20.6*log10(watershed size) + 0.02*elevation)	0.81

Table 5. Box plots scores based on Barbour et al. (1996) for all calibration sites combined and for each region separately. A score of (0) indicates total overlap of boxes, (1) overlap of one median, (2) overlap of boxes but not medians, (3) no overlap of boxes.

Metric	All sites	Northern region	Southern region
Total taxa*	3	1	3
EPT taxa*	3	1	3
Adj. Ephemeroptera taxa	1	0	1
Plecoptera taxa*	3	0	3
Trichoptera taxa	1	0	0
% EPT	2	0	1
% Ephemeroptera*	2	2	2
% Plecoptera	2	0	1
Adj. % Trichoptera	0	0	0
% EPT/(EPT + Chironomidae)	0	1	1
% Chironomidae	2	0	1
% Non-insects	1	1	0
% Dominant taxon	0	1	0
Collector-gatherer taxa richness	1	0	1
Collector-filterer taxa richness	0	0	1
Scraper taxa richness	0	0	0
Shredder taxa richness	2	1	2

Metric	All sites	Northern region	Southern region
% Collector-gatherers	1	1	1
Adj. % Collector-filterers	0	0	0
% Scrapers	0	0	0
% Shredders	1	1	1
% Predators	3	1	1
% Scrapers/(Scrapers + Collector-filterers)	0	0	0
Clinger taxa richness*	3	1	3
Sprawler taxa richness	0	1	0
% Clingers	1	0	1
% Burrowers	1	0	1
% Swimmers	2	0	1
% Sprawlers	0	1	0
Adj. Intolerant taxa richness*	3	0	2
Tolerant taxa richness	1	0	1
% Intolerant*	3	1	3
% Tolerant	0	0	0
HBI*	3	2	3

*Metric considered responsive because score of 2 or 3 overall and within Southern region.

Table 6. Summary of relationships of metrics with HDG variables and abiotic condition.

Metric	Correlations – HDG scores	Correlations – HDG raw	Box plots
Total taxa*	X	X	X
EPT taxa*	X	X	X
Adj. Ephemeroptera taxa			
Plecoptera taxa*	X	X	X
Trichoptera taxa			
% EPT*	X	X	
% Ephemeroptera*	X	X	X
% Plecoptera			
Adj. % Trichoptera			
% EPT/(EPT + Chironomidae)			
% Chironomidae*	X		
% Non-insects*	X	X	
% Dominant taxon			
Collector-gatherer taxa richness			
Collector-filterer taxa richness*	X		
Scraper taxa richness			
Shredder taxa richness			

Metric	Correlations – HDG scores	Correlations – HDG raw	Box plots
% Collector-gatherers			
Adj. % Collector-filterers*		X	
% Scrapers			
% Shredders			
% Predators*	X	X	
% Scrapers/(Scrapers + Collector-filterers)			
Clinger taxa richness*	X	X	X
Sprawler taxa richness			
% Clingers*	X	X	
% Burrowers			
% Swimmers*	X	X	
% Sprawlers			
Adj. Intolerant taxa richness*	X	X	X
Tolerant taxa richness*	X	X	
% Intolerant*	X	X	X
% Tolerant*	X		
HBI*	X	X	X

*Metrics considered for inclusion in index

Table 7. Redundancy among metrics and alternative sets of metrics tested.

Metric	Redundant metrics	Alternate indices											
		1	2	3	4	5	6	7	8	9	10	11	12
Total taxa	EPT taxa, Clinger taxa	X			X		X				X	X	X
EPT taxa	Total taxa, Clinger taxa		X		X		X	X	X				
Plecoptera taxa		X	X	X	X	X	X	X	X	X	X	X	X
% EPT	% Clingers, % Chironomidae	X	X	X									
% Ephemeroptera					X	X	X	X	X	X	X		X
% Chironomidae	% EPT				X	X	X				X	X	X
% Non-insects					X	X	X					X	X
Coll.-filterer taxa		X	X	X	X	X		X	X	X			
Adj. % coll.-filterer					X	X	X	X	X				
% Predators		X	X					X		X			
Clinger taxa	Total taxa, EPT taxa			X									
% Clingers	% EPT				X	X	X	X	X	X	X	X	X
% Swimmers		X	X	X									

Metric	Redundant metrics	Alternate indices											
		1	2	3	4	5	6	7	8	9	10	11	12
Adj. intol. taxa		X		X					X				
Tolerant taxa					X	X	X			X	X	X	X
% Intolerant	HBI				X	X	X				X	X	X
% Tolerant													
HBI	% Intolerant	X	X	X				X	X	X			

Table 8. Scoring thresholds for each metric used in alternate indices.

Metric	95th (* = 5th) percentile of all sites	75th (* = 25th) percentile of reference sites
Total taxa	21.5	16.9
EPT taxa	14.6	13.2
Plecoptera taxa	4.4	3.9
% EPT	96.7	92.7
% Ephemeroptera	68.1	56.0
% Chironomidae	*0.0	*1.1
% Non-insects	*0.0	*0.0
Collector-filterer taxa	5.3	3.7
Adj. % coll.-filterers (+ constant of 100)	99.9	79.4
% Predators	26.5	13.8
Clinger taxa	14.3	12.0
% Clingers	94.6	85.1
% Swimmers	37.8	12.0
Adj. intolerant taxa (+ constant of 11)	18.7	17.6
Tolerant taxa	*0	*0.25
% Intolerant	76.1	62.4
HBI	*2.0	*2.4

Table 9. Results of comparison of alternative indices and scoring methods using calibration data. Standard deviation is based on 37 reference sites, and correlations are based on 125 observations.

Alternative index	25th percentile	DE, HDG scores >14 (N=9)	DE, HDG scores >9 (N=35)	Standard deviation of reference	Spearman r with HDG score
All sites 1	55.6	0.78	0.66	12.4	-0.47
2	55.3	0.78	0.63	12.4	-0.43
3	63.3	0.78	0.69	11.5	-0.49
4	63.9	0.78	0.63	10.6	-0.48
5	65.1	0.89	0.66	11.0	-0.49
6	62.4	1.00	0.74	11.5	-0.55
7	58.0	0.89	0.63	11.9	-0.45
8	61.4	0.89	0.66	12.5	-0.50
9	59.7	0.89	0.63	11.5	-0.48
10	62.1	1.00	0.74	13.0	-0.55
11	70.9	1.00	0.80	10.4	-0.56
12	66.8	1.00	0.77	11.4	-0.54
Reference 1	63.9	0.78	0.60	15.0	-0.50
2	65.3	0.78	0.60	15.3	-0.46
3	71.6	0.78	0.69	13.4	-0.50
4	69.2	0.78	0.57	11.5	-0.49
5	69.4	0.78	0.51	11.9	-0.49

Alternative index	25th percentile	DE, HDG scores >14 (N=9)	DE, HDG scores >9 (N=35)	Standard deviation of reference	Spearman r with HDG score
6	69.2	1.00	0.74	12.2	-0.55
7	67.2	0.89	0.60	13.7	-0.48
8	67.2	0.89	0.63	13.5	-0.50
9	70.3	1.00	0.69	13.1	-0.50
10	70.9	1.00	0.77	13.6	-0.54
11	74.6	1.00	0.77	10.8	-0.58
12	73.3	1.00	0.77	12.0	-0.55

Table 10. Evaluation of alternative indices using validation data. Discrimination efficiencies (DEs) were based on the 25th percentile index values of calibration reference sites. Standard deviations are based on 9 reference sites, and correlations are based on 31 observations.

Alternative index	DE, HDG scores >14 (N=3)	DE, HDG scores >9 (N=13)	Standard deviation of reference	Spearman r with HDG score
All sites 1	1.00	0.92	7.4	-0.77
2	1.00	0.85	7.9	-0.78
3	1.00	0.85	7.2	-0.78
4	1.00	0.85	5.7	-0.74
5	1.00	0.92	6.1	-0.76
6	1.00	1.00	6.6	-0.79
7	1.00	1.00	6.3	-0.77
8	1.00	0.92	7.4	-0.80
9	1.00	1.00	5.8	-0.82
10	1.00	1.00	7.5	-0.76
11	1.00	1.00	7.4	-0.76
12	1.00	1.00	6.6	-0.79
Reference 1	1.00	0.62	8.6	-0.77
2	1.00	0.62	9.3	-0.73
3	1.00	0.85	8.0	-0.74
4	1.00	0.92	5.3	-0.75

Alternative index	DE, HDG scores >14 (N=3)	DE, HDG scores >9 (N=13)	Standard deviation of reference	Spearman r with HDG score
5	1.00	0.92	5.5	-0.78
6	1.00	1.00	5.8	-0.79
7	1.00	0.92	5.9	-0.81
8	1.00	0.92	6.4	-0.81
9	1.00	0.92	5.9	-0.83
10	1.00	1.00	7.1	-0.78
11	1.00	1.00	7.0	-0.77
12	1.00	1.00	6.2	-0.79

Table 11. Correlations of index alternatives with chemistry parameters for calibration and validation data.

	Calibration sites (N=105)			Validation sites (N=30)		
Alternative	pH	D.O.	Conductivity	pH	D.O.	Conductivity
All sites 1	0.38	0.43	-0.41	0.39	0.47	-0.55
2	0.35	0.41	-0.38	0.37	0.46	-0.56
3	0.39	0.45	-0.42	0.33	0.47	-0.60
4	0.39	0.53	-0.46	0.23	0.57	-0.55
5	0.39	0.52	-0.47	0.25	0.58	-0.55
6	0.39	0.53	-0.51	0.23	0.61	-0.59
7	0.36	0.47	-0.44	0.29	0.49	-0.54
8	0.40	0.50	-0.48	0.33	0.51	-0.56
9	0.37	0.50	-0.47	0.37	0.59	-0.57
10	0.41	0.51	-0.50	0.31	0.60	-0.54
11	0.40	0.58	-0.52	0.27	0.63	-0.58
12	0.40	0.52	-0.51	0.29	0.60	-0.55
Reference 1	0.39	0.42	-0.44	0.37	0.51	-0.54
2	0.38	0.41	-0.41	0.36	0.48	-0.50
3	0.39	0.44	-0.43	0.38	0.51	-0.55
4	0.38	0.52	-0.47	0.27	0.55	-0.53
5	0.39	0.52	-0.47	0.29	0.57	-0.53
6	0.40	0.53	-0.51	0.29	0.61	-0.55

	Calibration sites (N=105)			Validation sites (N=30)		
Alternative	pH	D.O.	Conductivity	pH	D.O.	Conductivity
7	0.38	0.49	-0.47	0.34	0.49	-0.53
8	0.39	0.49	-0.48	0.38	0.53	-0.55
9	0.38	0.52	-0.49	0.42	0.56	-0.53
10	0.42	0.52	-0.51	0.34	0.60	-0.55
11	0.40	0.58	-0.53	0.29	0.61	-0.59
12	0.42	0.52	-0.51	0.31	0.59	-0.55

Table 12. Number of samples processed using Imhoff and Caton laboratory subsampling and sorting methods, provided by region and human disturbance grouping.

Region	Method	Worst (HDG scores >14)	Fair (HDG scores 10-14)	Good (HDG scores 5-9)	Best (HDG scores 0-4)
Northern	Caton	0	0	3	0
	Imhoff	0	0	5	26
Southern	Caton	10	22	28	10
	Imhoff	2	14	26	10

Table 13. Equations for calculating Benthic Index of Biotic Integrity (B-IBI) scores for New Hampshire streams.

Metric	Scoring equation
Total taxa	*Total taxa*/21.5*100
Plecoptera taxa	*Plecoptera taxa*/4.4*100
% Chironomidae	(100 - *% Chironomidae*)/(100 – 0)*100
% Non-insects	(100 - *% Non-insects*)/(100 – 0)*100
% Clingers	*% Clingers*/94.6*100
% Intolerant	*% Intolerant*/76.1*100
Tolerant taxa	(6.2 - Tolerant taxa)/(6.2 – 0)*100

Figure 1. The stepwise development process used to develop macroinvertebrate index.

```
┌─────────────────────────────┐
│ Develop/identify stressor   │
│ (human influence) gradient  │
└─────────────────────────────┘
              │
┌─────────────────────────────┐
│ Identify reference and      │
│ impaired sites based on     │
│ gradient                    │
└─────────────────────────────┘
              │                          ┌─────────────────────────┐
┌─────────────────────────────┐          │ Adjust for watershed    │
│ Evaluate stream             │   ┌─────▶│ size, elevation,        │
│ classification based on     │   │      │ gradient as necessary   │
│ natural factors             │   │      └─────────────────────────┘
└─────────────────────────────┘   │                   │
              │                    │      ┌─────────────────────────┐
┌─────────────────────────────┐   │      │ • Range                 │
│ Identify candidate          │───┤      │ • Discriminatory power  │
│ metrics                     │   └─────▶│ • Relationship to       │
└─────────────────────────────┘          │   stressors             │
              │                          │ • Redundancy            │
┌─────────────────────────────┐          └─────────────────────────┘
│ Evaluate candidate metrics  │                     │
│ in stepwise process to      │◀────────────────────┘
│ reduce set                  │
└─────────────────────────────┘
              │
┌─────────────────────────────┐
│ Select metrics for          │
│ alternative indices from    │
│ those remaining             │
└─────────────────────────────┘
              │
┌─────────────────────────────┐
│ Set scoring of metrics      │
│ based on all sites or       │
│ reference sites             │
└─────────────────────────────┘
              │
┌─────────────────────────────┐
│ Evaluate alternative        │
│ indices with validation     │
│ data to select one index    │
└─────────────────────────────┘
              │
┌─────────────────────────────┐
│ Set biocriteria for aquatic │
│ life use attainment using   │
│ reference distribution      │
└─────────────────────────────┘
```
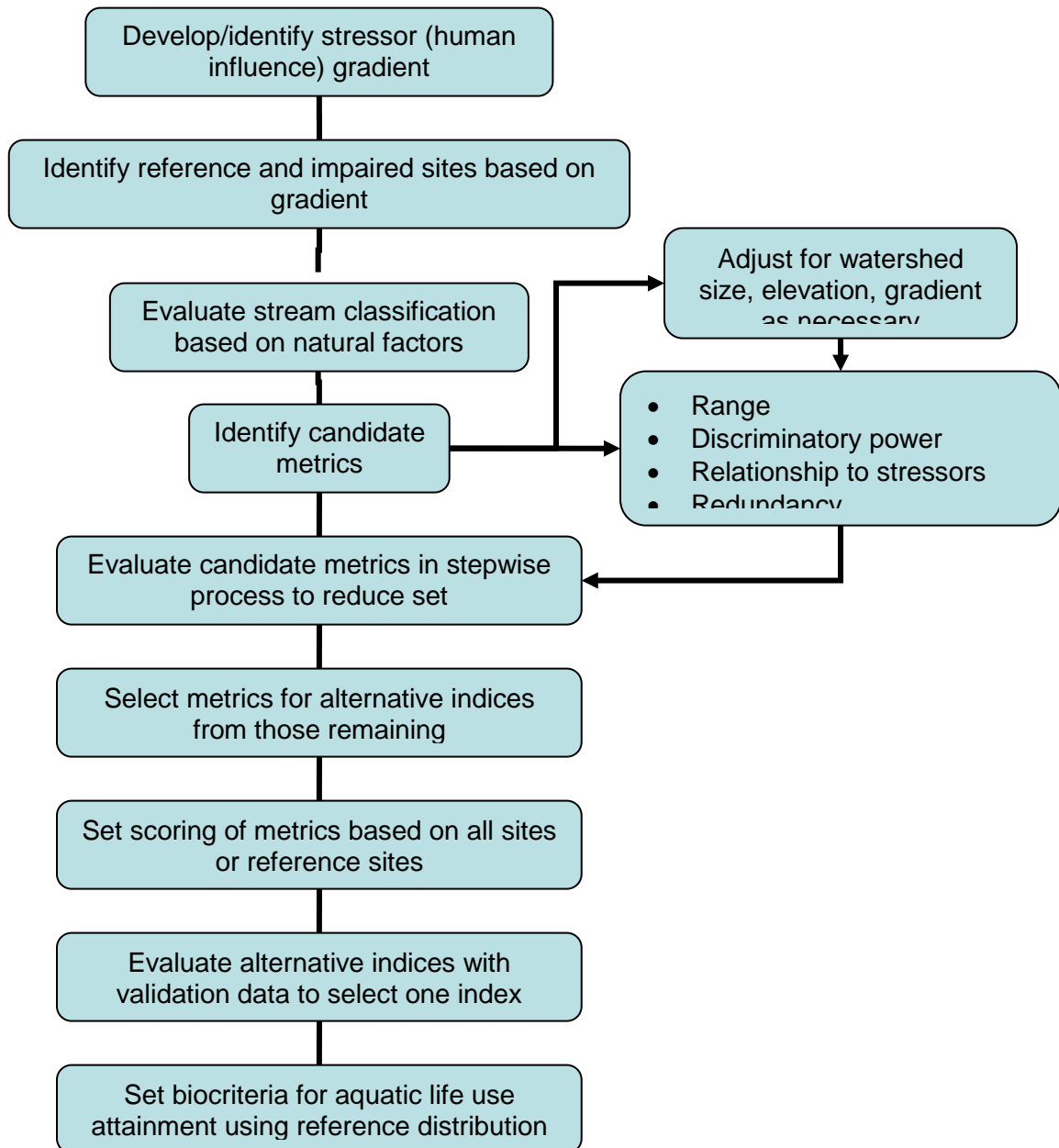
Figure 2. Nonmetric Multi-dimensional Scaling axes plotted by TNC regions and combined TNC regions.
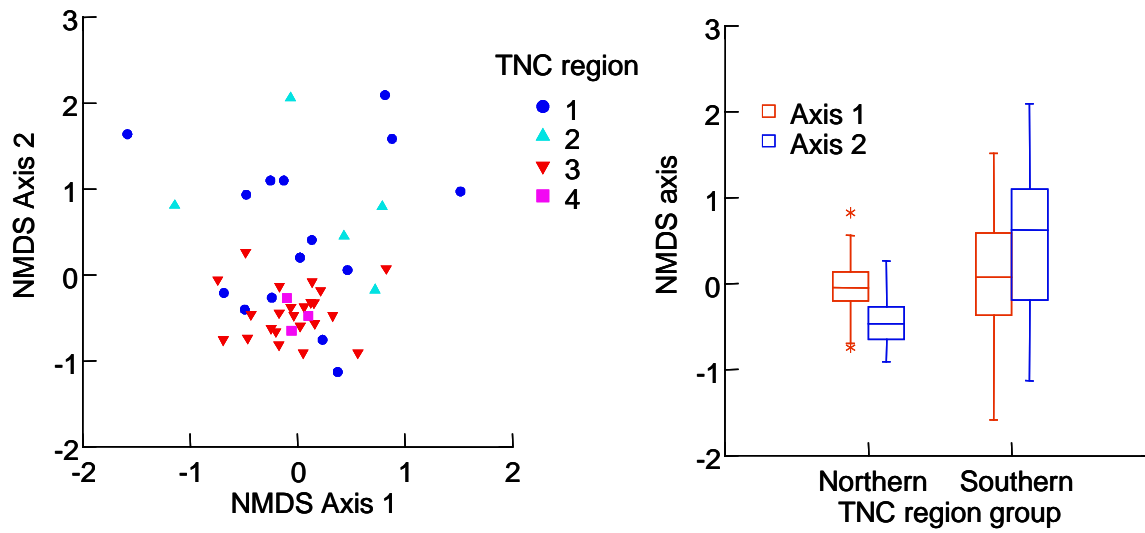
Figure 3. Correlation of NMDS axes with elevation and latitude, corresponding to northern and southern groups of TNC regions.
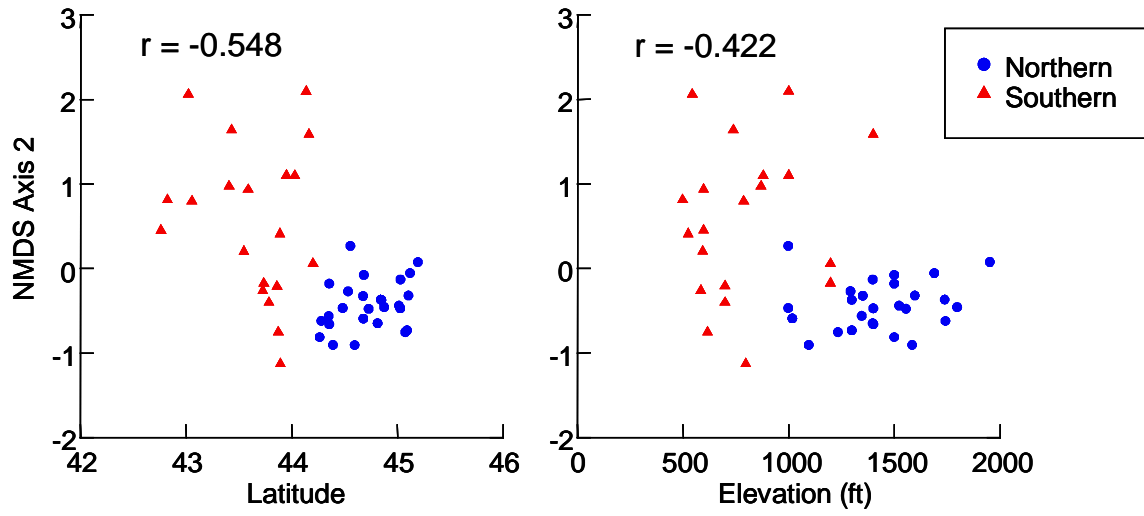
Figure 4. Comparison of alternative index 6 and 11 scores between Imhoff and Caton laboratory methods for Southern sites only.