Development of Indices of Biotic Integrity for Assessing Macroinvertebrate Assemblages in Massachusetts Freshwater Wadeable Streams

FINAL REPORT

Cold River - Charlemont, MA. Photo by MA DEP.

***Prepared for*:**

Massachusetts Department of Environmental Protection
Robert Nuzzo, Work Assignment Manager

***Prepared by*:**

Benjamin Jessup
Jen Stamp
Tetra Tech, Inc.
73 Main Street, Room 38, Montpelier, VT 05602

August 18, 2020

Executive Summary

The Massachusetts Department of Environmental Protection (MassDEP) is responsible for sampling and assessing Massachusetts's surface water quality pursuant to the Clean Water Act (CWA) Section 305(b). The Massachusetts Surface Water Quality Standards (SWQS) (314 CMR 4.00; MassDEP 2013) has narrative biological criteria that define biological integrity as "the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having species composition, diversity, and functional organization comparable to that of the natural habitat of the region." Waters supporting Aquatic Life Use should be suitable for "sustaining a native, naturally diverse, community of aquatic flora and fauna. This use includes reproduction, migration, growth and other critical functions" (MassDEP 2013).

To measure whether the macroinvertebrate communities in Massachusetts' freshwater wadeable streams exhibit biological integrity, Indices of Biotic Integrity (IBI) were calibrated using the reference condition approach to recognize characteristics of relatively undisturbed biological samples. The biological reference condition was calibrated for two naturally distinct regions of Massachusetts; the Western Highlands and the Central Hills. The multimetric indices for each region were comprised of biological metrics that were found to be responsive to a general stressor gradient. By scoring the metrics for each sample and averaging the scores in a multimetric index, the resulting index indicates the biological condition of the stream on a relative scale. The index values in the reference sites are reasonable expectations for any stream in the region and scores that do not resemble the reference scores indicate that there might be stressors influencing the biological condition.

Stream macroininvertebrate index development involved the following steps: compilation of MassDEP stream monitoring data, definition of site disturbance categories and criteria, establishment of stream classes, metric selection and scoring, index compilations, performance evaluation and selection of final indices. Through this process, two index formulations were developed for application in the Western Highlands and the Central Hills regions of Massachusetts, which were recognized for having naturally distinct biological expectations.

The discrimination efficiency of the indices showed that the separation of index values in least-disturbed reference and most disturbed stressed sites had minimal error (higher discrimination efficiency indicates that a greater percentage of stressed index values are outside of the reference inter-quartile range). The index selection process considered not only minimizing error, but also including metrics that were ecologically meaningful and diverse in response mechanisms. The metrics included in each index represented four categories of metric responses: taxa richness, individual composition, functional feeding groups, and pollution tolerance Table ES-1).

The IBI in the Central Hills was more sensitive to the stressor gradient than the IBI in the Western Highlands (Index DE: 100% and 84%, respectively). This difference is attributed to the difference in the general stressor intensity across the landscape of Massachusetts. For example, there are more areas with sparse development in the west and more ubiquitous and severe stressor conditions in the east. The indices were successfully validated with independent data. The error rate for validation reference and stressed data sets was within 10% of the calibration data. Additional checks that were performed on multihabitat and pre-2000 data (both of which were excluded from the calibration and validation datasets) also showed adequate validation of the indices.

The new RBP kick net IBIs improve MassDEP's diagnostic ability to identify degradation in biological integrity and water quality. The IBIs are modernized compared to past assessment indices used in

Massachusetts and make use of data that were collected from hundreds of sites in recent years. MassDEP will use the IBIs to assess stream degradation relative to least-disturbed streams in the Western Highlands and Central Hills and has begun to explore potential thresholds for four biological condition categories (Exceptional Condition, Satisfactory Condition, Moderately Degraded, and Severely Degraded). In the future, MassDEP may decide to work towards establishing numeric bio-criteria that would be integrated into the SWQS and would be used to evaluate Aquatic Life Use Attainment (ALU) decisions. If MassDEP decides to make this a future pursuit, the proposed criteria would need to go through a rule-making process that includes a period for public review and comment.

Table ES-1. Metrics included in the Central Hills and Western Highlands IBI. DE = discrimination efficiency. Trend is the direction of metric response with increasing stress.

Metric abbrev	Metric	DE	Trend
Central Hills IBI (Model 6_ 33344)			
nt_total	Number of taxa	66.7	Dec.
pt_EPT	Percent EPT taxa	76.7	Dec.
pi_Ephem NoCaeBae	Percent Ephemeroptera individuals excluding Caenidae and Baetidae	66.7	Dec.
pi_ffg_filt	Percent collector-filterer individuals	76.7	Inc.
pt_ffg_pred	Percent predator taxa	90	Dec.
pt_tv_intol	Percent intolerant taxa	100	Dec.
Western Highlands IBI (Model 6_32701)			
nt_total	Number of taxa	52.4	Dec.
pi_Pleco	Percent Plecoptera individuals	66.7	Dec.
pi_ffg_filt	Percent collector-filterer individuals	50	Inc.
pi_ffg_shred	Percent shredder individuals	61.9	Dec.
pi_tv_intol	Percent intolerant individuals	59.5	Dec.
x_Becks	Becks Biotic Index	57.1	Dec.

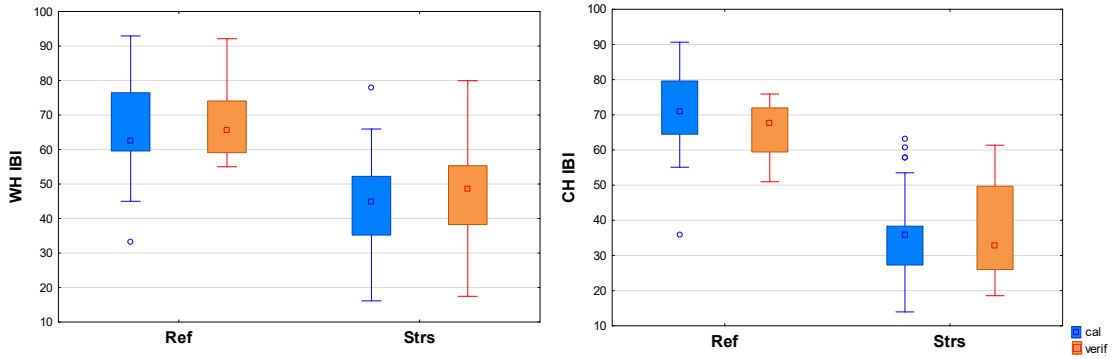


Figure ES-1. Distributions of Western Highland (WH, left) and Central Hills (CH, right) IBI values in reference (Ref) and stressed (Strs) sites in calibration (cal, blue) and verification (verif, orange) data sets.

Acknowledgments

The index development process was supported by the Massachusetts Department of Environmental Protection (MassDEP) through a contract with Tetra Tech, Inc. Robert Nuzzo and Kimberly Groff of MassDEP facilitated the contract. The index development team, which consisted of Robert Nuzzo, James Meek, Arthur Johnson, Robert Maietta, Joan Beskenis, Allyson Yarra, and Anna Mayor, participated in bi-weekly calls and provided feedback throughout the process. In addition, Robert Nuzzo provided additional levels of support by providing data and reviewing taxonomic nomenclature, taxa attribute assignments and index alternatives, James Meek reviewed site disturbance assignments and index alternatives. We owe thanks to our regional state partners as well for providing data. They included Steve Fiske and Aaron Moore from Vermont Department of Environmental Conservation, Mary Becker and Chris Bellucci from the Connecticut Department of Energy & Environmental Protection, and Katie DeGoosh from Rhode Island Department of Environmental Management. We are very grateful for the hard work and enthusiasm of all the project participants.

Table of Contents

Appendixes

List of Tables

List of Figures

1 Background

The Massachusetts Department of Environmental Protection (MassDEP) is responsible for sampling and assessing Massachusetts's surface water quality pursuant to the Clean Water Act (CWA) Section 305(b) as well as, according to Section 303(d) of the CWA, identifying waterbodies of the State that are not meeting water quality criteria and are in need of developing a Total Maximum Daily Load (TMDL). To help meet these requirements, MassDEP monitors biological conditions and assesses the integrity of macroinvertebrate assemblages in freshwater streams and rivers (Massachusetts Division of Watershed Management Watershed Planning Program 2016). MassDEP's biomonitoring program has been collecting macroinvertebrate data since the early 1980's.

The Massachusetts Surface Water Quality Standards (314 CMR 4.00; MassDEP 2013) has narrative biological criteria that defines biological integrity as "the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having species composition, diversity, and functional organization comparable to that of the natural habitat of the region." Waters supporting Aquatic Life Use should be suitable for "sustaining a native, naturally diverse, community of aquatic flora and fauna. This use includes reproduction, migration, growth and other critical functions" (MassDEP 2013).

To measure whether or not the macroinvertebrate communities in Massachusett's streams are meeting this definition, MassDEP has been using a multimetric index to rate samples. Multimetric indices are numeric representations of biological conditions based on the combined signals of several different assemblage measurements. The raw measurements are recalculated or standardized as biological metrics, or numerical expressions of attributes of the biological assemblage (based on sample data) that respond to human disturbance in a predictable fashion. Impacts to the benthic community may be indicated by the absence of generally pollution-sensitive macroinvertebrate taxa such as Ephemeroptera, Plecoptera, and Trichoptera (EPT); dominance of a particular taxon, especially pollution-tolerant taxa; low taxa richness; or shifts in community composition relative to the reference station (Barbour et al. 1999).

MassDEP's existing multimetric index is based on a modification of the Rapid Bioassessment Protocols (RPB) III metrics and scoring document (Plafkin et al. 1989). It is comprised of seven metrics that are recommended in the RBP document: taxa richness, biotic index, EPT index, EPT/Chironomidae, Scrapers/Filterers, % dominant taxon and reference site affinity. Metrics are calculated and then scored based on comparability to one or two high quality sites (referred to as "reference" sites) within the same watershed basin that have comparable drainage areas. MassDEP then assigns samples to one of four categories (% of reference condition): non-impaired (>83%), slightly impaired (54 – 79%), moderately impaired (21 – 50%), and severely impaired (<17%). Each impact category corresponds to a specific aquatic life use-support determination used in the CWA Section 305(b) water quality reporting process. Non-impacted and slightly impacted communities are assessed as "support" in the 305(b) report; moderately impacted and severely impacted communities are assessed as "impaired."

MassDEP started using the RBP III index in the late 1980s. Since that time, its biomonitoring program has collected hundreds of macroinvertebrate samples that capture a wide range of human disturbance (from minimally disturbed to highly stressed). In addition, new tools have become available for assessing human disturbance. Examples include the Massachusetts Human Disturbance Index (HDI) (Weiskel et al. 2010), the USEPA's Stream-Catchment (StreamCat) Dataset (Hill et al. 2016) and the Indices of Catchment and Watershed Integrity (ICI and IWI) (Thornbrugh et al. 2018). The goal of this project was to use these data to develop new, more modern biotic indices for

MassDEP's macroinvertebrate samples. The new indices, which are called Indices of Biotic Integrity (IBIs), were developed using a well-established, multimetric approach pioneered by Karr (1981). The IBIs were calibrated for freshwater wadeable streams in all but the southeastern portion of the state[1], and are broken into two stream classes: Western Highlands and Central Hills. The new IBIs will supplement the existing RPB III index and improve MassDEP's diagnostic ability to identify degradation in biological integrity and water quality. In this report we describe the development of the two IBIs, which involved the following steps: data compilation and preparation, definition of site disturbance categories and criteria, establishment of stream classes, metric selection and scoring, index compilations, performance evaluation and selection of the final IBI for each stream class.

2 Data Compilation and Preparation

IBI development began with the assembly and analysis of macroinvertebrate and environmental data, including habitat, water quality data and GIS-derived landscape-level data such as land cover. The data were compiled into a MS Access relational database.

2.1 Macroinvertebrates

Regional dataset
The Massachusetts macroinvertebrate data came from two sources: MassDEP and the Deerfield River Watershed Association (DRWA). Sampling locations are shown in Figure 1. The majority of samples were collected with the Rapid Bioassessment Protocol (RBP) kick net method, which involves kicking or disturbing bottom sediments and catching the dislodged organisms in a net as the current carries them downstream (Barbour et al. 1999). Other collection methods in the MassDEP dataset included the RBP multihabitat method and pilot methods unique to specific projects. Ultimately only RBP kick net samples were used for IBI calibration[2].

In addition to the MassDEP and DRWA samples, macroinvertebrate data were obtained from the Connecticut Department of Energy & Environmental Protection (CT DEEP) and Rhode Island Department of Environmental Management (RI DEM) (sampling locations are shown in Figure 1). The objective was to evaluate whether CT DEEP and RI DEM samples were comparable enough to the MA samples to use in the IBI calibration dataset for southeastern MA, where high quality "reference" sites were known to be lacking. Differences between three datasets were too large to utilize the combined datasets for calibration of the IBIs (see Text Box #1). However, all of the data were retained in the MS Access database for potential future applications.

[1]The Narragansett/Bristol Lowlands, Cape Cod, and the Islands were excluded because they had insufficient RBP kick net data to develop an IBI at this time.
[2] Exploratory analyses were performed to evaluate differences between samples collected with the MassDEP multi-habitat method vs. the kick net method. Differences were large enough to warrant exclusion of the multi-habitat samples from the IBI calibration dataset (see Supplemental Materials #1) but further exploration is encouraged as more multi-habitat samples are collected in the future.

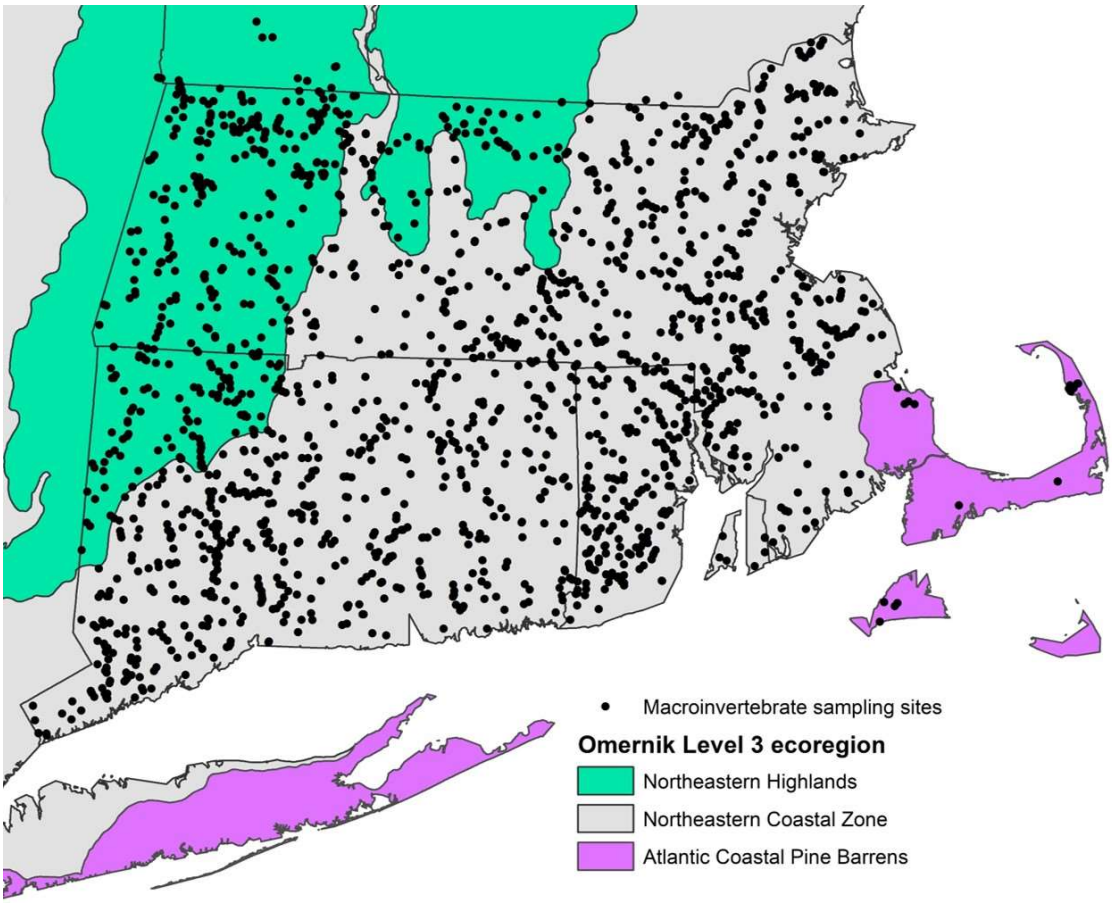


Figure 1. Sites with macroinvertebrate data that are included in the regional dataset (MA/CT/RI).

TEXT BOX #1 – Assessing comparability of MA, CT DEEP and RI DEM samples

To assess comparability, we reconciled differences in taxonomic nomenclature across the three datasets, performed Non-metric Multidimensional Scaling (NMS) ordinations of taxa and metrics, calculated commonly-used metrics, generated box plots and looked for patterns. Analyses were limited to reference sites only. Differences were evident in the ordination (Figure 2) as well as the box plots. The CT DEEP samples had higher median EPT taxa richness, the MA samples had higher median Chironomid taxa and the RI DEM samples had lower median total taxa (Figure 3). We concluded that samples from the three entities were not comparable enough to combine for the IBI calibration dataset.

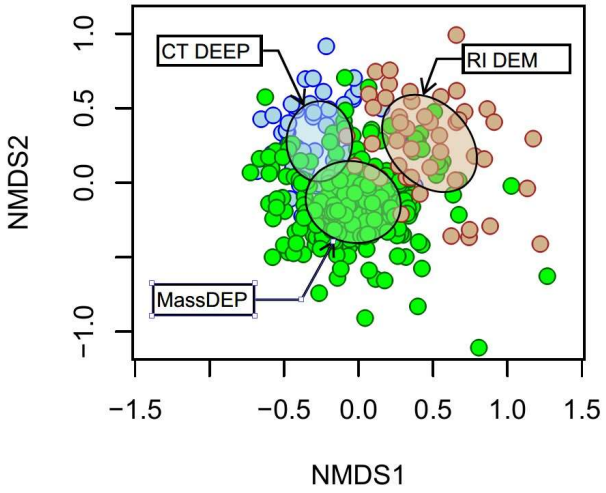


Figure 2. The NMDS ordination showed differences in taxonomic composition across entities. This plot is based on presence/absence data from the 'All Ref Samps dataset (PA, 312)'.

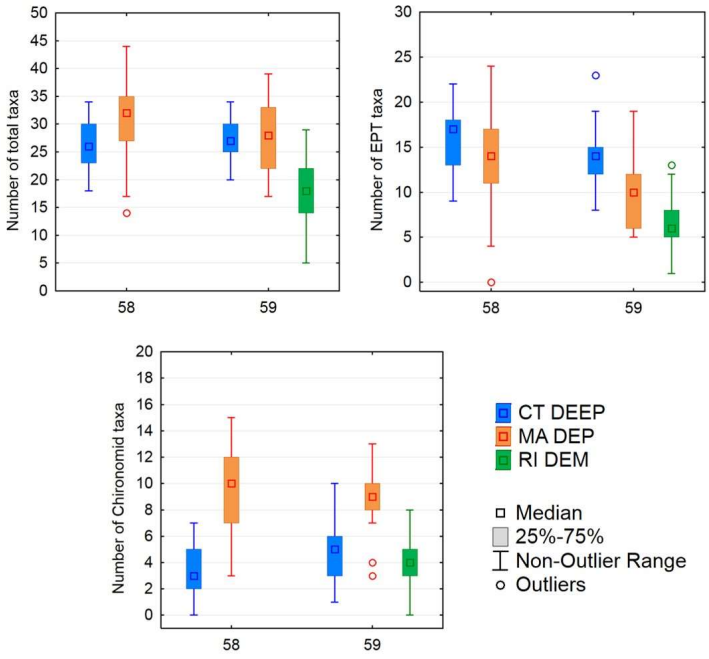


Figure 3. Box plots showing distributions of metric values (number of EPT taxa, total taxa and Chironomid taxa) across entities and Level III ecoregions (58 = NE Highlands, sample size CT DEEP = 29, MA DEP 63; 59 = NE Coastal Zone, sample size CT DEEP = 33, MA DEP = 26, RI DEM = 53). The box plots are based on kick net samples collected from 2000 onward at reference sites.

Massachusetts RPB kick net dataset

The IBIs were calibrated using MassDEP and DRWA RBP kick net samples collected from freshwater perennial wadeable streams. The dataset spanned 34 years (1983-2017). Samples were collected during the months of July through September when baseflows are typically at the lowest of the year and levels of stress to aquatic organisms are presumed to be at peak. The MassDEP samples were collected in accordance with MassDEP's standard operating procedures (Nuzzo 2003) and Quality Assurance Project Plan (QAPP) (MassDEP 2004). Some sites were located just across the Massachusetts border. These samples were included in the analyses because they were sampled by MassDEP field crews using MassDEP kick net methods. Sample collection was conducted throughout a 100-m reach, in riffle/run areas with fast currents and rocky (cobble, pebble, and gravel) substrate, which are generally the most productive habitats, supporting the most diverse communities in the stream system. Field crews used a kick-net with a 46-cm wide opening and 500-µm mesh. Ten kicks in squares approximately 0.46 m x 0.46 m were composited for a total sample area of about 2 m^2. Samples were labeled and preserved in the field with denatured 95% ethanol, then brought to the MassDEP lab for sorting. The sorting procedure entailed distributing whole samples in pans, selecting grids within the pans at random, and sorting specimens from the other materials in the sample until, depending on the project, either 100 organisms (±10%) or 300 organisms (±20%) were extracted. Specimens were identified to genus or species as allowed by available keys, specimen condition, and specimen maturity. Table 1 summarizes the main elements of the RBP kick net protocol.

Table 1. Summary of the macroinvertebrate sample protocol elements for the MassDEP Rapid Bioassessment Protocol (RBP) kick net method.

Effort	Gear	Habitat	Sampling area	Index period	Target # organisms	Taxonomic resolution
10 kick-samples are taken in riffle habitats within the sampling reach and composited	Kick-net, 46-cm wide opening, 500-mm mesh	Riffle/run is the preferred habitat	Approximately 2 m^2	July 1–September 30	100	Lowest practical level

Data Preparation

After the MA, CT and RI macroinvertebrate data were compiled, the following steps were performed to prepare the data for analyses (see Appendix A for more detailed descriptions of these steps):

- **Compile master taxa list and reconcile differences in nomenclature**
- **Assemble Benthic Master Taxa table**
  - Phylogenetic information (Phylum, Class, Order, Family, Subfamily, Tribe, Genus, Species)
  - Attributes (functional feeding group (FFG), tolerance value, habit, life cycle/voltinism and thermal preferences) – see Table 2
  - Non-target designations (e.g., Hemiptera, crayfish; Appendix A)
- **Assemble Benthic table (taxa and counts), based on 100-count samples**
  - Rarify/randomly subsample data as needed
  - Exclude redundant taxa on a sample-by-sample basis so that they are not included in richness calculations (decision critieria are shown in Appendix A)
- **Assess suitability of samples for the IBI calibration dataset**

  - Considerations: geographic location, number of samples per site (to avoid bias, only one sample per site was included), number of total individuals (not too many or too few), collection method, month, year, level of taxonomic resolution
- **Assess need for Operational Taxonomic Units (OTUs) (Cuffney et al. 2007)**
  - Evaluate whether any taxonomic groups need to be collapsed to a higher level of taxonomic resolution due to inconsistencies over time (e.g., should mites be collapsed to Order-level, Chironomids to tribe or subfamily-level or worms to family-level?)
- **Calculate metrics**
  - Table 2 contains a summary of attributes that were used in the metric calculations. Metrics were calculated with the BioMonTools R package (https://github.com/leppott/BioMonTools). Appendix B contains the list of metrics that were calculated and considered as candidates for inclusion in the IBIs.

The regional (CT/RI/MA) version of the Benthic Master Taxa table included 1320 unique taxa. When the taxa list was limited to MA samples only, there were 661 taxa. Table 2 shows what percentage of the taxa in the MA IBI calibration dataset had attribute assignments. Metrics based on habit, thermal preference and life cycle were considered exploratory in this phase of work. Only the FFG and tolerance value metrics (which MassDEP had prior experience with) were ultimately used in the IBIs.

Appendix A contains a more detailed description of the data preparation steps.

Table 2. Summary of the types of taxa attribute data that were compiled for the Benthic Master Taxa table. The total number of taxa (514) was derived from the Central Hills + Western Highlands IBI kick-net dataset (both calibration and verification samples) collected from 2000 onward. Metrics based on habit, thermal preference and life cycle were considered exploratory in this phase of work.

Attribute	Description	Categories	Sources	Number of taxa with attribute assignments (out of 514)	Percent of total
Functional feeding group (FFG)	Refers to the primary type of food resource that a particular species utilizes in the stream	PR = predator, CG = collector-gatherer, SH = shredder, SC = scraper, CF = collector-filterer	MassDEP, CT DEEP, VT DEC, NRSA, Poff et al. 2006, Tetra Tech	497	96.7%
Tolerance values (TolVal)	Relative sensitivity to pollution. In this dataset, the tolerance values are oriented toward detection of organic pollution (per Hilsenhoff 1987)	Values range from 0 (most *intolerant*) to 10 (most *tolerant*). Intolerant taxa 0 to 3. Tolerant taxa 7 to 10	MassDEP*, VT DEC, CT DEEP, NRSA, Tetra Tech	489	95.1%
Habit	Distinguishes the primary mechanism a particular species utilizes for maintaining position and moving in the aquatic environment (Merritt et al. 1996)	SP = sprawler, SW = swimmer, CN = clinger, CB = climber, BU = burrower	NRSA, VT DEC, Poff et al. 2006, Vieira et al. 2006, Tetra Tech	469	91.2%
Life Cycle/ Voltinism	Number of broods or generations a species typically produces in a year	Uni (one), semi, multi (multiple)	NRSA, Poff et al. 2006	290	56.4%
Thermal preference	thermal preference/optima	cold_cool or warm	U.S. EPA 2012, U.S. EPA 2016	66 taxa were assigned to the cold/cool group; 32 taxa were assigned to the warm group	NA**

*The MassDEP tolerance values came from the following sources: NYS DEC (Bode 1996, 2002), Lenat 1993, Hilsenhoff 1987, Robert Nuzzo, VT DEC (Steve Fiske, personal communications). They were not generated from analyses of MassDEP data.

**Only the number of taxa assigned to the cold/cool and warm groups are reported here; the total number of taxa assessed during this pilot study were not available.

2.2 Habitat and water quality

Qualitative habitat and water quality data from MassDEP were obtained and added into the Microsoft Access relational database for use in site disturbance characterizations (Section 3). These data were collected by field crews at the time of the biological sampling events. MassDEP assessed habitat qualities for two stream types (RR = riffle/run; GP = glide-pool) using a modified version of the RBP evaluation procedure in Barbour et al. (1999). Field crews performed visual assessments and assigned scores to ten parameters: bank stability (left and right bank), bank vegetative protection (left and right bank), riparian vegetative zone (left and right bank), bottom substrate/available cover, channel flow status, channel alteration, channel sinuosity, pool substrate characterization, pool variability, and sediment deposition. Each metric was scored on a scale of either 0-10 or 0-20, then summed to get a total score (higher scores indicated better habitat quality). In addition to the RBP habitat assessment, some sites had visual estimates of substrate composition (clay, sand, gravel, cobble, boulder, bedrock) and qualitative assessments of water and sediment quality (water odor, color, surface oil and turbidity, sediment odors, oils, deposits).

2.3 Landscape-scale (GIS-based)

Several datasets with landscape-scale metrics were obtained for site disturbance characterization (Section 3) and classification (Section 4). One was the Human Disturbance Index (HDI), which is available for USGS-delineated hydrologic units in Massachusetts (Weiskel et al. 2010). The HDI is based on seven indicators of human disturbance: three streamflow alteration indicators (August flow, water-use intensity, dam storage ratio) and four landscape indicators (impervious cover, local impervious cover, agriculture – local & watershed scale). The indicator metrics were converted to unitless scores and scaled from 1 to 5, with 5 being the most disturbed.

Because HDI scores were only available for MA, to characterize disturbance on a consistent regional scale we had to find an alternate dataset that covered all three states (MA, CT and RI). We selected the USEPA's Stream-Catchment (StreamCat) Dataset (Hill et al. 2016), which covers the contiguous US. StreamCat is an extensive database of natural and anthropogenic landscape metrics that are associated with the National Hydrography Dataset (NHD) Plus Version 2 (NHDPlusV2) stream segments (McKay et al. 2012). StreamCat data are available at two spatial scales: local catchment and full upstream watershed (Figure 4). Some variables address site disturbance characterization (e.g., % agricultural cover, % urban cover, road density, and specific discharges or activities (National Pollutant Discharge Elimination System discharges, Confined Animal Feeding Operations, mining activity, etc.). Natural (classification) variables include geologic types, elevation, stream slope, catchment size, ecoregion, temperature and precipitation, among others. A list of the candidate StreamCat metrics that were considered can be found in Appendix C, Table C1.

In 2018, the indices of catchment and watershed integrity (ICI and IWI, respectively) (Thornbrugh et al. 2018, Johnson et al. 2019) [3] were added to the StreamCat dataset. The ICI and IWI provide a relative quantification of human-related stress to the ecosystem services provided by watersheds (Flotemersch et al. 2016, Thornbrugh et al. 2018). The ICI is scaled to the local catchment and the IWI to the total watershed (Figure 4). IWI and ICI scores are often (but not always) similar.

[3] In 2019, EPA released two updates to the ICI/IWI dataset (version 2 and 2.1) (Johnson et al. 2019). We used Version 1 of the ICI and IWI for this project (Thornbrugh et al. 2018) because the later versions did not become available until after the site review had been completed (Section 3). When comparisons were made between the Version 1 and Version 2.1 datasets, relative patterns across catchments and watersheds were similar, but scores in the IWI version 2.1 were generally lower).

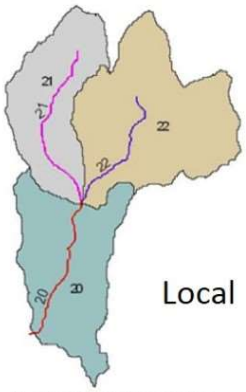
Differences between ICI and IWI scores are more likely to occur in catchments with large versus small drainage areas. ICI and IWI scores range from 0 to 1, with higher values having greater integrity/better watershed condition.

The ICI and IWI scores are based on six components: hydrologic regulation (HYD), regulation of water chemistry (CHEM), sediment regulation (SED), hydrologic connectivity (CONN), temperature regulation (TEMP), and habitat provision (HABT). The six components are scored based on StreamCat metrics that have been shown within the literature to be associated with degraded key watershed functions and have corresponding geospatial datasets that can be mapped. As with the ICI and IWI, component scores range from 0 to 1. Scores for the six components are multiplied together to get the overall ICI and IWI scores, so that a poor score for any one of the six metrics may be enough to produce a low IWI score (Thornbrugh et al. 2018).

To associate the biological sampling sites with the StreamCat and ICI/IWI data, an intersect procedure was performed with Geographic Information System software (ArcGIS 10.3.1), which created an attribute table with a list of the biological sampling stations and unique identifiers for the NHDPlusV2 catchments (COMID/FEATUREID). The COMID was then used to link the biological sampling sites with the StreamCat data tables, which were downloaded from the StreamCat website (http://www2.epa.gov/nationalaquatic-resource-surveys/streamcat). The data were uploaded to MS Access and queries were created to generate tables with the desired StreamCat metrics. Because the ICI and IWI had not been tested in MA yet, comparative analyses were performed to evaluate differences between the ICI, IWI and HDI scores for the biological sampling sites. Overall there was good correspondence (Appendix D).

There are limitations with using the HDI or StreamCat data for characterizing sites. One is that the data are not based on exact watershed delineations (except in instances where the site happens to be located at the downstream end of the NHDPlusV2 local catchment for the ICI/IWI or USGS-delineated catchment for the HDI). Instead, a site is characterized based on whatever attributes are associated with the catchment in which the site is located. In some cases, this may cause inaccuracies (e.g., if the site is located upstream of urban land cover, but the urban land cover is located within the same catchment, the urban land cover data are wrongly associated with the site). Another limitation stems from the resolution (1:100K) and accuracy of the NHDPlusV2 dataset. Not all of the sites match with NHDPlusV2 flowlines. This issue tends to arise with sites on very small tributaries (that are too small to show up in the NHD 1:100K) as well as streams in high intensity urban areas that have been altered.

In addition to StreamCat data, NHDPlusV2 attribute data for flowline type (stream/river, canals/ditches, coastline, and artificial pathway; Appendix C, Table C2) and slope were associated with biological sampling sites, as were EPA level III and IV ecoregions and hydrologic basins (Hydrologic Units Codes – HUCs).



A. Local Catchments for Reaches 20, 21, and 22

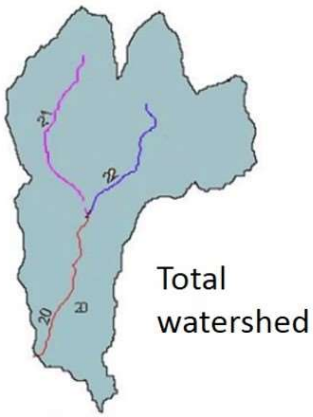
Local catchment

Definition: the landscape area draining to a single stream segment, excluding upstream contributions.

In this example, there are three local catchments (associated with unique flowline segments) –

- # 20 (green)
- # 21 (gray)
- # 22 (brown)

Each local catchment has a unique identifier (COMID or FEATUREID).



B. Total Upstream Watershed for Reach 20

Watershed-level

Definition: the local catchment plus the accumulated area of all upstream catchments

In this example there is one total watershed, comprised of the three local catchments (#20 + #21 + #22).

Figure 4. USEPA's StreamCat metrics (Hill et al. 2016) cover two spatial scales: local catchment and total watershed.

3 Site Disturbance Characterization

Purpose

Bioassessment is based on comparison of conditions in assessable waterbodies to sites with relatively natural reference conditions, which are often referred to as reference sites. Reference sites serve several purposes, including index calibration, site classification and setting of biocriteria thresholds. Biotic indices (like IBIs) are calibrated based on a disturbance gradient. Capturing the full gradient, from best to worst, is important for index calibration. Reference sites are used to identify metric expectations with the least levels of disturbance. When a set of stressed sites are identified using criteria at the opposite end of the disturbance scale, the response of metrics along the resulting stressor gradient can be detected. The direction and strength of response can be used for selecting candidate metrics for inclusion in an assessment index (like an IBI) and properly scoring them.

Reference sites are also used for classification. The biological characteristics associated with the natural environmental setting are best recognized when they are not confounded by the effects of human disturbance. In the site classification process, the distribution and abundance of biota or the

distribution of metric values in minimally or least disturbed sites are used to identify biological groups and responses to natural gradients. By accounting for such natural biological variability, an IBI can be specifically calibrated to the natural stream type and the responses to disturbance that might be unique to each stream type.

Once site classes are established and indices are calibrated, some entities establish thresholds for numeric biocriteria. The reference condition (RC) approach is the most commonly used method to derive biological thresholds (e.g., Yoder and Rankin 1995, DeShon 1995, Barbour et al. 1996, Roth et al. 1997). With the RC approach, IBI scores are calculated from a reference site dataset, and then a percentile of the IBI scores, such as the 25th or 10th, is chosen to represent the RC.

Approaches

In order to develop a disturbance gradient for a population of sites, it is necessary to specify criteria for the least disturbed and most disturbed sites. The criteria should be clearly defined and documented, and should be based on a priori measures of condition that are independent of the biology (U.S. EPA 2013). There is no universal method for designating reference sites but most entities use a combination of desktop screening of landscape-scale factors (watershed and local scale), water quality, habitat scores, best professional judgment (BPJ) and site visits. The land use/land cover criteria (whether single index or multiple measures) may be based on partial catchments, buffers around a stream, or for the entire watershed. Land use categories that are commonly summarized and used as criteria include forest, natural cover, agriculture, and urban (U.S. EPA 2013).

States biomonitoring programs have used a variety of methods to define reference and stressed sites. Some have developed a generalized disturbance index and set thresholds to designate reference and stressed sites. Examples include Minnesota's Human Disturbance Score (Bouchard et al. 2016) and the Landscape Development Index (Brown and Vivas 2005, Fore et al. 2007). In recent years, several states (Illinois (Tetra Tech 2015), Indiana (Jessup and Stamp 2017) and Michigan (Tetra Tech in progress)) have used an approach that designates multiple categories of disturbance, with up to three levels of reference (Best Reference (BestRef), Reference (Ref) and Sub-Reference (SubRef)) - which allows for recognition that reference sites in some areas do not represent pristine or nearly pristine natural landscapes; three levels of stress (Some Stress (SomeStrs), Stress (Strs) and High Stress (HighStrs)); and one category in the middle (Other). This approach is similar in concept to the minimally disturbed, least disturbed and best attainable categories proposed in Stoddard et al. (2006).

MA approach

A wide range of disturbance is represented in the MA dataset, on a noticeable east-west gradient. Eastern MA generally has higher levels of disturbance than western MA (Figures 5 & 6), with a number of large urban areas occurring in eastern MA (including Boston) (Figure 5).

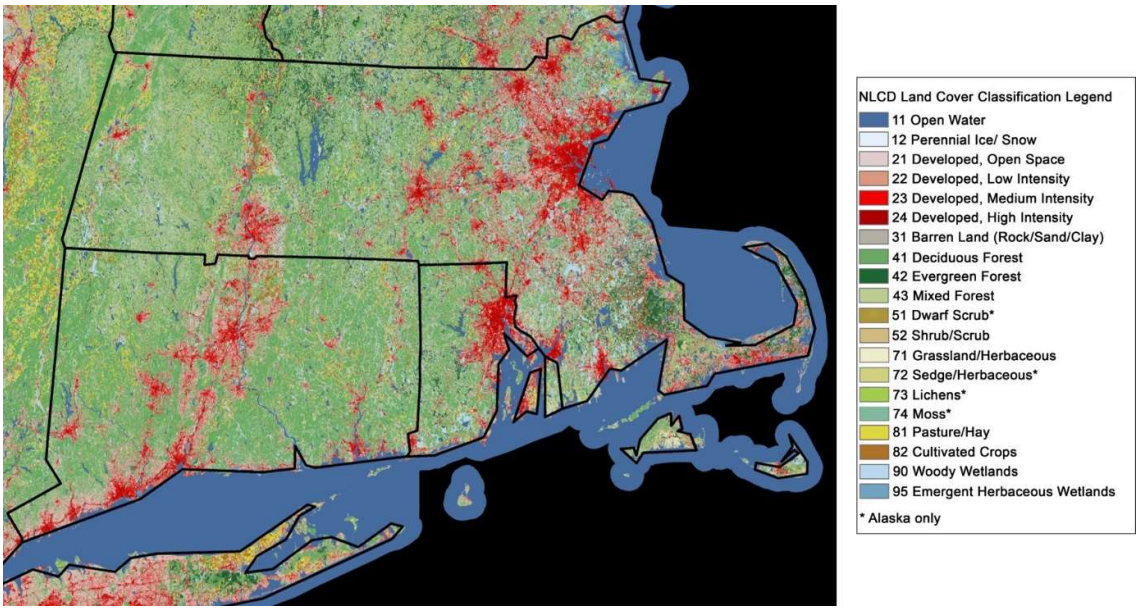


Figure 5. Eastern Massachusetts has higher levels of urban land cover and human disturbance than western MA (source: National Land Cover dataset (NLCD) 2011).

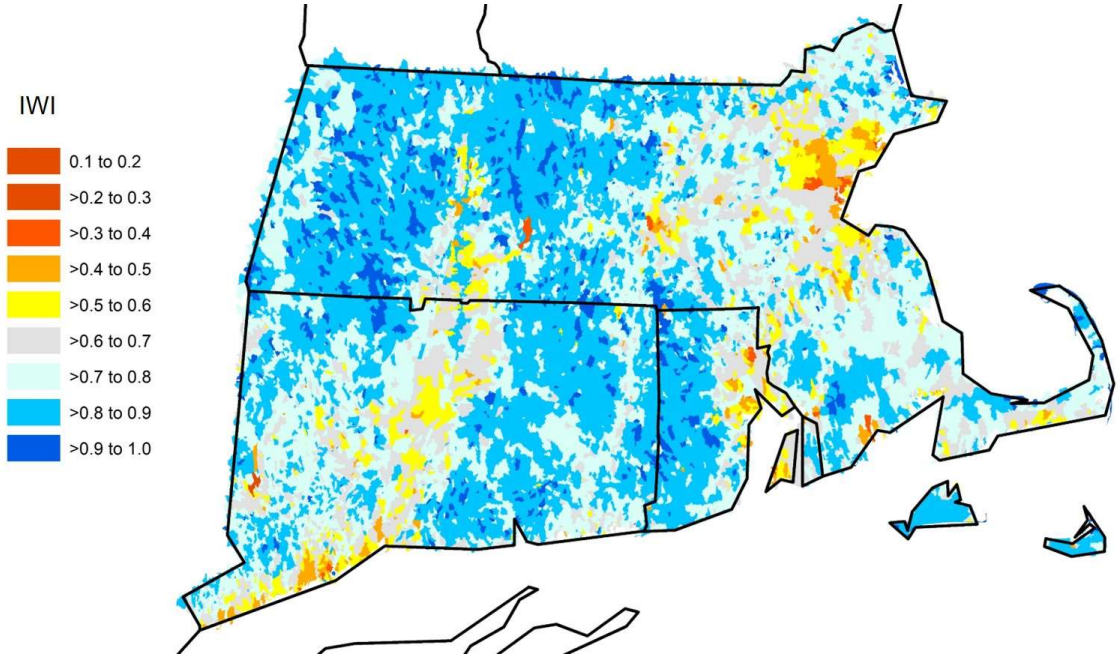


Figure 6. Index of Watershed Integrity (IWI) scores (version1; Thornbrugh et al. 2018) for NHDPlusV2 catchments in MA, CT and RI. Higher integrity catchments (which have higher scores) are in blue colors, moderate integrity in gray and yellow, and lower integrity in orange and red colors.

The 'multiple categories of disturbance' approach was used to establish the disturbance gradient in the MA IBI calibration dataset. Seven categories of disturbance were defined (BestRef, Ref, SubRef, Other, SomeStrs, Strs, HiStrs). The gradient was based on selected StreamCat metrics and the ICI and IWI scores.

Analytical steps were as follows:

- **Run a Principal Components Analysis (PCA) to narrow down the list of candidate stressor metrics (Figure 7).** Strong candidates explain high amounts of variation on the major stressor axes; are not redundant (Spearman $|r| < 0.8$-0.9); cover different spatial scales (local and watershed); are meaningful and stressor-based. PCAs were run on the statewide MA dataset as well as the two main EPA Level III ecoregions (Northeastern Highlands and Northeastern Coastal Zone).
- **Select the strongest candidate variables**. Based on the considerations listed above, seven metrics were selected: ICI, IWI, % urban land cover (watershed-scale), density of roads, dam storage volume, % agricultural land cover (local catchment), modeled mean rate of fertilizer application + biological nitrogen fixation + manure application (Table 3).
- **Establish thresholds/disturbance criteria for each metric**. Thresholds were set for each metric across seven disturbance levels (ranging from least disturbed to most disturbed) (Table 4). The 10th, 25th, 50th, 75th and 90th percentiles served as a starting point. Adjustments were made as needed to attain adequate sample sizes in the least and most disturbed categories in each Level III ecoregion (due to the east-west gradient of disturbance in MA, it was challenging to find adequate numbers of reference sites in the east and sufficient numbers of highly stressed sites in the west). Ecoregions were chosen to illustrate and select thresholds because they were likely classification variables; thus, using them allowed spatial distribution of reference and stressed sites throughout the state, and resulted in representation from various reference site types.
- **Assign sites to preliminary disturbance categories.** Each of the seven variables were scored based on their value in relation to the thresholds in Table 4. For example, if a site had a IWI of 0.9, it received an IWI score of +3; or if it had an IWI score of 0.55, it got an IWI score of -2. Each metric was scored like this. Then the scoring criteria in Table 5 were applied to obtain the preliminary disturbance category assignments for each site. Sites were then mapped and color-coded by disturbance category to ensure that their spatial distribution matched with expectations (Table 6 and Figure 8).
- **MassDEP biologists review the disturbance category assignments.** MassDEP staff were provided with a MS Excel 'disturbance worksheet' that included preliminary designations, metric scores, raw metric values, site information (like ecoregion), and secondary considerations (such as the NHDPlusV2 FTYPE/FCODE, habitat and water quality data, and HDI scores). In some cases, MassDEP staff changed designations based on local knowledge or other information that were not available in the GIS data. The following components of the site review process were documented in the spreadsheet: 1) whether MassDEP staff confirmed or changed the designations; and 2) MassDEP comments explaining rationale for changing designations.

- Strong for the axis
- Not redundant
  Spearman |r| < 0.8-0.9
- Meaningful
- Stressor-based

All Sites	Urban	Ag	Natural	Dams	Natural	
Variable	PC1	PC2	PC3	PC4	PC5	Select
ICI	**0.13**	0.09	-0.08	0.04	-0.11	**PC1**
IWI	**0.15**	0.06	-0.06	-0.05	0.03	**PC1**
PctUrbLMH2011Ws	**-0.16**	0.00	0.02	0.08	-0.06	**PC1**
RdDensCat	**-0.15**	-0.03	0.05	-0.03	0.05	**PC1**
PctHayCrop2011Cat	0.05	**-0.19**	-0.10	0.01	0.01	**PC2**
AllAgNWs	0.03	**-0.19**	-0.12	-0.01	0.07	**PC2**
DamNrmStorWs	-0.04	-0.04	0.04	**-0.29**	-0.08	**PC4**

Coastal	Urban	Ag	Ag Nutrients	Dams	Natural	
CostlVariable	PC1	PC2	PC3	PC4	PC5	Select
ICI	**0.15**	0.09	0.00	0.06	-0.06	PC1
IWI	**0.16**	0.06	0.02	-0.04	0.05	PC1
PctUrbLMH2011Ws	**-0.16**	0.02	-0.04	0.05	-0.08	PC1
RdDensCat	**-0.15**	-0.03	0.02	-0.05	0.04	PC1
PctHayCrop2011Cat	0.08	**-0.15**	-0.13	0.00	-0.06	PC2
AllAgNWs	0.07	**-0.15**	**-0.15**	-0.04	-0.01	PC3
DamNrmStorWs	-0.04	-0.08	0.03	**-0.21**	0.07	PC4

Highlands	Urban	Ag	Dams & natural	Dams & natural	Urban	
HighVariable	PC1	PC2	PC3	PC4	PC5	Select
ICI	**-0.14**	0.00	-0.07	0.03	**-0.15**	PC1
IWI	**-0.15**	0.02	-0.03	-0.05	0.02	PC1
PctUrbLMH2011Ws	**0.13**	-0.06	0.09	-0.01	-0.02	PC1
RdDensCat	**0.12**	-0.08	0.00	-0.03	0.10	PC1
PctHayCrop2011Cat	0.10	**0.13**	0.00	0.07	-0.07	PC2
AllAgNWs	0.09	**0.14**	0.07	0.13	-0.08	PC2
DamNrmStorWs	0.06	-0.12	0.01	**-0.10**	-0.12	PC4

Figure 7. Results from the Principal Components Analysis (PCA) helped inform variable selection when developing the disturbance gradient. This output is limited to the selected disturbance variables (the full output includes a much longer list of variables). See Table 3 for code descriptions.

Table 3. The following seven metrics were selected to characterize site disturbance in Massachusetts.

Category	Variable	Description	Scoring
Overall watershed condition	ICI	Index of catchment integrity (Thornbrugh et al. 2018)	Higher score = less disturbance
	IWI	Index of watershed integrity (Thornbrugh et al. 2018)	
Urban	PctUrbLMH2011Ws	% of watershed area classified as developed, high + medium + low-intensity land use (NLCD 2011 class 24+23+22)	Higher value = more disturbance
	RdDensCat	Density of roads (2010 Census Tiger Lines) within catchment (km/square km)	
Dam	DamNrmStorWs	Volume all reservoirs (NID_STORA in NID) per unit area of watershed (cubic meters/square km)	
Ag	PctHayCrop2011Cat	% of catchment area classified as hay and crop land use (NLCD 2011 class 82+81)	
	AllAgNWs	[CBNFWs]+[FertWs]+[ManureWs]*	

*CBNFWs = Mean rate of biological nitrogen fixation from the cultivation of crops in kg N/ha/yr, within watershed

*FertWs = Mean rate of synthetic nitrogen fertilizer application to agricultural land in kg N/ha/yr, within watershed

*ManureWs = Mean rate of manure application to agricultural land from confined animal feeding operations in kg N/ha/yr, within watershed

Table 4. Thresholds and scoring scheme (in parentheses) that were used to define the disturbance gradient. See Table 3 for code descriptions

Category (score)	IWI v.1	ICI v.1	PctUrbLMH 2011Ws	PctHayCrop 2011Cat	AllAgNWs	RdDensCat	DamNrmStorWs
Disturb Level 1 (least disturbed) (+3)	≥0.875	≥0.875	≤1%	≤1%	≤0.5	≤1.5	≤0.1
Disturb Level 2 (+2)	≥0.85	≥0.85	≤2%	≤2%	≤1	≤2	≤1,000
Disturb Level 3 (+1)	≥0.80	≥0.80	≤5%	≤5%	≤2.5	≤3	≤10,000
Disturb Level 4 (0)	>0.75 and <0.80	>0.75 and <0.80	>5 and <10%	>5 and <10%	>2.5 and <5	>3 and <5	>10,000 and <50,000
Disturb Level 5 (-1)	≤0.75	≤0.75	≥10%	≥10%	≥5	≥5	≥50,000
Disturb Level 6 (-2)	≤0.60	≤0.60	≥40%	≥15%	≥7.5	≥7.5	≥100,000
Disturb Level 7 (most disturbed) (-3)	≤0.50	≤0.50	≥60%	≥20%	≥10	≥10	≥200,000

Table 5. Sites were placed into disturbance categories based on the criteria below.

Disturbance category	Scoring criteria (based on scores for the seven metrics)
Best Reference	Minimum (min) score >= 2
Reference	Min score = 1
Sub Reference	All but 1 or 2 scores are > 0
Other	If other criteria are not met (min score = 0)
Some Stress	If min score = -1 OR count of negative (strs) scores < 2
Stress	If (min score < -1 AND >1 negative (strs) scores) OR (min score = -1 AND >3 negative (strs) scores)
High Stress	If >3 negative (strs) scores AND min score = -3

Table 6. Distribution of sites across disturbance categories and Level III ecoregions (Highlands = eco 58; Coastal = eco 59).

Category	Number of sites	
	MA Highlands	MA Coastal
Best Reference	12	5
Reference	45	9
Sub Reference	41	17
Other	23	22
Some Stress	57	126
Stress	54	159
High Stress	10	75
Total Reference*	***98***	***31***
Total Stress*	***121***	***360***

*Reference = Best Reference + Reference + Sub Reference

**Stressed = Some Stress + Stress + High Stress

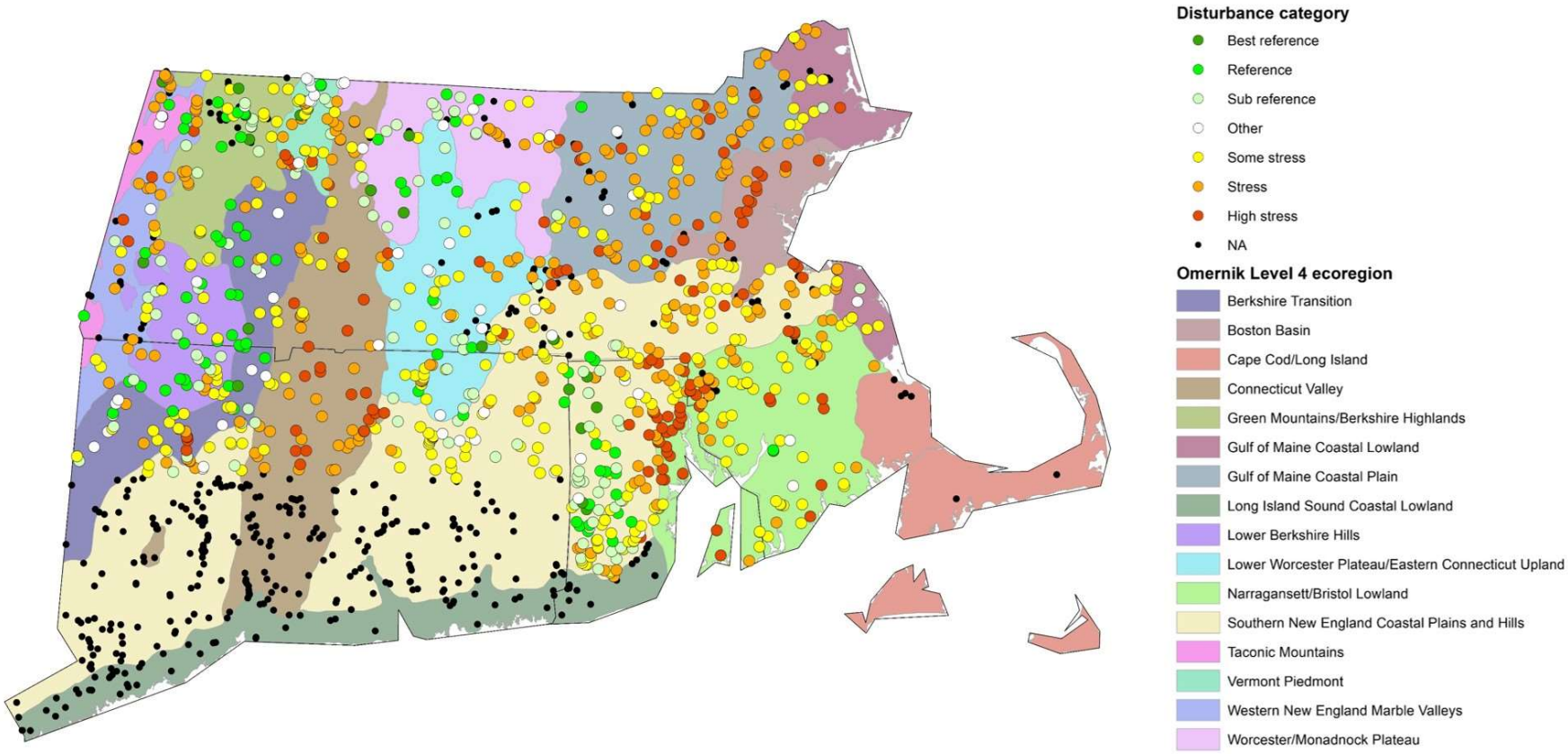


Figure 8. Spatial distribution of sites with valid samples in the regional (MA/CT/RI) dataset. Sites are color-coded by disturbance category and overlaid on Omernik Level IV ecoregions.

4 Classification

The purpose of classification is to develop classes that will minimize natural variability among sites, which will generate more robust relationships between biological conditions and human disturbance. Classification can be a balancing act. With too few classes, it may be difficult to distinguish between natural stream variability and human induced variability. With too many classes, there may be too many sites to monitor (which spreads program's resources too thin).

EPA (Omernik) Level III and IV ecoregions were used as a starting point for classification in MA because they incorporate important classification characteristics such as topography, soils, vegetation, elevation, latitude, longitude and more. In addition, they are easy to communicate because they are regional units of the landscape that are generally understood to have distinct characteristics that are important to aquatic organisms. Level III and IV ecoregions have been used for classifying stream biota in several statewide assessment programs (Weigel 2003, Gerritsen et al. 2000, Van Sickle and Hughes 2000, Barbour et al. 1996, Bryce and Clarke 1996, Rohm et al. 1987).

In MA, there are three Level III ecoregions: Northeastern Highlands (code 58), Northeastern Coastal Zone (code 59) and Atlantic Coastal Pine Barrens (code 84) (Figure 1. The Atlantic Coastal Pine Barrens Level III ecoregion was excluded from the analyses because it lacked sufficient numbers of sites. Within the Northeastern Highlands and Northeastern Coastal Zone, patterns in data from various combinations of Level IV ecoregions were explored to look for indications that the macroinvertebrate assemblages had distinct characteristics warranting the establishment of different stream classes. Initially the CT and RI data were included in the classification analyses, along with MassDEP multihabitat samples. However, in the end, only the MA kick net samples were considered, and analyses were limited to reference sites only (BestRef, Ref, SubRef) (Section 3).

To begin, Non-metric Multidimensional Scaling (NMS) ordinations of taxa and metrics were performed. NMS allows for visualization of patterns in taxonomic composition related to grouping variables such as entity, collection method, ecoregions, basin, sampling year and month, and more. Box plots showing metric distributions in Level III and IV ecoregions were also generated. Figure 9 shows results from a NMS ordination in which samples were color-coded by Level III ecoregion. The dataset included CT, RI and MA reference samples. Samples generally formed distinct groups, suggesting that macroinvertebrate assemblages in the two Level III ecoregions were different enough to consider breaking out into separate classes. Box plots like the one shown in Figure 10 (number of EPT taxa) helped corroborate and refine this pattern. Overall, metrics like number of EPT taxa had noticeable differences in distributions across the Northeastern Highlands and Northeastern Coastal Zone. There was an exception - the Worcester/Monadnock Plateau Level IV ecoregion (58g) – which had lower mean EPT richness than the other Northeastern Highlands Level IV ecoregions. This pattern (in which ecoregion 58g samples grouped more closely with Coastal Zone vs. Northeastern Highland samples) held true with several other metrics as well. Thus, ecoregion 58g was grouped with Coastal Zone sites for classification.

Other patterns were less clear in the data. For example, the Southern New England Coastal Plains and Hills (59c) was difficult to classify. It showed similarities with ecoregions in the central part of the state, as well as with areas in the southeastern part of MA ("Eastern Coastal Zone"), and had limited numbers of reference sites to confirm this pattern. At one point the following three classes were being considered: 1) Western Highlands = Northeastern Highlands Level III ecoregion minus the Worcester/Monadnock Plateau Level IV ecoregion (58g); 2) Eastern Coastal Zone = Southern New England Coastal Plains and Hills (59c), Boston Basin (59d), Gulf of Maine Coastal Lowland (59f) and Narragansett/Bristol Lowland (59e); and 3) Central Hills = Worcester/Monadnock Plateau (58g)

Connecticut Valley (59a), Lower Worcester Plateau/Eastern Connecticut Upland (59b) and Gulf of Maine Coastal Plain (59h). In the end, after running some additional analyses (including a PCA and Classification and Regression Tree (CART) Analysis) and consulting with MassDEP staff, two classes were defined: Western Highlands and Central Hills. Table 7 lists the combinations of Level IV ecoregions that comprise each class, and Figures 11 & 12 show the delineations of the classes. The two Level IV ecoregions in southeastern MA (Narragansett/Bristol Lowland (59e) and Cape Cod/Long Island (84a) were left unassessed in this phase of work due to insufficient RBP kick net data.

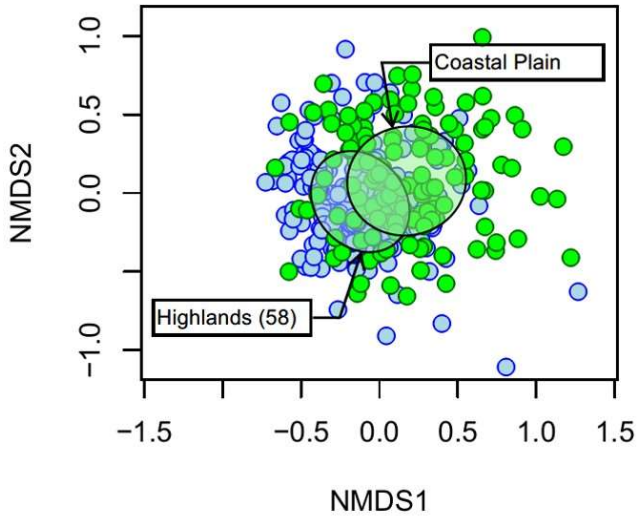


Figure 9. Non-metric Multidimensional Scaling (NMS) ordinations were used to evaluate patterns in taxonomic composition related to EPA Level III ecoregions. This ordination is based on the regional reference site, presence-absence dataset (MA/CT/RI) (n=312) (Section 3).

Table 7. The two macroinvertebrate stream classes (Central Hills and Western Highlands) are comprised of combinations of Level IV ecoregions. The Narragansett/Bristol Lowland and Cape Cod/Long Island ecoregions were left 'unassessed' due to limited numbers of sites and unique features. The Worcester/Monadnock Plateau ecoregion (58g) was the only Northeastern Highland Level IV ecoregion that was grouped with the Central Hills.

Class	Level IV code	Level IV ecoregion name
Central Hills	58g	Worcester/Monadnock Plateau
	59a	Connecticut Valley
	59b	Lower Worcester Plateau/Eastern Connecticut Upland
	59c	Southern New England Coastal Plains and Hills
	59d	Boston Basin
	59f	Gulf of Maine Coastal Lowland
	59h	Gulf of Maine Coastal Plain
Western Highlands	58a	Taconic Mountains
	58b	Western New England Marble Valleys
	58c	Green Mountains/Berkshire Highlands
	58d	Lower Berkshire Hills
	58e	Berkshire Transition
	58f	Vermont Piedmont
Unassessed	59e	Narragansett/Bristol Lowland
	84a	Cape Cod/Long Island

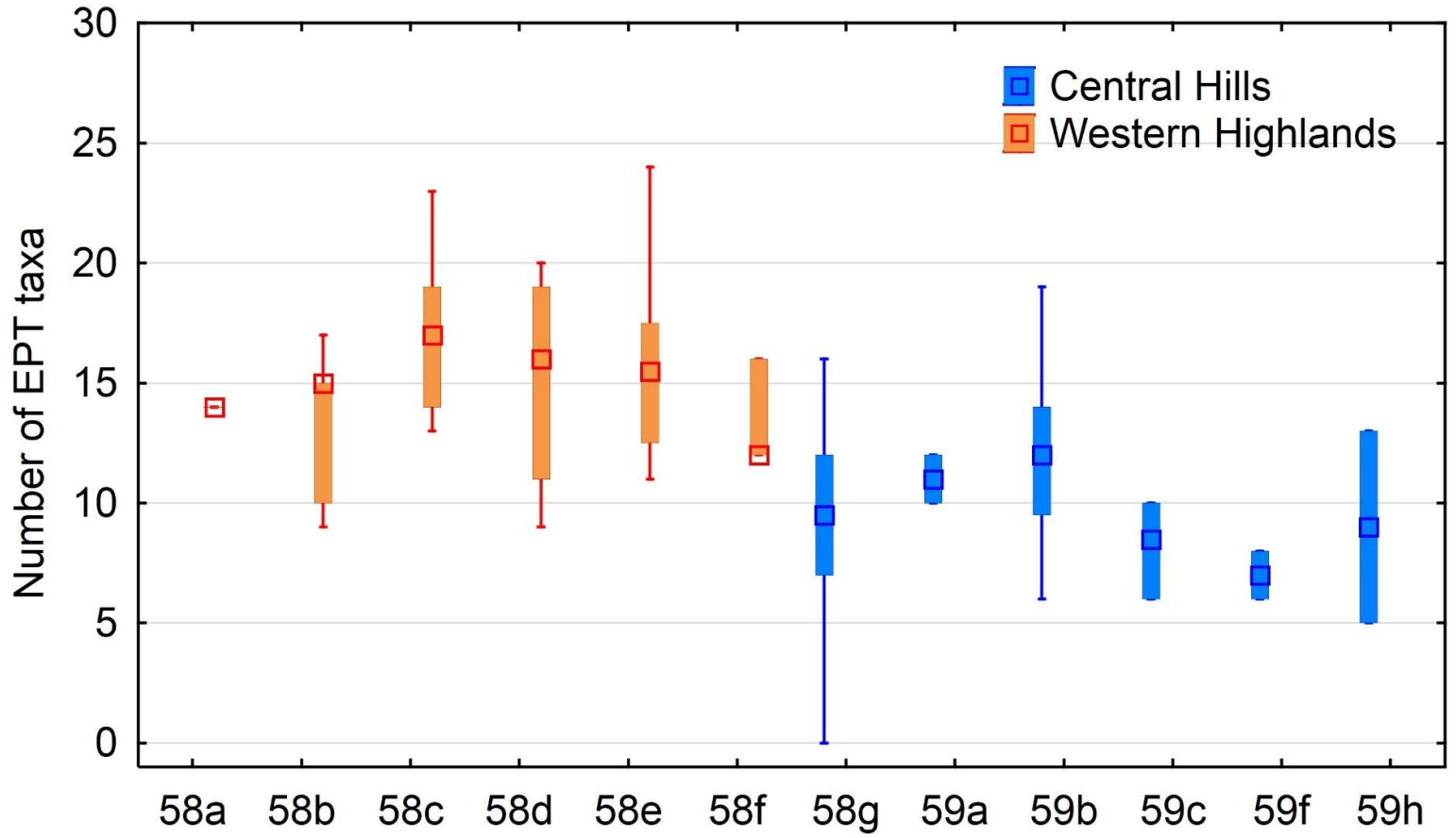


Figure 10. Box plot showing the distribution of EPT richness metric values in reference sites across Level IV ecoregions (see Table 7 for code descriptions). This is one of the plots that supported our decision to group 58g (Worcester/Monadnock Plateau) with the Northeastern Coastal Zone. This plot was generated with the Massachusetts kick net data from the calibration/validation dataset collected from 2000 onward.

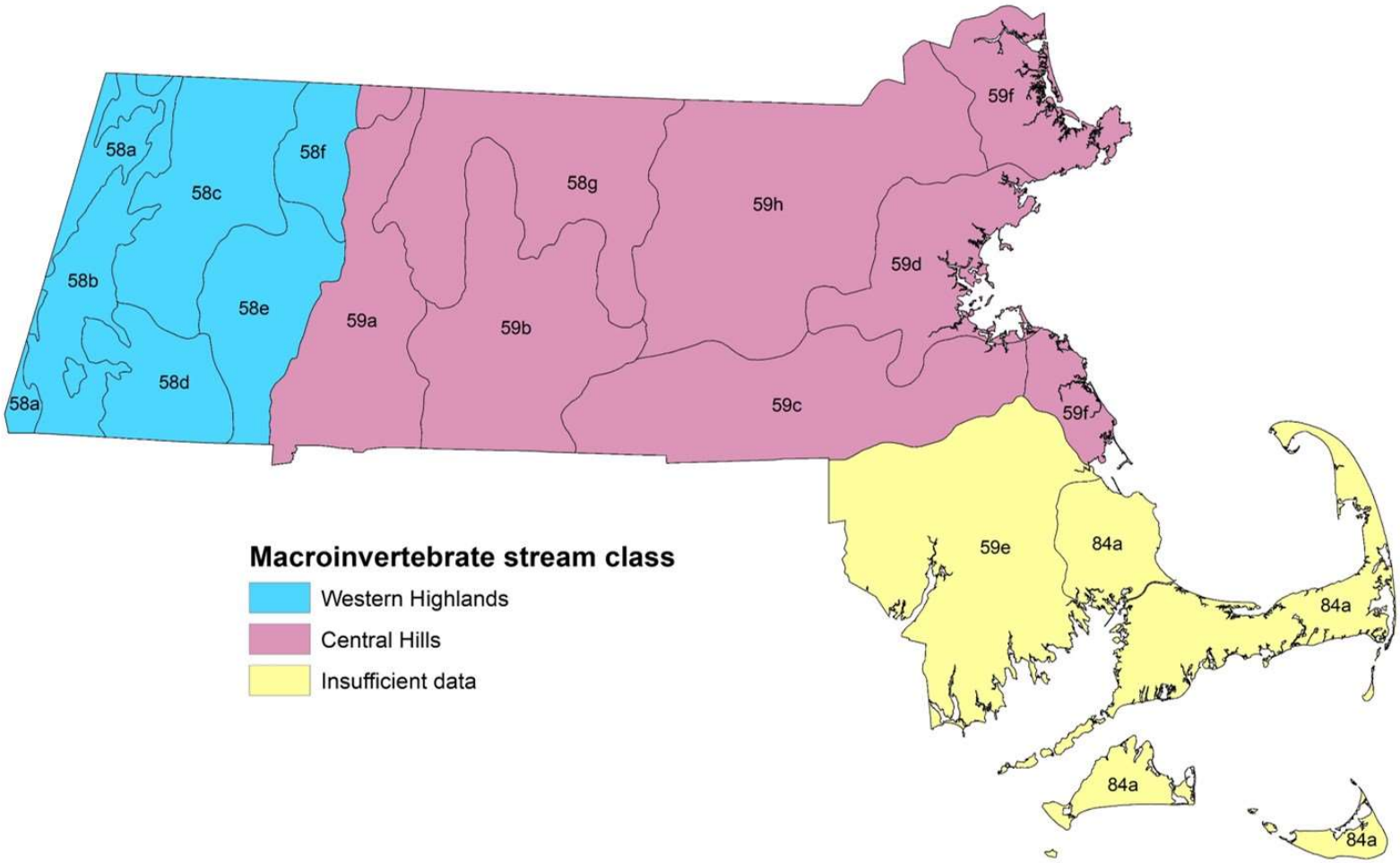


Figure 11. For IBI development, Omernik Level IV ecoregions were grouped into two stream classes: Western Highlands and Central Hills. See Table 7 for code descriptions.

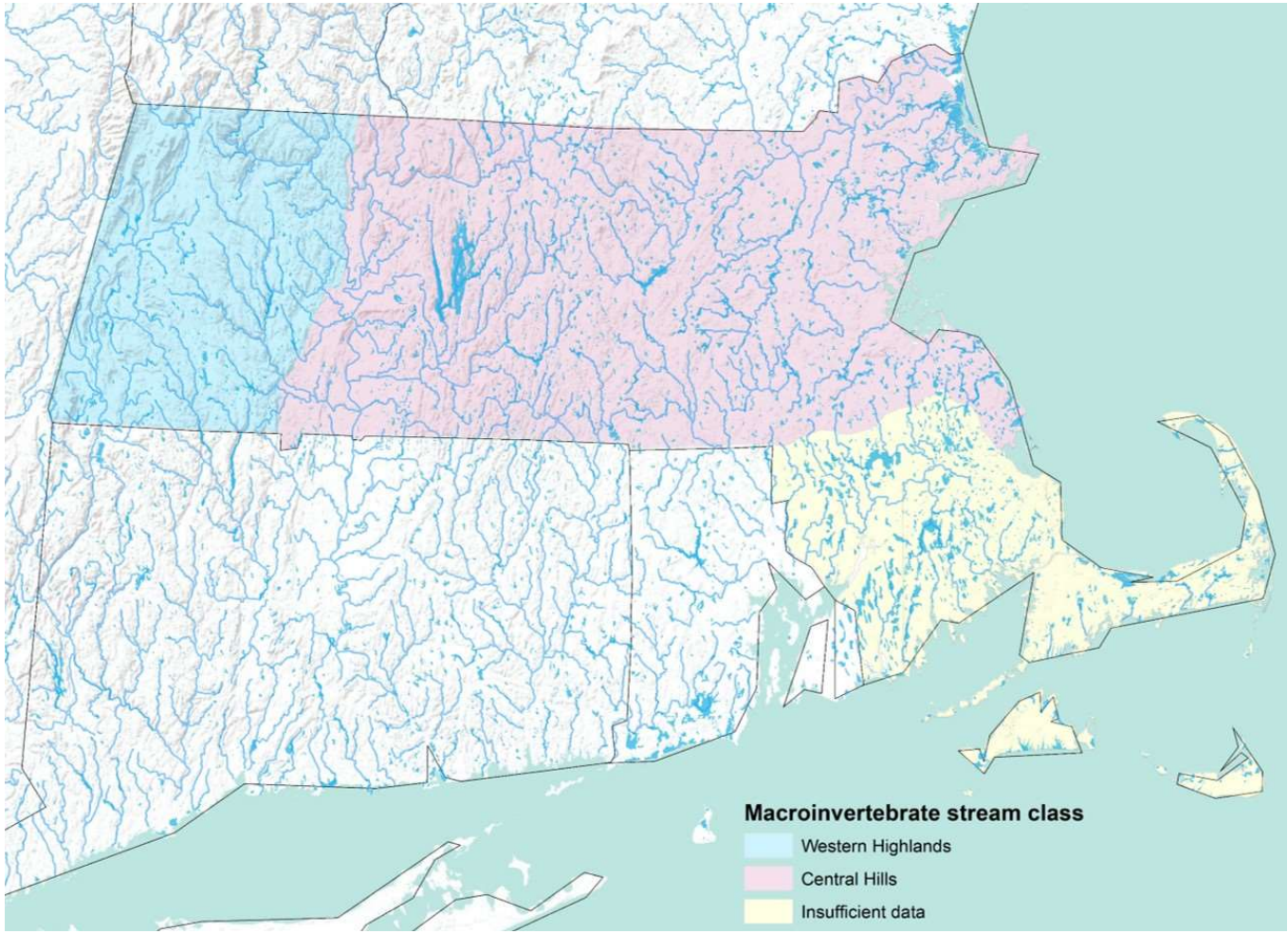


Figure 12. There are differences in topography and hydrology across the two macroinvertebrate stream classes. The Western Highlands have steeper, more complex terrain that includes the Berkshire Mountains. The Central Hills has more low gradient streams, wetlands and lakes, including the Quabbin Reservoir.

5 Index Development

Index development consisted of four main steps:

- Selection of samples for IBI development
- Metric scoring and selection
- Index compilations and performance evaluation
- Selection of final IBIs

5.1 Site selection for calibration and verification

The following criteria were used to select samples for IBI development:

- Geographic location – in the Western Highlands (WH) or Central Hills (CH)
- Entity and collection method – MassDEP and DRWA kick net samples only
- Sampling years – 2000-2017[4]
- Sampling months - July through September
- One sample per site

After the screening criteria were applied, the dataset consisted of 274 samples in the Central Hills and 170 in the Western Highlands. Samples were grouped into reference and stressed categories. In the Central Hills class, where there were higher levels of disturbance (Figure 13), reference was defined as BestRef + Ref + SubRef (Section 3), which provided 42 reference sites. All the stressed sites were taken from the High Strs group (n=44). In the Western Highlands, reference was defined as BestRef + Ref (n=38) and stressed sites were taken from both the Strs + High Strs groups (n=56). After samples were assigned to reference and stressed groups, the samples were randomly assigned to calibration and verification datasets. Some of the DRWA samples were moved from the calibration to the verification dataset so that the Deerfield River watershed would not have a biased influence on metric performance in the Western Highlands. The target was to have 20% of the samples in the validation dataset (at a minimum, 10%). Only the calibration samples were used for IBI development. Table 8 shows the distribution of sites across class, disturbance category and dataset (calibration vs. verification). Figures 14 & 15 show how the reference and stressed sites in the calibration and verification datasets are spatially distributed across the landscape. Appendix E contains a list of the reference and stressed sites that were used for IBI development.

Table 8. Distribution of IBI calibration and verification samples across class and disturbance categories.

Class	Dataset	Disturbance sub-category					Total	
		BestRef	Ref	SubRef	Strs	High Strs	Reference	Stressed
Western Highlands	Calibration	5	21	--	35	7	26	42
	Verification	2	10	--	11	3	12	14
	Total	*7*	*31*	--	*46*	*10*	*38*	*56*
Central Hills	Calibration	3	10	17	--	30	30	30
	Verification	1	3	8	--	14	12	14
	Total	4	13	25	--	44	42	44

[4]Prior to 2000, some taxonomic standards were slightly different. To reduce variability associated with these differences, we limited the calibration/verification dataset to 2000-2017.

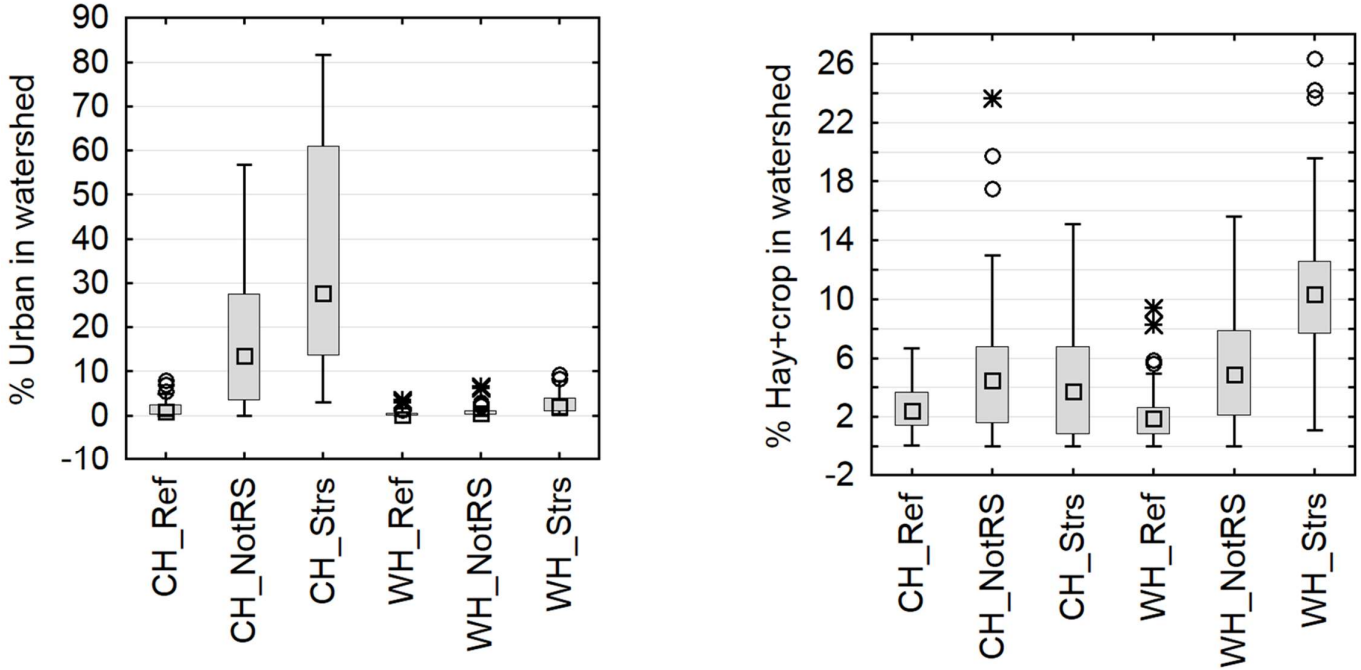


Figure 13. The Central Hills (CH) dataset has higher levels of urban disturbance than the Western Highlands (WH). The WH has slightly higher distrubance from agricultural land use, but this number is still relatively low (<30%). Codes are as follows: _Ref = reference dataset, _NotRS = not Reference or Stressed, _Strs = stressed

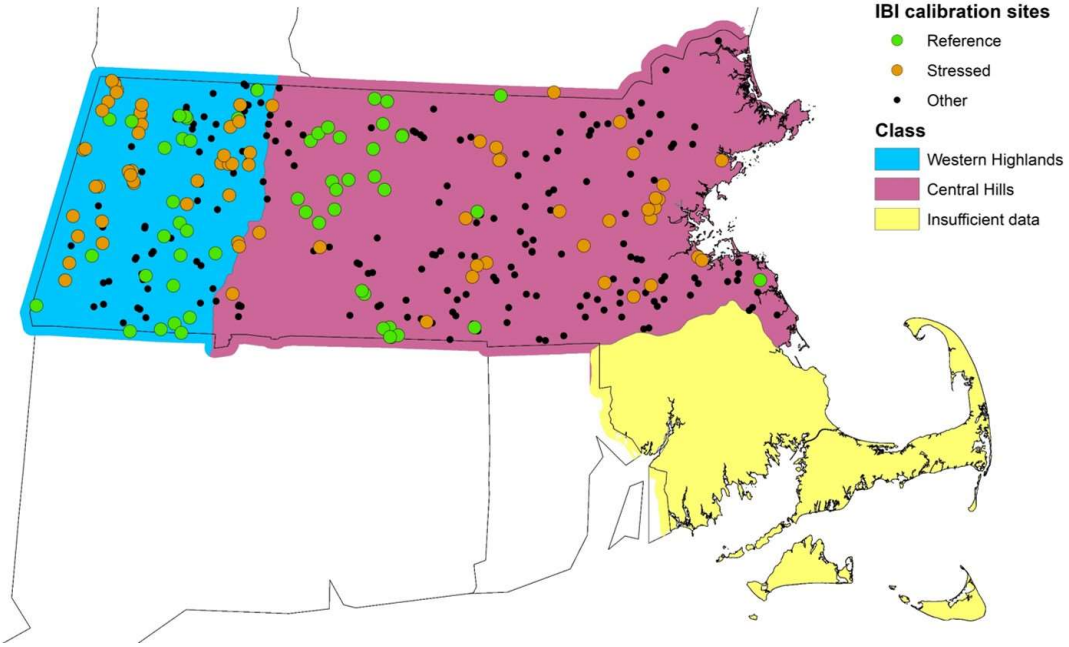


Figure 14. Locations of sites that were used to calibrate the CH and WH IBIs. The sites are color-coded by disturbance category (Reference (Ref) or Stressed (Strs)), as defined in Section 5.1.

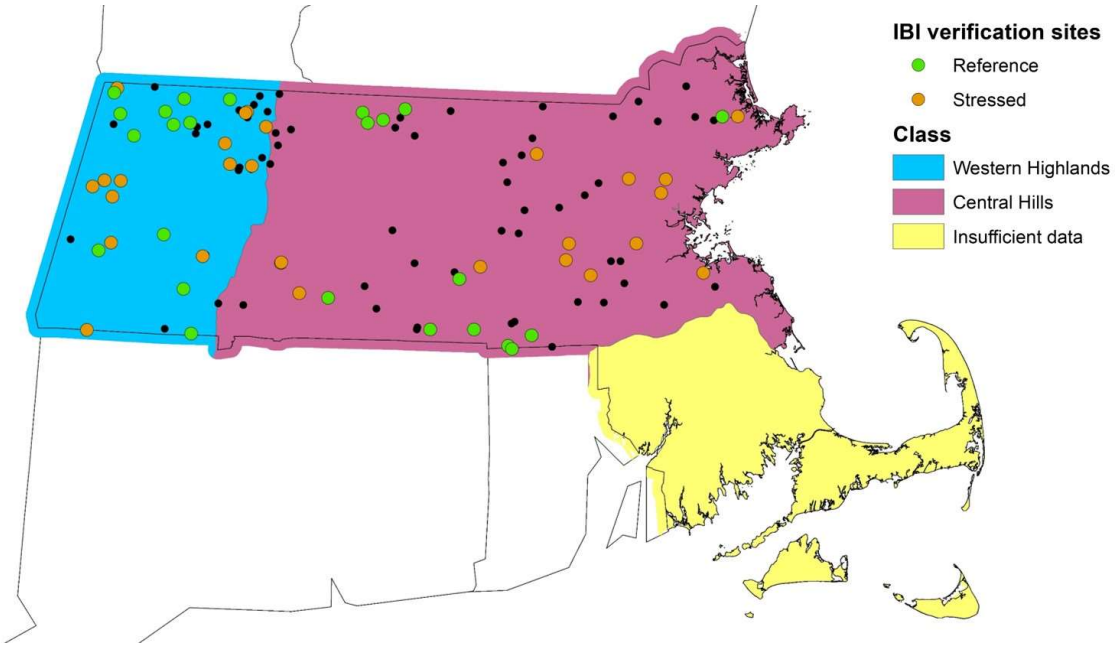


Figure 15. Locations of sites that were used to verify the CH and WH IBIs. The sites are color-coded by disturbance category (as defined in Section 5.1).

5.2 Metric Scoring and Selection

Evaluation and selection of metrics typically involves testing of many more metrics than end up going into the final index. Tt calculated and evaluated 100 metrics (Appendix B), which were grouped into the following metric categories: richness, composition, tolerance, functional attribute, habit, thermal preference and life cycle. Metrics were then evaluated for the following -

- Sensitivity
    - How well does the metric distinguish between reference and stressed sites?
    - What is the relationship between the metric and the disturbance variables?
        - Direction of response
        - Strength/significance
- Redundancy
- Representation across metric categories
- Metrics currently being used in MA's RBP III index
- Precision

Metric scoring
Before these evaluations were done, metric scores were calculated because metrics are mostly on different scales and thus cannot be directly aggregated. To address this, formulas were applied to the metrics to convert them to a 0-100-point scoring scale (as in Hughes et al. 1998, and Barbour et al. 1999). The scoring scale was based on the distribution of metric values across all sites (versus reference sites only).

For metrics that decreased with increasing stress (referred to as 'decreasers'; an example is the number of intolerant taxa metric), we used the following equation, in which the 95th percentile was the upper end of the scoring scale and the minimum possible value (0) was the lower end:

'Decreaser' metric score = 100*(Metric value – minimum)/(95th percentile-minimum)

Scores for metrics that increased with stress (referred to as 'increasers'; an example is the number of tolerant taxa metric) were calculated with this equation:

'Increaser' metric score = 100*(95th percentile-metric value)/(95th percentile-5th percentile)

Appendix F contains more detailed information on the selection of the scoring formulas.

Metric scoring adjustments
Relationships between metric values and natural variables were evaluated within the classes to see whether adjustments were needed for metric scoring. These analyses were limited to the reference datasets only. Scatterplots and Spearman rank correlation analyses were used to look for patterns. The correlated environmental variables that were tested included latitude, longitude, watershed area, stream slope, elevation, base flow index (BFI), modelled summer stream temperature, average air temperature, average precipitation, and collection date. This exploratory analysis showed that there were metrics in both site classes that were correlated at Spearman $|r| > 0.50$. The variables that were correlated included collection date, modeled temperature, average temperature, watershed area, elevation, latitude, and BFI.

Scatter plots were examined to confirm correlation patterns in reference and all sites. The patterns were generally un-convincing because of one or more reasons, such as: they appear to be driven by a few points, they were not consistent among similar variables (e.g., among temperature measures),

they were not confirmed or understandable in non-reference sites, or the environmental variable was poorly understood or had a short range of conditions. Adjustments were attempted when there was a consistent slope or wedge for at least a portion of the gradient, the ecological mechanism was interpretable, reference and non-reference gave similar signals, and similar classification variables gave similar signals.

In the Western Highlands, several adjustments were considered. However, in consultation with MassDEP biologists, none of the adjustments were convincing and none were used in index trials. In the Central Hills, only collection date was correlated with Plecoptera and sprawler metrics. These relationships were also weak and unconvincing, so they were not used.

Metric selection

The ability of each metric to distinguish between reference and stressed sites within a site class was measured as discrimination efficiency (DE) (Flotemersch et al. 2006, Maxted et al. 2000, Ofenböck et al. 2004). The simplest distinction between reference and stressed sites is shown with box plots that show several attributes of the distribution graphically: median, upper and lower quartiles, tails, outliers and/or minimum and maximum (Barbour et al. 1999). The box plots show exactly how much the distributions differ or overlap. The DE is a quantification of the visually apparent distinctions. DE was calculated as the percentage of metric scores in stressed sites that were worse than the worst quartile of those in the reference sites. For metrics with a pattern of decreasing value with increasing environmental stress, DE is the percentage of stressed values below the 25th percentile of reference site values. For metrics that increase with increasing stress, DE is the percentage of stressed sites that have values higher than the 75th percentile of reference values. DE can be visualized on box plots of reference and stressed metric or index values with the inter-quartile range plotted as the box (Figure 16). Higher DE denotes more frequent correct association of metric values with site conditions. DE values ≤25% show no discriminatory ability in one direction. Metrics with DE values ≥50% were generally considered for inclusion in an index. However, in a site class, metric selection was usually dependent on relative DE values within a metric category.

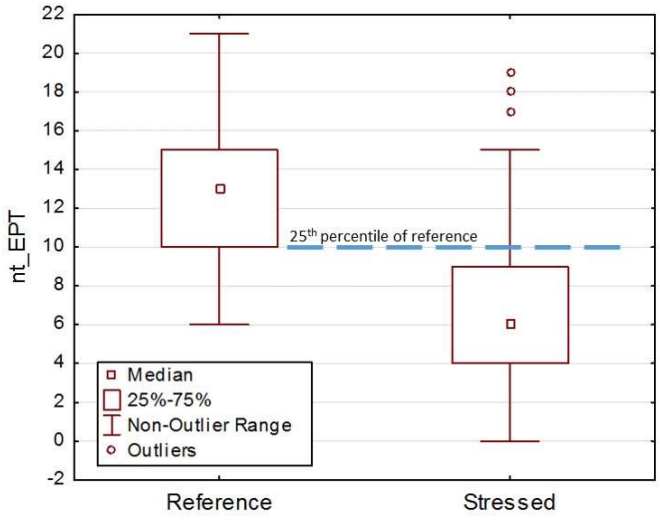


*Figure 16. Discrimination efficiency (DE). In this example, which uses # EPT taxa (a metric that decreases with stress), the 25th percentile of the reference distribution is used as the standard (and we calculate what % of stressed sites were below that threshold). If it was a metric that increased with stress, we would have used the 75th percentile of the reference distribution as the standard (and calculated what % of stressed sites were above that threshold). The formula is: DE = a/b*100, where a = number of a priori stressed sites identified as being below the degradation threshold (in this example, 25th percentile of the LD site distribution) and b = total number of stressed sites. The higher the DE, the better (the more frequent the correct association of metric values with site conditions).*

A second measure of metric discrimination was the *Z*-score, which was calculated as the difference between reference and stressed metric or index values divided by the standard deviation of reference values. The *Z*-score is similar to Cohen's D (Cohen 1992) and gives a combined measure of index sensitivity and precision. There is no absolute *Z*-score value that indicates adequate metric performance, but among metrics or indices, higher *Z*-scores suggest better separation of reference and stressed values. Cohen proposed that Z values ≥ 0.80 indicated a "large" effect.

The DE and *Z*-scores epitomize the difference in distributions at critical potential threshold levels and incorporate precision of the reference distribution. They were used in favor of a t-test or signal:noise (S:N) ratio. The DE is an estimate of the percentage of correct impaired assessments and can be interpreted for management applications. While the *t*-test has been used elsewhere (Stoddard et al. 2008), we did not use it because we are not testing a hypothesis about the difference between reference and stressed sites. The *Z*-score and S:N ratio are similar measures of responsiveness as a function of variability.

Table 9 contains a list of the metrics that had the best performance (showed sensitivity and consistency among site classes) within each metric category and were selected to be tested in the index compilations (Section 5.4). Metrics that performed well in one class, but showed opposite response in the other class, were sidelined in favor of more dependable responses. In the WH, the best performing metrics had DE of 67%. These included percent Plecoptera individuals and number of cold/cool indicator taxa. Metrics with DE > 50% were represented from each metric category. In the Central Hills, the best performing metrics had DE > 95%, including percent Plecoptera individuals, percent intolerant individuals, and number of predator taxa. Metrics with DE > 80% were represented from each metric category except habit.

There were some response patterns that are worth noting because they were unusual or difficult to explain. Thermal metrics appeared to perform as well as tolerance metrics. The thermal traits were developed to describe taxa associated with the thermal gradient. The thermal gradient might be a natural gradient and is therefore not an appropriate indicator of human disturbance. Taxa that are sensitive to thermal gradient might also be sensitive to organic or other types of pollution. Because the tolerance metrics were available and were calibrated to pollution tolerance (not a natural gradient), the tolerance metrics were preferred over thermal metrics in index development.

Chironomids and Diptera metrics decrease with increasing stress in the Massachusetts data sets. These groups are typically thought of as tolerant of stressful conditions, which would imply that they should increase with increasing stress. For example, the Rapid Bioassessment protocols (RBP; Barbour et al. 1999), recommends using the increase in Chironomids as an indicator of increasing stress. The taxonomic groups are diverse and include some sensitive taxa, but the decreasing trend is uncommon. This leads to less confidence in the metric responses and avoidance during metric selection.

The Habit metrics are not easily explained in that the clingers increase with increasing stress and the sprawler are strong decreasers. Clingers typically decrease with stress because they indicate turbulent flowing waters that are common in high-gradient streams. Human disturbance might be more prevalent in lower gradient landscapes where flow is less steep, and the clingers decrease. This expected pattern is not evident in the Massachusetts data sets and the metric response is difficult to explain. The sprawlers include several Chironomid taxa, that, as mentioned previously, also decrease with stress with a poorly understood mechanism.

The EPT metrics are typically reliable decreasers with increasing stress. In the MA data set they are not strong indicators of stress, except for the Plecoptera, which were the strongest metrics in both site classes. While EPT are typically sensitive organisms, there are also some unclear and tolerant taxa in the group. For example, the net-spinning Hydropsychid caddisflies can increase with increasing nutrients. The Plecoptera are dependent on cool, well-aerated waters and are the most sensitive of the insect orders. In the Central Hills, they might only be expected in the higher-gradient hills, regardless of human disturbances.

The list of metrics in Table 9 was further refined for each class by checking for redundancy (top candidate metrics should not be redundant with each other). Spearman correlation analyses were performed on all pairwise combinations of candidate metrics within each stream class. Metric pairs with Spearman $|r| \geq 0.85$ were considered redundant and were not both used in any index alternative. Metrics correlated at Spearman $|r| \geq 0.75$ were evaluated for possible exclusion. The mean redundancy among index metrics was calculated for final index selections

Tables 10 & 11 contain the refined list of top candidate metrics in the Central Hills and the associated correlation matrix, respectively. Tables 12 & 13 contain the list of top metrics and associated correlations in the Western Highlands. Appendix G contains more information on the candidate metrics and their performance in each class.

Table 9. Macroinvertebrate metric discrimination efficiency (DE), trend with increasing stress, and Z-score for metrics used in IBI development in each site class.

		Western Highlands				Central Hills			
Metric Category	Metric Code	Selected	DE	Trend with stress	*Z-score*	Selected	DE	Trend with stress	*Z-score*
RICH	nt_total	Yes	52.4	(dec)	0.66	Yes	66.7	(dec)	1.21
	nt_EPT	Yes	57.1	(dec)	0.60	Yes	70.0	(dec)	1.30
	pt_Pleco	Yes	59.5	(dec)	0.81	Yes	90.0	(dec)	1.50
	nt_Insect	Yes	42.9	(dec)	0.69	Yes	76.7	(dec)	1.31
	pt_EPT		40.5	(dec)	0.31	Yes	76.7	(dec)	1.20
	pi_dom04	Yes	33.3	(inc)	-0.44	Yes	60.0	(inc)	-0.61
COMP	pi_NonIns		45.2	(inc)	-0.34		56.7	(inc)	-0.48
	pi_Pleco	Yes	66.7	(dec)	1.03		100.0	(dec)	1.21
	pi_EphemNoCaeBae		35.7	(inc)	0.00	Yes	66.7	(dec)	0.64
FFG	nt_ffg_pred	Yes	45.2	(dec)	0.60	Yes	96.7	(dec)	1.81
	pi_ffg_filt	Yes	50.0	(inc)	-0.69	Yes	76.7	(inc)	-1.78
	pi_ffg_pred	Yes	40.5	(dec)	0.62	Yes	80.0	(dec)	1.36
	pt_ffg_pred	Yes	42.9	(dec)	0.49	Yes	90.0	(dec)	1.81
	pi_ffg_shred	Yes	61.9	(dec)	0.84		53.3	(dec)	0.68
TOLER	nt_tv_intol	Yes	57.1	(dec)	1.02	Yes	90.0	(dec)	2.08
	x_Becks	Yes	57.1	(dec)	1.11	Yes	86.7	(dec)	1.83
	x_HBI	Yes	42.9	(inc)	-0.58	Yes	83.3	(inc)	-0.89
	pt_tv_intol	Yes	50.0	(dec)	0.85	Yes	100.0	(dec)	2.74
	pi_tv_intol	Yes	59.5	(dec)	0.83	Yes	93.3	(dec)	1.88
HABIT	nt_habit_sprawl	Yes	47.6	(dec)	0.61		43.3	(dec)	0.67
	pi_habit_cling	Yes	54.8	(inc)	-0.75		46.7	(inc)	-0.50
	pi_habit_sprawl	Yes	54.8	(dec)	0.57		50.0	(dec)	0.48
	nt_habit_cling		31.0	(dec)	0.18	Yes	73.3	(dec)	1.05
VOLT	nt_volt_semi	Yes	57.1	(dec)	0.58	Yes	80.0	(dec)	1.37
	pt_volt_semi		52.4	(dec)	0.44	Yes	76.7	(dec)	1.39
	nt_volt_uni		31.0	(dec)	0.10	Yes	63.3	(dec)	1.02

Table 10. Top candidate metrics in the Central Hills index. The scoring formula for 'decreaser' metrics = 100(Metric value – minimum possible value)/(95th percentile-minimum) and the formula for 'increaser' metrics = 100*(95th percentile-metric value)/(95th percentile-5th percentile). The minimum possible value for these metrics is 0. To simplify the formulas, the 0's in the 'decreaser' formulas are not shown. All values that calculate to < 0 or >100 are re-set to the 0-100 scale.*

Metric abbreviation	Metric description	Category	5th	95th	Scoring formula	DE	Trend
nt_total	Number of taxa - total	RICH	11	34.9	100*(metric)/34.9	67	Dec.
nt_EPT	Number of taxa - Orders Ephemeroptera, Plecoptera & Trichoptera (EPT)	RICH	1.1	15	100*(metric)/15	70	Dec.
pt_EPT	Percent taxa - Orders EPT	RICH	11	54.5	100*(metric)/54.5	77	Dec.
nt_Insect	Number of taxa - Class Insecta	RICH	8	32	100*(metric)/32	77	Dec.
nt_Pleco	Number of taxa - Order Plecoptera	RICH	0	3	100*(metric)/3	90	Dec.
pi_dom04	Percent individuals - four most dominant taxa	RICH	33	84.6	100*(84.6-metric)/51.5	60	Inc.
pi_EphemNoCaeBae	Percent individuals - Order Ephemeroptera, excluding Families Caenidae and Baetidae	COMP	0	13.9	100*(metric)/13.9	67	Dec.
nt_ffg_pred	Number of taxa - Functional Feeding Group (FFG) - predator (PR)	FFG	0	8	100*(metric)/8	97	Dec.
pi_ffg_filt	Percent individuals - FFG - collector-filterer (CF)	FFG	13	79.9	100*(79.9-metric)/66.9	77	Inc.
pi_ffg_pred	Percent individuals - FFG - predator (PR)	FFG	0	17	100*(metric)/17.0	80	Dec.
pt_ffg_pred	Percent taxa - FFG - predator (PR)	FFG	0	28.5	100*(metric)/28.5	90	Dec.
nt_habit_cling	Number of taxa - Habit - clingers (CN)	HABIT	4	20	100*(metric)/20	73	Dec.
nt_tv_intol	Number of taxa - tolerance value - intolerant ≤ 3	TOLER	0	12	100*(metric)/12	90	Dec.
pi_tv_intol	Percent individuals - tolerance value - intolerant ≤ 3	TOLER	0	36.9	100*(metric)/36.9	93	Dec.
pt_tv_intol	Percent taxa - tolerance value - intolerant ≤ 3	TOLER	0	39.1	100*(metric)/39.1	100	Dec.
x_Becks	Becks Biotic Index	TOLER	1	25	100*(metric)/25	87	Dec.
x_HBI	Hilsenhoff Biotic Index	TOLER	3.8	6.07	100*(6.07-metric)/2.23	83	Inc.
nt_volt_semi	Number of taxa - semivoltine (SEMI)	VOLT	0	6.9	100*(metric)/6.90	80	Dec.
nt_volt_uni	Number of taxa - univoltine (UNI)	VOLT	2	11	100*(metric)/11	63	Dec.
pi_volt_semi	Percent individuals - semivoltine (SEMI)	VOLT	0	27.9	100*(metric)/27.9	73	Dec.

5^{th}: 5^{th} percentile of all sample metrics in the site class.
95^{th}: 95^{th} percentile of all sample metrics in the site class
DE: Discrimination Efficiency.
Scoring Formula: Replace "metric" with the sample metric value for calculation of an index
Trend: Decreasing (Dec.) or increasing (Inc.) trend with increasing stress

Table 11. Correlation (Spearman rho) among metrics of the top Central Hills macroinvertebrate index candidates.

		Metric #														
#	Metric	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	nt_total	1														
2	nt_Insect	0.97	1													
3	pt_EPT	0.31	0.44	1												
4	pt_Pleco	0.52	0.57	0.33	1											
5	pi_dom04	-0.85	-0.81	-0.27	-0.42	1										
6	pi_EphemNoCaeBae	0.49	0.56	0.58	0.33	-0.43	1									
7	nt_ffg_pred	0.71	0.70	0.09	0.63	-0.53	0.32	1								
8	pi_ffg_filt	-0.40	-0.38	0.05	-0.40	0.34	-0.09	-0.42	1							
9	pi_ffg_pred	0.65	0.64	0.03	0.56	-0.56	0.31	0.91	-0.43	1						
10	pt_ffg_pred	0.38	0.38	-0.04	0.52	-0.23	0.14	0.9	-0.33	0.81	1					
11	nt_habit_cling	0.83	0.88	0.63	0.45	-0.74	0.6	0.47	-0.19	0.42	0.17	1				
12	pi_tv_intol	0.69	0.76	0.51	0.64	-0.60	0.55	0.61	-0.43	0.56	0.41	0.67	1			
13	pt_tv_intol	0.62	0.71	0.60	0.70	-0.47	0.51	0.58	-0.40	0.5	0.41	0.66	0.89	1		
14	nt_volt_semi	0.68	0.74	0.43	0.58	-0.62	0.50	0.55	-0.31	0.49	0.34	0.77	0.70	0.70	1	
15	nt_volt_uni	0.75	0.79	0.54	0.37	-0.66	0.48	0.46	-0.27	0.38	0.21	0.78	0.61	0.60	0.56	1
16	pt_volt_semi	0.42	0.51	0.41	0.49	-0.41	0.41	0.38	-0.19	0.33	0.26	0.62	0.57	0.61	0.94	0.38

Table 12. Top candidate metrics in the Western Highlands index. The scoring formula for 'decreaser' metrics = 100(Metric value – minimum possible value)/(95th percentile-minimum) and the formula for 'increaser' metrics = 100*(95th percentile-metric value)/(95th percentile-5th percentile). The minimum possible value for these metrics is 0. To simplify the formulas, the 0's in the 'decreaser' formulas are not shown. All values that calculate to < 0 or >100 are re-set to the 0-100 scale before averaging in the index.*

Metric abbrev	Metric description	Category	5th	95th	Scoring formula	DE	Trend
nt_total	Number of taxa - total	RICH	21	38.8	100*(metric)/38.8	52.4	Dec.
nt_EPT	Number of taxa - Orders Ephemeroptera, Plecoptera & Trichoptera (EPT)	RICH	8	23	100*(metric)/23	57.1	Dec.
nt_Insect	Number of taxa - Class Insecta	RICH	20	37	100*(metric)/37	42.9	Dec.
pt_Pleco	Percent taxa - Order Plecoptera	RICH	0	21.1	100*(metric)/21.1	59.5	Dec.
x_Shan_2	Shannon Wiener Diversity Index (log base 2) - x_Shan_Num/log(2)	RICH	3.58	4.97	100*(metric)/4.97	42.9	Dec.
pi_Pleco	Percent individuals - Order Plecoptera	COMP	0	18.3	100*(metric)/18.3	66.7	Dec.
nt_ffg_pred	Number of taxa - Functional Feeding Group (FFG) - predator (PR)	FFG	2	11	100*(metric)/11	45.2	Dec.
pi_ffg_filt	Percent individuals - FFG - collector-filterer (CF)	FFG	9.76	50.5	100*(50.5-metric)/40.7	50	Inc.
pi_ffg_pred	Percent individuals - FFG - predator (PR)	FFG	2.29	23.8	100*(metric)/23.8	40.5	Dec.
pi_ffg_shred	Number of taxa - FFG - shredder (SH)	FFG	1.17	23	100*(metric)/23	61.9	Dec.
pt_ffg_pred	Percent taxa - FFG - predator (PR)	FFG	9.18	31.9	100*(metric)/31.9	42.9	Dec.
nt_habit_sprawl	Number of taxa - Habit - sprawlers (SP)	HABIT	2	11	100*(metric)/11	47.6	Dec.
pi_habit_cling	Percent individuals - Habit - clingers (CN)	HABIT	44.5	86.8	100*(86.8-metric)/42.3	54.8	Inc.
pi_habit_sprawl	Percent individuals - Habit - sprawlers (SP)	HABIT	2.23	35.9	100*(metric)/35.9	54.8	Dec.
nt_tv_intol	Number of taxa - tolerance value - intolerant ≤ 3	TOLER	3.2	17	100*(metric)/17	57.1	Dec.
pi_tv_intol	Percent individuals - tolerance value - intolerant ≤ 3	TOLER	6.09	51.5	100*(metric)/51.5	59.5	Dec.
pt_tv_intol	Percent taxa - tolerance value - intolerant ≤ 3	TOLER	14.8	51.4	100*(metric)/51.4	50	Dec.
x_Becks	Becks Biotic Index	TOLER	12	36.8	100*(metric)/36.8	57.1	Dec.
x_HBI	Hilsenhoff Biotic Index (references the TolVal field)	TOLER	3.02	4.9	100*(4.90-metric)/1.88	42.9	Inc.
nt_volt_semi	Number of taxa - semivoltine (SEMI)	VOLT	1	8.8	100*(metric)/8.8	57.1	Dec.

5^{th}: 5^{th} percentile of all sample metrics in the site class.
95^{th}: 95^{th} percentile of all sample metrics in the site class
DE: Discrimination Efficiency.
Scoring Formula: Replace "metric" with the sample metric value for calculation of an index
Trend: Decreasing (Dec.) or increasing (Inc.) trend with increasing stress

Table 13. Correlation (Spearman rho) among metrics of the Western Highlands macroinvertebrate index candidates.

		Metric #															
#	Metric	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	nt_total	1															
2	nt_EPT	0.68	1														
3	pt_Pleco	0.17	0.48	1													
4	x_Shan_2	0.9	0.61	0.17	1												
5	pi_Pleco	0.23	0.38	0.86	0.26	1											
6	pi_ffg_filt	-0.13	-0.14	-0.48	-0.13	-0.48	1										
7	pi_ffg_pred	0.32	0.32	0.64	0.32	0.67	-0.45	1									
8	pi_ffg_shred	0.14	0.04	0.21	0.15	0.34	-0.32	0.13	1								
9	pt_ffg_pred	0.28	0.4	0.66	0.23	0.57	-0.33	0.72	0.17	1							
10	nt_habit_sprawl	0.26	-0.2	-0.18	0.27	-0.01	-0.17	0.07	0.32	-0.01	1						
11	pi_habit_cling	-0.27	0.01	-0.06	-0.29	-0.25	0.44	-0.23	-0.33	-0.05	-0.54	1					
12	pi_habit_sprawl	0.1	-0.33	-0.13	0.16	0.1	-0.29	0.1	0.34	-0.07	0.82	-0.67	1				
13	nt_tv_intol	0.57	0.8	0.71	0.5	0.64	-0.34	0.56	0.16	0.61	-0.14	-0.11	-0.18	1			
14	pi_tv_intol	0.17	0.45	0.67	0.17	0.67	-0.37	0.59	0.07	0.56	-0.09	-0.18	-0.08	0.76	1		
15	x_Becks	0.68	0.88	0.67	0.61	0.62	-0.32	0.52	0.15	0.56	-0.12	-0.13	-0.2	0.94	0.66	1	
16	x_HBI	-0.15	-0.49	-0.67	-0.15	-0.6	0.31	-0.54	-0.03	-0.61	0.22	0.03	0.23	-0.74	-0.92	-0.68	1
17	nt_volt_semi	0.43	0.46	0.6	0.37	0.51	-0.25	0.48	0.01	0.56	-0.2	0.04	-0.22	0.59	0.42	0.62	-0.47

5.4 Index Compilations and Performance

Alternative index compositions were formulated from the best performing metrics in each metric category. The metrics were combined by scoring each on the 0 to 100 scale and then averaging the scores. Each alternative index was then evaluated for discrimination efficiency and other measures of representativeness and sensitivity. Multiple index formulations were created and evaluated in two ways: manual metric substitutions and automatic all-subsets modeling.

In the manual substitution method, the best metrics were included in an initial index and subsequent index alternative were formulated by removing, adding, or substituting metrics. The performance of each index formulation was used to determine whether the latest alternative was an improvement. New formulations were built upon the best performing alternatives. The manual substitution method of index development that was a preliminary index development strategy was not exhaustive and included less than 30 alternatives derived using metrics with good performance in each metric category.

The all-subsets analysis allowed consideration of diverse alternatives that were not considered in the manual method. The best 10-20 candidate metrics in each site class were selected for inclusion in index trials based on discrimination efficiency, Z-score, and professional opinion of the MassDEP biologists. An "all subsets" routine in R software (R Core Team 2013) was used to combine up to 10 metrics in multiple index trials. Each of the index alternatives was evaluated for performance using DE, Z-score, number of metric categories, and redundancy of component metrics. Those models including two or more correlated metrics (Spearman $|r| \geq 0.80$) were excluded from consideration. As many metric categories as practical were represented in the index alternatives so that signals of various stressor-response relationships would be integrated into the index. While several metrics should be included to represent biological integrity, redundant metrics can bias an index to show responses specific to certain stressors or taxonomic responses.

The metrics shown in Table 9 were included in the all-subsets analysis for the two classes. The candidate index metrics included 15 that were common to both site classes. The uncommon metrics might be uniquely responsive in either class due to unique stressors or natural background conditions. The all-subsets model calculation and screening resulted in multiple valid index combinations to evaluate for each site class. After removing index combinations that included redundant metrics, there were 12,799 indices evaluated in the Central Hills and 48,215 indices in the Western Highlands. There were more highly correlated metrics in the Central Hills, resulting in fewer valid index combinations to be evaluated.

To identify the most sensitive, comprehensive, and practical index alternatives, the characteristics of the alternatives were screened for discrimination performance and other favorable characteristics; including critical metric categories, excluding metrics with conceptual redundancy, and excluding metrics with unexplained response mechanisms. Plots were also generated to evaluate the distribution of DEs across index alternatives comprised of different numbers of metrics, which helped reviewers decide on a minimum and maximum number of metrics to include in their top picks. In the Western Highlands, consistently high DEs were observed in index alternatives with 5 – 8 metrics (Figure 17). The median of DEs increased with additional metrics, though the maximum DE decreased with more than 8 metrics.The two highest DEs (>90%) occurred in index combinations with 3 and 4 metrics. In the Central Hills, DEs of 100% were observed in indices with 1-8 metrics (Figure 18). The highest median DEs were observed in indices with 5-8 metrics.

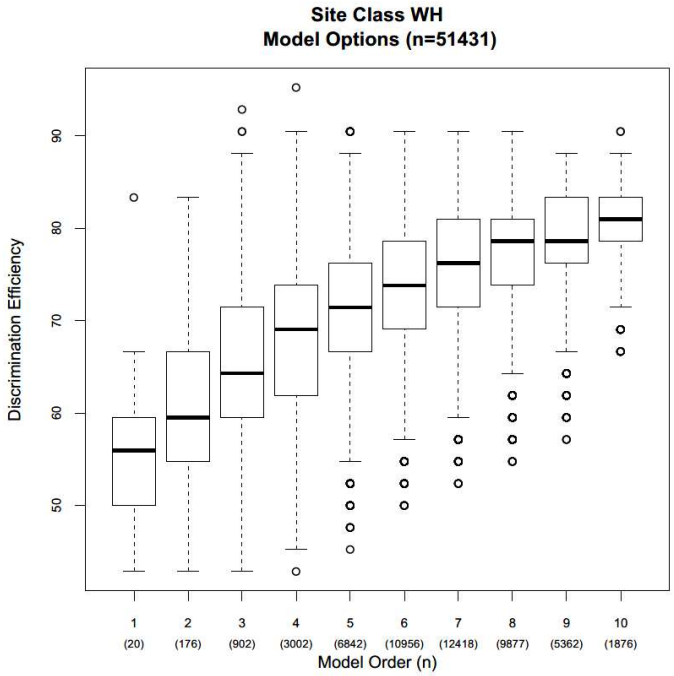


Figure 17. Distribution of Discrimination Efficiency scores (DEs) for index alternatives evaluated in the all-subsets analysis in the Western Highlands, grouped by the number of metrics included in the alternative. For example, the first box on the left is labeled '1' – this means there is 1 metric in that index group; the (20) below it means there are 20 index alternatives comprised of 1 metric; the box plot shows the distribution of DEs for the 1-metric group.

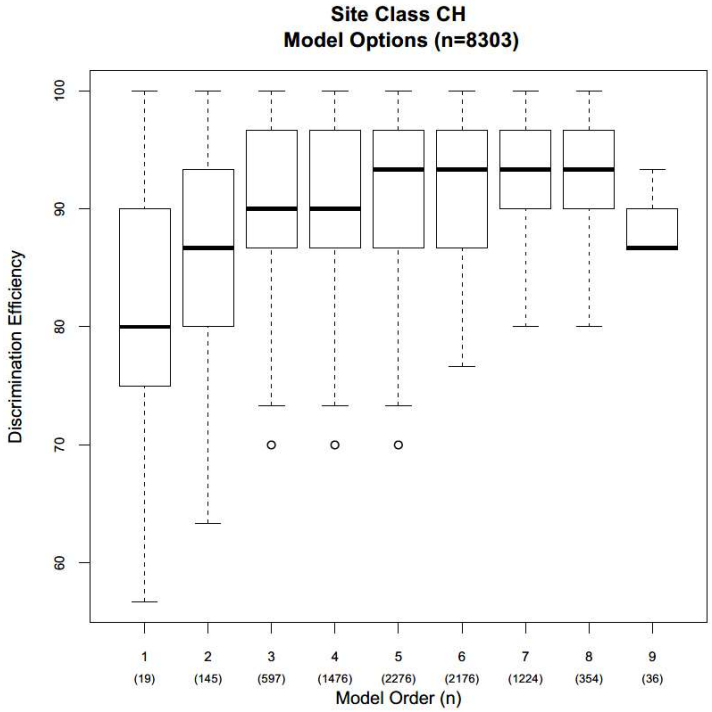


Figure 18 . Distribution of Discrimination Efficiency scores (DEs) for index alternatives evaluated in the all-subsets analysis in the Central Hills, grouped by the number of metrics included in the alternative.

To narrow down the long list of index alternatives, two reviewers from MassDEP (James Meek and Bob Nuzzo) were provided with an Excel worksheet with results from the all-subsets analysis. They were asked to select their top twenty index combinations for each class and document the rationale behind their choices They worked independently. Their screening and decision criteria are summarized in Table 14. The 20 index alternatives that they selected had similar performance statistics. They then narrowed it down to their top five picks. These final selections included subjective decisions per site class and reviewer.

Table 14. MassDEP reviewer screening and decision criteria for narrowing down the list of index candidates.

Reviewer #1	Reviewer #2
Western Highlands • Discrimination Efficiency > 85 • At least 6 metrics • Excluding pi_habit_cling due to unexplained increasing trend with stress • At least one metric in richness, composition, FFG, and tolerance categories • No metric category with more than 2 metrics to reduce conceptual redundancy • Includes either nt_total or nt_EPT • Z-score ≥ 1.4	• Discrimination Efficiency > 85 • At least 6 metrics • Avoiding habit metrics • At least 5 metric categories
Central Hills • Discrimination Efficiency > 95 • At least one metric in the richness, composition, FFG, and tolerance categories • Excluding pt_Pleco due to uneven reference/stressed site distribution • At least 6 metrics • Z-score ≥ 2.6	• Discrimination Efficiency > 93 • At least 6 metric categories • Includes pi_dom04

The top ten models (5 from each reviewer) for each class are shown in Tables 15 (CH) and 16 (WH). The tables include the input metrics as well as the performance statistics (DE and z-scores). In the the Central Hills indices, the top ten index candidates had DEs > 90% and Z-scores >2.2, indicating excellent separation between reference and stressed index scores. The candidate indices met the reviewers' criteria. All metrics showed individual DEs greater than 60% (Table 10). Metrics included in any index were not redundant, having correlation coefficient magnitudes $|r| < 0.85$ (Table 11). In the Western Highlands site class, the top ten index candidates had DEs > 85% and Z-scores >1.2, indicating good separation between reference and stressed index scores. The candidate indices met the reviewers' criteria. All metrics showed individual DEs greater than 40% (Table 12). Metrics included in any index were not redundant, having correlation coefficient magnitudes $|r| < 0.85$ (Table 13).

MassDEP biologists selected a final index from among these alternatives (which are highlighted in green in Tables 15 & 16). We refer to the final choices as the CH IBI and the WH IBI. The biologists rationalized their choice based on empirical performance and ecological characteristics of the individual and combined metrics. In the end, they decided to pick indices with metrics they were most familiar with (tolerance, FFG, richness and composition) and to leave Habit, Life Cycle and thermal preference as exploratory metrics. The CH IBI and WH IBI are discussed in more detail in Section 5.5.

Index Verification

Index scores were also calculated for the verification samples. Their performance statistics were compared to calibration results. Verification data were expected to perform nearly as well as calibration data, with DEs not more than 10% less than the calibration data. This was evaluated by comparing reference and stressed validation data to the reference calibration 25th percentiles.

In the Central Hills top ten indices, all had at least 90% of stressed verification samples below the calibration reference 25th percentile (Table 17). Two of the indices had high percentages of reference sites with values below the 25th percentile of calibration reference. The selected index (CH IBI; model 6_ 33344) had a calibration stressed DE of 100% and 33% verification reference below the 25th percentile of calibration reference, which are acceptable statistics for verification (no more than 10% more errros than calibration). In the Western Highlands, the selected index was verified, though some other alternatives had low percentages of stressed verification sites with values below the 25th calibration reference value. Box plots showing the distributions of scores for all top 10 index combinations in the calibration and verification datasets for each class can be found in Appendix H.

As a final check, Spearman rank correlations were performed to evaluate the responsiveness of the indices to stressors. Results showed that all ten indices in each class were responsive to stressors, though not always equally responsive to the same stressors in both site classes (Table 19 – CH; Table 20 - WH). The strongest correlations were with the IWI and urban land uses. The indices were weakly positively related to agricultural uses in the Central Hills.

Table 15 . The ten best macroinvertebrate index alternatives for the Central Hills site class (selected by MassDEP reviewers). Metrics used in each alternative are listed as "1". 0 = not included. The final model that was selected is highlighted in green (model 6_33344). See Table 10 for code descriptions.

Metric	Model ID									
	8_ 61881	**8_ 100881**	**8_ 100980**	**8_ 60267**	**8_ 99023**	**6_ 33344**	**6_ 25973**	**6_ 30341**	**6_ 37425**	**7_ 76362**
nt_total	0	0	0	0	0	1	0	0	0	0
nt_Insect	0	0	0	0	0	0	0	1	0	0
pt_EPT	0	0	0	0	0	1	1	1	1	1
pt_Pleco	1	1	1	1	1	0	0	0	0	0
pi_dom04	1	1	1	1	1	0	0	0	0	0
pi_Ephem NoCaeBae	1	1	1	1	1	1	1	1	1	1
nt_ffg_pred	1	0	0	1	0	0	0	0	0	0
pi_ffg_filt	1	1	1	1	1	1	1	1	1	1
pi_ffg_pred	0	1	0	0	1	0	0	0	0	0
pt_ffg_pred	0	0	1	0	0	1	1	1	1	1
nt_habit_cling	1	1	1	1	1	0	1	0	0	0
pi_tv_intol	1	1	0	1	1	0	0	0	0	0
pt_tv_intol	0	0	1	0	0	1	1	1	1	1
pt_volt_semi	1	1	1	0	0	0	0	0	0	1
nt_volt_uni	0	0	0	0	0	0	0	0	1	1
nt_volt_semi	0	0	0	1	1	0	0	0	0	0
Str.DE	***93.3***	***93.3***	***93.3***	***93.3***	***93.3***	***100***	***100***	***100***	***100***	***96.7***
z	***2.29***	***2.24***	***2.63***	***2.22***	***2.18***	***3.02***	***2.94***	***2.96***	***2.78***	***2.73***

Table 16. The ten best macroinvertebrate index alternatives for the Western Highlands site class (selected by MassDEP reviewers). Metrics used in each alternative are listed as "1". 0 = not included. The final model that was selected is highlighted in green (model 6_32701). See Table 12 for code descriptions.

Metric	Model ID									
	7_ 74572	**7_ 74172**	**7_ 24379**	**8_ 119923**	**7_ 20203**	**6_32701**	**6_32713**	**6_31111**	**6_31836**	**7_68339**
nt_total	0	0	0	0	1	1	1	1	1	1
nt_EPT	0	0	1	0	1	0	0	0	0	0
pt_Pleco	0	0	0	1	0	0	0	0	0	1
x_Shan_2	1	1	1	1	0	0	0	0	0	0
pi_Pleco	1	1	1	0	1	1	1	1	0	0
pi_ffg_filt	1	1	1	1	1	1	1	1	1	1
pi_ffg_pred	0	1	0	0	0	0	0	0	0	0
pi_ffg_shred	1	1	1	1	1	1	1	1	1	1
pt_ffg_pred	0	0	0	0	0	0	0	0	0	0
nt_habit_sprawl	0	0	0	0	0	0	0	0	0	0
pi_habit_cling	0	0	0	1	0	0	0	0	0	0
pi_habit_sprawl	0	0	0	0	0	0	0	0	0	0
nt_tv_intol	0	0	0	1	0	0	0	1	0	0
pi_tv_intol	1	0	0	1	0	1	0	1	1	1
pt_tv_intol	0	0	0	0	0	0	1	0	0	0
x_Becks	1	1	0	0	0	1	1	0	1	1
x_HBI	0	0	1	0	1	0	0	0	0	0
nt_volt_semi	1	1	1	1	1	0	0	0	1	1
Str.DE	***85.7***	***81.0***	***85.7***	***85.7***	***83.3***	***88.1***	***85.7***	***85.7***	***85.7***	***78.6***
z	***1.33***	***1.27***	***1.21***	***1.41***	***1.20***	***1.40***	***1.33***	***1.35***	***1.36***	***1.33***

Table 17 . Verification statistics for the macroinvertebrate index alternatives in the Central Highlands. The final model that was selected is highlighted in green (model 6_33344). See Table 15 for more information on the indices.

Statistic	Model ID									
	8_ 61881	8_ 100881	8_ 100980	8_ 60267	8_ 99023	6_ 33344	6_ 25973	6_ 30341	6_ 37425	7_ 76362
Verif.Ref.DE	16.7	16.7	16.7	16.7	16.7	33.3	41.7	33.3	41.7	25.0
Verif.Str.DE	92.9	92.9	92.9	92.9	92.9	100.0	100.0	100.0	100.0	100.0

Table 18. Verification statistics for the macroinvertebrate index alternatives in the Western Highlands. The final model that was selected is highlighted in green (model 6_32701). See Table 16 for more information on the indices.

Statistic	Model ID									
	7_ 74572	7_ 74172	7_ 24379	8_ 119923	7_ 20203	6_32701	6_32713	6_31111	6_31836	7_68339
Verif.Ref.DE	33.33	33.33	33.33	16.67	33.33	25.00	16.67	25.00	25.00	16.67
Verif.Str.DE	78.57	71.43	71.43	78.57	57.14	85.71	78.57	85.71	78.57	57.14

Table 19 . Central Hills. Correlation coefficients (Spearman rank r) for the indices and disturbance variables. The final model that was selected is highlighted in green (model 6_33344). All of the correlations are significant (p<0.05). See Table 3 for variable descriptions.

Model ID	ICI v1	IWI v1	% Urban	% Agricultural
8_ 61881	0.53	0.68	-0.69	0.21
8_ 100881	0.53	0.68	-0.69	0.21
8_ 100980	0.54	0.69	-0.71	0.22
8_ 60267	0.52	0.68	-0.69	0.21
8_ 99023	0.52	0.68	-0.69	0.20
6_ 33344	0.54	0.70	-0.72	0.20
6_ 25973	0.52	0.70	-0.72	0.23
6_ 30341	0.54	0.71	-0.73	0.21
6_ 37425	0.53	0.69	-0.70	0.21
7_ 76362	0.53	0.68	-0.70	0.23

Watershed-scale, Source: StreamCat, based on the NLCD 2011 land cover dataset. Urban includes low + medium + high intensity. Ag = hay + crop

Table 20. Western Highlands. Correlation coefficients (Spearman rank r) for the indices and disturbance variables. All of the correlations are significant (p<0.05). The final model that was selected is highlighted in green (model 6_32701). See Table 3 for variable descriptions.

Model ID	ICI v1	IWI v1	% Urban	% Agricultural
7_ 74572	0.47	0.45	-0.48	-0.36
7_ 74172	0.43	0.44	-0.45	-0.34
7_ 24379	0.43	0.41	-0.44	-0.30
8_ 119923	0.49	0.46	-0.49	-0.36
7_ 20203	0.43	0.41	-0.44	-0.29
6_32701	0.50	0.46	-0.50	-0.34
6_32713	0.46	0.44	-0.46	-0.29
6_31111	0.50	0.45	-0.50	-0.33
6_31836	0.46	0.45	-0.47	-0.34
7_68339	0.45	0.43	-0.45	-0.31

Watershed-scale, Source: StreamCat, based on the NLCD 2011 land cover dataset. Urban includes low + medium + high intensity. Ag = hay + crop

5.5 Final Index Selection and Performance

The team of MassDEP biologists used the following empirical and logical criteria to select their final top picks (the CH IBI and WH IBI):

- Relatively high index DE and Z-scores
- Index metrics representing as many metric categories as practical
- Not including redundant metrics
- Inclusion of individual metrics having the following characteristics:

  - High overall DE within and among site classes
  - Response mechanisms that were plausible and ecologically important
  - Not conceptually redundant with other index metrics regardless of statistical correlations
  - Straightforward metric calculations

The CH IBI and WH IBI each have six metrics. The component metrics, performance statistics, scoring formulas and correlation matrixes for each index are listed in Tables 21-24. The CH and WH IBIs share two of the same metrics (number of total taxa and percent collector-filterer individuals). The box plots in Figure 19 show how IBI scores are distributed across the reference and stressed groups. Calibration and verification scores are shown side by side. The median IBI scores in the reference calibration samples are lower in the WH than in the CH (median=55 in the WH versus 70 in the CH) and the median WH scores in the stressed calibration group are slightly higher (approximately 36 vs. 30). This is a function of the range of scores represented in the two calibration datasets (as an example, in the WH dataset, the 5th percentile of the number of total taxa metric ranges from 21 (5th percentile) to 39 (95% percentile) (Table 21) vs. the CH, where it ranges from 11 to 35 (Table 23). This is not a surprise given how the CH has higher levels of disturbance than the WH (Figure 13). Thus, the scoring scales should not be interpreted to be directly comparable, even though they are both on a scale of 0 to 100.

We also performed a secondary verification check on the CH IBI and WH IBI using kick net data collected prior to 2000 (when some taxonomic standards were a little different). The stressed sites were all below 50 index units and all three reference samples had index values above 45 index units (Figure 20). There were few reference samples collected before 2000, so the comparison of that distribution is not a reliable representation of the data type.

Table 21. Metrics in the Central Hills index, with scoring formulas, Discrimination Efficiency (DE) scores and trend. Metrics in bold text are also used in the Western Highlands IBI.

Metric abbreviation	Metric	Category	5th	95th	Scoring formula	DE	Trend
nt_total	**Number of taxa - total**	RICH	11	34.9	100*(metric)/ 34.9	66.7	Dec.
pt_EPT	Percent taxa - Orders Ephemeroptera, Plecoptera & Trichoptera (EPT)	RICH	10.6	54.5	100*(metric)/ 54.5	76.7	Dec.
pi_Ephem NoCaeBae	Percent individuals - Order Ephemeroptera, excluding Families Caenidae and Baetidae	COMP	0	13.9	100*(metric)/ 13.9	66.7	Dec.
pi_ffg_filt	**Percent individuals - Functional Feeding Group (FFG) - collector-filterer (CF)**	**FFG**	**13**	**79.9**	**100*(79.9-metric)/ 66.9**	**76.7**	**Inc.**
pt_ffg_pred	Percent taxa - Functional Feeding Group (FFG) - predator (PR)	FFG	0	28.5	100*(metric)/ 28.5	90	Dec.
pt_tv_intol	Percent taxa - tolerance value - intolerant ≤ 3	TOLER	0	39.1	100*(metric)/ 39.1	100	Dec.

5^{th}: 5^{th} percentile of all sample metrics in the site class.
95^{th}: 95^{th} percentile of all sample metrics in the site class
DE: Discrimination Efficiency.
Scoring Formula: Replace "metric" with the sample metric value for calculation of an index
Trend: Decreasing (Dec.) or increasing (Inc.) trend with increasing stress

Table 22. Central Hills IBI. Correlation (Spearman rho) among metrics of the Central Hills macroinvertebrate index.

Metric	nt_total	pt_EPT	pi_Ephem NoCaeBae	pi_ffg_filt	pt_ffg_pred	pt_tv_intol
nt_total	1					
pt_EPT	0.31	1				
pi_Ephem NoCaeBae	0.49	0.58	1			
pi_ffg_filt	-0.40	0.05	-0.09	1		
pt_ffg_pred	0.38	-0.04	0.14	-0.33	1	
pt_tv_intol	0.62	0.60	0.51	-0.40	0.41	1

Table 23. Metrics in the Western Highlands index, with scoring formulas, Discrimination Efficiency (DE) scores and trend. Metrics in bold text are also used in the Central Hills IBI.

Metric abbreviation	Metric	Category	5th	95th	Scoring formula	DE	Trend
nt_total	**Number of taxa - total**	RICH	21	38.8	100*(metric)/ 38.8	52.4	Dec.
pi_Pleco	Percent individuals - Order Plecoptera	COMP	0	18.3	100*(metric)/ 18.3	66.7	Dec.
pi_ffg_filt	**Percent individuals - Functional Feeding Group (FFG) - collector-filterer (CF)**	FFG	9.76	50.5	100*(50.5-metric)/ 40.7	50	Inc.
pi_ffg_shred	Percent individuals - FFG - shredder (SH)	FFG	1.17	23	100*(metric-)/ 23	61.9	Dec.
pi_tv_intol	Percent individuals - tolerance value - intolerant ≤ 3	TOLER	6.09	51.5	100*(metric-)/ 51.5	59.5	Dec.
x_Becks	Becks Biotic Index*	TOLER	12	36.8	100*(metric)/ 36.8	57.1	Dec.

Beck's Biotic Index (Terrell and Perfetti 1996) = 2[Class 1 Taxa]+[Class 2 Taxa] where Class 1 taxa have tolerance values of 0 or 1 and Class 2 taxa have tolerance values of 2, 3 or 4.

5^{th}: 5^{th} percentile of all sample metrics in the site class.

95^{th}: 95^{th} percentile of all sample metrics in the site class

DE: Discrimination Efficiency.

Scoring Formula: Replace "metric" with the sample metric value for calculation of an index

Trend: Decreasing (Dec.) or increasing (Inc.) trend with increasing stress

Table 24. Correlation (Spearman rho) among metrics of the Western Highlands macroinvertebrate index.

Metric	nt_total	pi_Pleco	pi_ffg_filt	pi_ffg_shred	pi_tv_intol	x_Becks
nt_total	1					
pi_Pleco	0.23	1				
pi_ffg_filt	-0.13	-0.48	1			
pi_ffg_shred	0.14	0.34	-0.32	1		
pi_tv_intol	0.17	0.67	-0.37	0.07	1	
x_Becks	0.68	0.62	-0.32	0.15	0.66	1

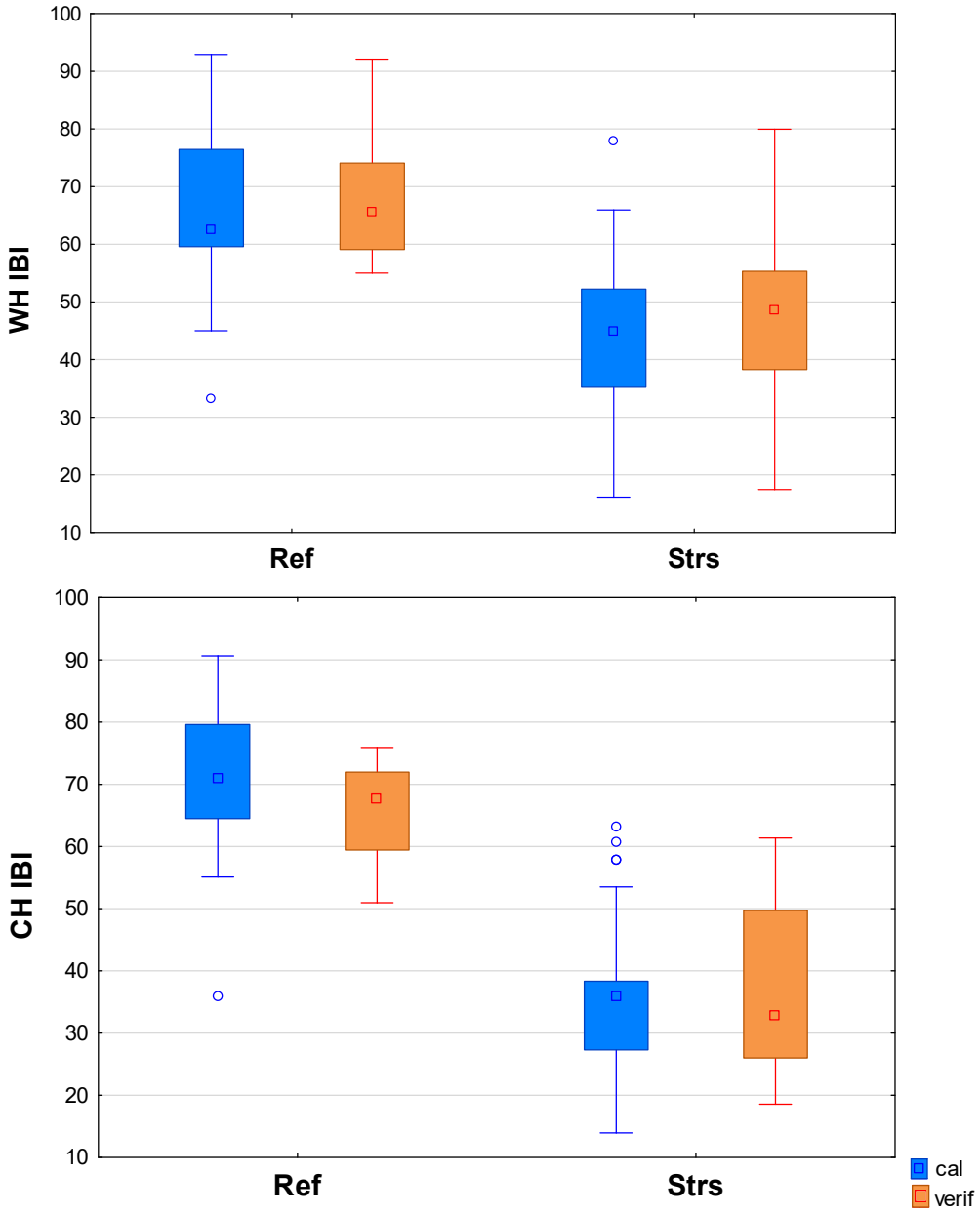


Figure 19. Box plots showing the distribution of Western Highland (WH) IBI scores (top) and Central Hills (CH) IBI scores (bottom) in the reference and stressed calibration and verification datasets.

Central Hills		
Dataset	**Number of sites**	
	Ref	**Strs**
cal	30	30
verif	12	14
verif pre-2000	3	19

Western Highlands		
Dataset	**Number of sites**	
	Ref	**Strs**
cal	26	42
verif	12	14
verif pre-2000	3	14

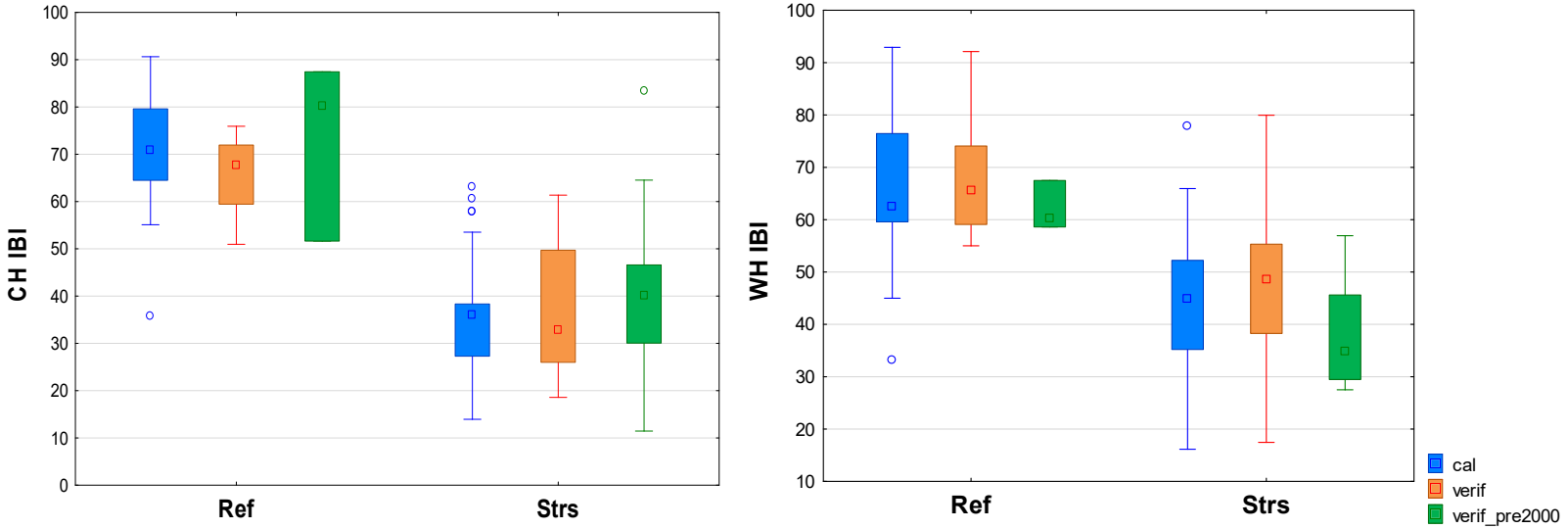


Figure 20. Box plots showing the distribution of Central Hills (CH) IBI (left) and Western Highland (WH) IBI (right) scores in the reference and stressed calibration and verification datasets (2000-2017) and the pre-2000 verification kick net datasets.

6 Discussion

IBIs were calibrated for MassDEP benthic macroinvertebrate kick net samples in freshwater perennial wadeable streams in two naturally distinct regions: Western Highlands and Central Hills. The southeast portion of Massachusetts, which includes the Narragansett/Bristol Lowlands, Cape Cod, and the Islands, had insufficient data to develop an IBI for kick net samples. However, a statewide IBI for low gradient streams (which are prevalent in southeastern MA) that are sampled with the MassDEP multihabitat collection method is currently under development.

The new RBP kick net IBIs improve MassDEP's diagnostic ability to identify degradation in biological integrity and water quality. The IBIs are modernized compared to past assessment indices used in Massachusetts and make use of data that were collected from hundreds of sites in recent years. They are multimetric indices comprised of biological metrics that were found to be responsive to a general stressor gradient, are ecologically meaningful and diverse in response mechanisms. The metrics included in each regional index represented four categories of metric responses: taxa richness, individual composition, functional feeding groups, and pollution tolerance. Two of the six metrics in each index are common to both regions (number of total taxa and percent collector-filterer individuals).

The IBIs were calibrated using the Reference Condition (RC) approach, which bases biological expectations on reference sites within each region (Western Highlands and Central Hills). If a site receives an IBI score that does not resemble reference scores, it indicates that there might be stressors influencing the biological condition at that site. During calibration, the IBIs had minimal error when discriminating between least-disturbed reference and most disturbed stressed sites. During validation with independent data, the IBIs also performed well, with an error rate within 10% of the calibration dataset. Additional checks that were performed on multihabitat and pre-2000 data (both of which were excluded from the calibration and validation datasets) also showed adequate validation of the indices.

The IBIs as they currently stand can be used to assess stream degradation relative to least-disturbed streams in each of the two regions. Some state biomonitoring programs take the additional step of establishing numeric IBI thresholds to designate different categories of biological condition and to assess attainment of aquatic life use standards. The Massachusetts Surface Water Quality Standards (SWQS) (314 CMR 4.00; MassDEP 2013) currently has narrative biological criteria that define biological integrity as "the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having species composition, diversity, and functional organization comparable to that of the natural habitat of the region." In addition, the SWQS designate specific uses for surface water classes. For inland waters, Class A must sustain excellent habitat, while Class B waters must sustain habitat for aquatic life and wildlife. Waters supporting Aquatic Life Use should be suitable for "sustaining a native, naturally diverse, community of aquatic flora and fauna. This use includes reproduction, migration, growth and other critical functions" (MassDEP 2013).

MassDEP does not currently have plans to pursue numeric bio-criteria but has begun to explore potential thresholds for four biological condition categories (Exceptional Condition, Satisfactory Condition, Moderately Degraded, and Severely Degraded). A separate report (Stamp and Jessup 2020) describes the analyses that were performed to start exploring potential thresholds. A number of state biomonitoring programs have integrated numeric biocriteria into their WQS (e.g., Maine,

Minnesota[5]). If MassDEP decides to make this a future pursuit, the proposed criteria would need to go through a rule-making process that includes a period for public review and comment. Any proposed amendments to use numeric biocriteria as the basis for water quality management actions under the CWA would need to be approved by the U.S. Environmental Protection Agency (EPA) following promulgation.

Moving ahead, in addition to exploring potential IBI thresholds, MassDEP will continue to evaluate the kick net IBIs as new data are collected (are results are in keeping with expectations? Where are the IBIs performing well? Where are they performing poorly and why?). In addition, MassDEP is planning to conduct targeted sampling to broaden the disturbance gradients represented in each region, which will be important for future IBI recalibrations (which are recommended periodically; e.g., Jessup and Stribling 2008, Stribling et al. 2016). Currently the IBI in the Central Hills is more sensitive to the stressor gradient than the IBI in the Western Highlands. This difference is attributed to the difference in the general stressor intensity across the landscape of Massachusetts. For example, there are more areas with sparse development in the west and more severe stressor conditions in the east. The difference was recognized when establishing the reference and stressed data sets, so that the standards for "least-disturbed" were higher in the Western Highlands than in the Central Hills. MassDEP recently identified candidate monitoring sites in both regions that would help provide a more complete characterization of the disturbance gradient. The targeted sites are intended to bolster the reference dataset in the Central Hills and the highly stressed dataset in the Western Highlands.

In addition, MassDEP is considering strategies being used by other states to address limitations in biocriteria development stemming from differences in disturbance patterns. For example, Minnesota (MN) has a north-south disturbance gradient, where northern MN has limited numbers of highly stressed sites (similar to the Western Highlands) and southern MN has relatively high levels of anthropogenic disturbance with few least- or minimally disturbed reference sites (similar to the Central Hills). MN used the Biological Condition Gradient (BCG) (U.S. EPA 2016) to supplement their IBIs and help address shortcomings in the RC approach, which, if used alone, would have limitatations for setting protective goals and for ensuring consistency among stream types (Bouchard et al. 2016). The BCG is a conceptual model describing how ecological attributes change in response to increasing levels of human-caused stress (Davies and Jackson 2006, U.S. EPA 2016, Hausmann et al. 2016, Gerritsen et al. 2017). Calibration of the BCG is a collective exercise among biologists to develop consensus assessments of samples, and then to elicit the rules that the biologists use to assess the samples (Davies and Jackson 2006). The BCG has a universal scale that is typically divided into six levels of biological condition along a generalized stressor-response curve, ranging from observable biological conditions found at no or low levels of stressors (level 1) to those found at high levels of stressors (level 6) (US EPA 2016). In southern MN, there were no BCG level 1 or level 2 sites; rather the best were BCG level 3 or 4. In contrast, northern MN had few BCG level 5 or 6 sites. MN was able to use this universal scale to synchronize biocriteria for stream types from its southern versus northern regions, as well as to communicate these differences to the public. MN also used the BCG to inform its IBI-based numeric biocriteria thresholds, which were more ecologically meaningful when linked to the BCG biological narratives (Bouchard et al. 2016). If MassDEP decided to pursue a similar approach, it could use the existing New England high gradient BCG model (Snook et al. 2007) as a starting point for calibrating a BCG model for its kick net samples.

[5] More information on numeric biocriteria in Maine and Minnesota can be found at the following links: Maine - https://www.maine.gov/dep/water/monitoring/biomonitoring/retro/pt1ch1pref.pdf; Minnesota - https://www.pca.state.mn.us/sites/default/files/wq-bsm4-02.pdf

7 Literature Cited

Barbour, M. T., J. Gerritsen, G.E. Griffith, R. Frydenborg, E. McCarron, J.S. White, and M.L. Bastian. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. Journal of the North American Benthological Society 15(2):185-211.

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

Bode, R.W., Novak, M.A., and Abele, L.E. 1996. Quality Assurance Work Plan for Biological Stream Monitoring in New York State. NYS Department of Environmental Conservation, Albany, NY. 89p.

Bode, R.W., Novak, M.A., Abele, L.E., Heitzman, D.L., and Smith, A.J. 2002. Quality Assurance Work Plan for Biological Stream Monitoring in New York State. NYS Department of Environmental Conservation, Albany, NY. 115p.

Bouchard, R.W., S. Niemela, J.A. Genet, C.O. Yoder, J. Sandberg, J.W. Chirhart, M. Feist, B. Lundeen, and D. Helwig 2016. A novel approach for the development of tiered use biological criteria for rivers and streams in an ecologically diverse landscape. Environmental monitoring and assessment 188: 196.

Brown, M.T. and M.B. Vivas. 2005. Landscape Development Intensity Index. Environmental Monitoring and Assessment 101: 289-309.

Bryce, S.A. and S.E. Clarke. 1996. Landscape-level ecological regions: linking state-level ecoregion frameworks with stream habitat classifications. Environmental Management, 20(3):297-311.
Cohen, J. 1992. A power primer. Psychological Bulletin 112(1): 155.

Cohen, J. 1992. A power primer. Psychological Bulletin, 112(1):155.

Cuffney, T.F., M.D. Bilger and A.M. Haigler. 2007. Ambiguous taxa: effects on the characterization and interpretation of invertebrate assemblages. J N Am Benthol Soc 26(2): 286–307.

Davies, S. B., and S. K. Jackson. 2006. The Biological Condition Gradient: A descriptive model for interpreting change in aquatic ecosystems. Ecological Applications 16(4):1251–1266.

DeShon, J.E. 1995. Development and Application of the Invertebrate Community Index (ICI). In: Davis, W.S. and Simon, T.P., Eds., Biological Assessment and Criteria—Tools for Water Resource Planning and Decision Making, Lewis Publ., Boca Raton, 217-244.

Flotemersch, J.E., Leibowitz, S.G., Hill, R.A., Stoddard, J.L., Thomas, M.C., & Tharme, R.E. 2016. A watershed Integrity Definition and Assessment approach to Support Strategic Management of Watersheds. River Research and Applications, 32, 1654–1671.

Fore, L.S., R. Frydenborg, D. Miller, T. Frick, D. Whiting, J. Espy, and L. Wolfe 2007. Development and testing of biomonitoring tools for macroinvertebrates in Florida streams (Stream Condition Index and Biorecon). Final Report, Prepared for: Florida Department of Environmental Protection, Tallahassee, FL.

Gerritsen, J., M.T. Barbour, and K. King. 2000. Apples, oranges, and ecoregions: on determining pattern in aquatic assemblages. Journal of the North American Benthological Society 19(3): 487-496.

Gerritsen, J., R.W. Bouchard Jr., L. Zheng, E.W. Leppo, and C.O. Yoder. 2017. Calibration of the biological condition gradient in Minnesota streams: a quantitative expert-based decision system. Freshwater Science, 36(2), 427-451.

Hausmann, S., Charles D.F., Gerritsen J., and T.J. Belton. 2016. A diatom-based biological condition gradient (BCG) approach for assessing impairment and developing nutrient criteria for streams. Sci Total Environ. 562:914-927. doi: 10.1016/j.scitotenv.2016.03.173.

Hill, R.A., Weber, M.H., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., 2016. The Stream-Catchment (StreamCat) dataset: a database of watershed metrics for the Conterminous United States. J. Am. Water Res. Assoc. 52 (1), 120–128

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

Hughes, R. M., P. R. Kaufmann, A. T. Herlihy, T. M. Kincaid, L. Reynolds, and D. P. Larsen, 1998. A process for developing and evaluating indices of fish assemblage integrity. Canadian Journal of Fisheries and Aquatic Sciences, 55(7):1618-1631.

Jessup, B. and J.B. Stribling. 2008. Evaluation and Recalibration of the Mississippi Benthic Index of Stream Quality (M-BISQ). Prepared for: the Mississippi Department of Environmental Quality, Office of Pollution Control, Jackson, Mississippi. Prepared by: Tetra Tech, Inc., Owings Mills, Maryland (for further information, contact Ms. Valerie Alley, MDEQ, 601-961-5182).

Jessup, B. and J. Stamp. 2017. Development of Multimetric Indices of Biotic Integrity for Assessing Macroinvertebrate and Fish Assemblages in Indiana Streams. Prepared for: US EPA Region 5, Chicago, IL and Indiana Department of Environmental Management, Indianapolis, IN.

Johnson, Z.G., Leibowitz, S. and R. Hill. 2019. Revising the index of watershed integrity national maps. Science of The Total Environment. 10.1016/j.scitotenv.2018.10.112.

Karr, J.R., and D.R. Dudley. 1981. Ecological perspectives on water quality goals. *Environ. Manage.* 5:55-68.

Lenat, David R. 1993. A Biotic Index for the Southeastern United States: Derivation and List of Tolerance Values, with Criteria for Assigning Water-Quality Ratings. Journal of the North American Benthological Society 12(3): 279-290.

MassDEP. 2004. CN 187.1. QAPP for 2004 Biological Monitoring and Habitat Assessment. Massachusetts Department of Environmental Protection, Division of Watershed Management. Worcester, MA. 99 p.

Massachusetts Division of Water Pollution Control. 2013. 314 CMR 4.00: Massachusetts Surface Water Quality Standards. Available online: https://www.mass.gov/files/documents/2016/11/nv/314cmr04.pdf

Massachusetts Division of Watershed Management Watershed Planning Program. 2016. Massachusetts Consolidated Assessment and Listing Methodology (CALM) Guidance Manual for the

2016 Reporting Cycle. Available online:
https://www.mass.gov/files/documents/2016/10/wy/2016calm.pdf

Maxted, J.R., M.T. Barbour, J. Gerritsen, V. Poretti, N. Primrose, A. Silvia, D. Penrose, and R. Renfrow. 2000. Assessment framework for mid-Atlantic coastal plain streams using benthic macroinvertebrates. Journal of the North American Benthological Society 19(1):128–144.

McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., Reah, A., 2012. NHDPlus Version 2: User Guide. U.S. Environmental Protection Agency

Merritt, R.W. and K.W. Cummins (editors). 1996. An introduction to the aquatic insects of North America, 3rd ed. Kendall/Hunt Publishing Company, Dubuque, Iowa.

Nuzzo, R. 2003. CN 32.2. Standard Operating Procedures: Water Quality Monitoring in Streams Using Aquatic Macroinvertebrates. Massachusetts Department of Environmental Protection, Division of Watershed Management. Worcester, MA. 35 p.

Ofenböck, T., O. Moog, J. Gerritsen, and M. Barbour. 2004. A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. In Integrated Assessment of Running Waters in Europe (pp. 251-268). Springer Netherlands.

Plafkin, J.L., M.T. Barbour, K.D. Porter, S.K. Gross, and R.M. Hughes. 1989. Rapid bioassessment protocols for use in streams and rivers: benthic macroinvertebrates and fish. EPA/444/4-89-001. U.S. Environmental Protection Agency, Washington, D.C.

Poff, NL; Olden, JD; Vieira, NKM; Finn, DS; Simmons, MP; Kondratieff, BC. 2006. Functional trait niches of North American lotic insects: traits-based ecological applications in light of phylogenetic relationships. J N Am Benthol Soc 25(4):730−755.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rohm, C. M., J.W. Giese and C.C. Bennett. 1987. Evaluation of an aquatic ecoregion classification of streams in Arkansas. Journal of Freshwater Ecology 4(1):127-140.

Roth, N.E., M.T. Southerland, J.C. Chaillou, J.H. Vølstad, S.B. Weisberg, H.T. Wilson, D.G. Heimbuch, J.C. Seibel. 1997. Maryland Biological Stream Survey: Ecological status of non-tidal streams in six basins sampled in 1995. Maryland Department of Natural Resources, Chesapeake Bay and Watershed Programs, Monitoring and Non-tidal Assessment, Annapolis, Maryland. CBWP-MANTA-EA-97-2.

Snook, H., S.P. Davies, J. Gerritsen, B.K. Jessup, R. Langdon, D. Neils, and E. Pizutto. 2007. The New England Wadeable Stream Survey (NEWS): Development of Common Assessments in the Framework of the Biological Condition Gradient. Prepared for USEPA Office of Science and Technology and USEPA Office of Watersheds Oceans and Wetlands, Washington, DC.

Stamp, J. and B. Jessup. 2020. Establishing Numeric Biological Condition Thresholds. Prepared for: Massachusetts Department of Environmental Protection. Prepared by: Tetra Tech Center for Ecological Sciences, Montpelier VT.

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

Stoddard, J.L., A. T. Herlihy, D. V. Peck, R. M. Hughes, T. R. Whittier, and E. Tarquinio. 2008. A process for creating multimetric indices for large-scale aquatic surveys. J. N. Am. Benthol. Soc., 2008, 27(4):878–891

Stribling, J.B., B.K. Jessup, and E.W. Leppo. 2016. The Mississippi-Benthic Index of Stream Quality (M-BISQ): Recalibration and Testing. Prepared for: Mississippi Department of Environmental Quality, Office of Pollution Control, P.O. Box 2261, Jackson, Mississippi 39225. Prepared by: Tetra Tech, Inc., Center for Ecological Sciences, Owings Mills, Maryland, and Montpelier, Vermont

Terrell, C.R. and P.B. Perfetti. 1996. Water quality indicators guide. Surface waters. Terrene Institute. Washington, D.C.

Tetra Tech. 2015. Illinois Stream Macroinvertebrate Multimetric Index Development. Prepared for the U.S. Environmental Protection Agency, Region 5, Chicago, IL and the Illinois Environmental Protection Agency, Springfield, IL.

Thornbrugh, D. J., Leibowitz, S.G., Hill, R. A., Weber, M. H., Johnson, Z.C. Olsen, A. R., Flotemersch, J. E., Stoddard, J. L., & Peck, D. V. 2018. Mapping watershed integrity for the conterminous United States. Ecological Indicators, 85, 1133-1148.

U.S. Environmental Protection Agency (U.S. EPA). 2012. Implications of Climate Change for Bioassessment Programs and Approaches to Account for Effects. Global Change Research Program, National Center for Environmental Assessment, Washington, DC; EPA/600/R-11/036F. Available from the National Technical Information Service, Springfield, VA, and online at http://www.epa.gov/ncea. Authorship: Anna Hamilton, Jen Stamp, Mike Paul, Jeroen Gerritsen, Lei Zheng, Erik Leppo and Britta Bierwagen.

U.S. Environmental Protection Agency (U.S. EPA). 2013. Biological Assessment Program Review: Assessing Level of Technical Rigor to Support Water Quality Management. EPA 820-R-13-001. U.S. Environmental Protection Agency, Washington, DC.

U.S. Environmental Protection Agency (U.S. EPA). 2016. Regional Monitoring Networks (RMNs) to detect changing baselines in freshwater wadeable stream. (EPA/600/R-15/280). Washington, DC: Office of Research and Development, Washington. Available online at http://www.epa.gov/ncea. Authorship: Jen Stamp, Anna Hamilton, Britta G. Bierwagen, Debbie Arnwine, Margaret Passmore and Jonathan Witt.

Van Sickle, J., and R.M. Hughes. 2000. Classification strengths of ecoregions, catchments, and geographic clusters for aquatic vertebrates in Oregon. Journal of the North American Benthological Society 19(3):370-384.

Vieira, Nicole K.M., Poff, N. LeRoy, Carlisle, Daren M., Moulton, Stephen R., II, Koski, Marci L. and Kondratieff, Boris C., 2006, A database of lotic invertebrate traits for North America: U.S. Geological Survey Data Series 187, http://pubs.water.usgs.gov/ds187.

Weigel, B.M. 2003. Development of stream macroinvertebrate models that predict watershed and local stressors in Wisconsin. Journal of the North American Benthological Society 22: 123-142.

Weiskel, P.K., Brandt, S.L., DeSimone, L.A., Ostiguy, L.J., and Archfield, S.A., 2010, Indicators of streamflow alteration, habitat fragmentation, impervious cover, and water quality for Massachusetts stream basins: U.S. Geological Survey Scientific Investigations Report 2009–5272, 70 p., plus CD–ROM. (Also available at http://pubs.usgs.gov/sir/2009/5272/)

Yoder, C. O., & Rankin, E. T. 1995. Biological criteria program development and implementation in Ohio. In W. S. Davis & T. P. Simon (Eds.), Biological assessment and criteria: tools for water resource planning and decision making (pp. 109–144). Boca Raton: Lewis Publishers.

Appendix A

1. Macroinvertebrate Data Preparation Steps
2. List of Taxonomists IDing MassDEP samples over time
3. Non-target Taxa
4. Excluded Taxa Decision Criteria

A1 Data preparation steps

Step 1 – Compile master taxa list

Compile list of unique taxa IDs from the four datasets (MassDEP, DRWA, CT DEEP, RI DEM)

- 1475 original TaxaIDs

Check for differences in nomenclature, misspellings, etc. (are the same taxa being called different things?)

Examples of changes that were made:

- Ablabesmyia sp. -> Ablabesmyia
- Baetis (cerci only) sp. -> Baetis
- Ceratopsyche -> Hydropsyche

Check TaxaIDs against trusted sources

- EPA NRSA Benthic Master Taxa table (provided by Richard Mitchell, personal communication 12/5/2018).
- USGS BioData https://my.usgs.gov/confluence/display/biodata/BioData+Taxonomy+Downloads
- Integrated Taxonomic Information System (itis.gov)

Look for temporal patterns

- Examine taxa matrices. Do certain taxa appear or disappear over time?
- Calculate metrics and examine scatterplots of metric values vs. year. Do metric values show noticeable increases or decreases over time?
- Create NMDS ordinations with taxa and metrics. Code samples by year. Are there any noticeable patterns?
- If yes, are these due to -
    - Changes in taxonomists? (see Section A2 for list of taxonomists over time)
    - Changes in taxonomic keys?
        - Example: Conchapelopia, Helopelopia, Meropelopia, Rheopelopia & Telopelopia - names changed over time; changed to Thienemannimyia genus group
    - Sampling locations? During some years, reference sites were targeted and EPT metric values were higher. During other years, efforts were focused in eastern MA (where streams have higher levels of disturbance). See 'SamplingLocationVsYr' maps in the Supplemental materials

Look for spatial patterns

- Create taxa distribution maps (see Supplemental materials). Do certain taxa only occur in certain parts of MA? Do some only occur in one state vs. all three (MA, RI and CT)? If so, why? (Is the pattern real or does it stem from differences in nomenclature?

Reconcile differences and assign FinalID

- Used multiple lines of evidence (where available) to inform decisions. Decisions were reviewed and approved by Bob Nuzzo, MassDEP. 1320 Final IDs in regional dataset (MA, CT, RI). 661 taxa in MA IBI dataset only.
- Original TaxaIDs were retained in "taxa translator" tables so that the data can be linked back to original source files if needed.

A1 continued...

Step 2 – Assemble Benthic Master Taxa table

- Add in phylogenic information (Phylum, Class, Order, Family, Subfamily, Tribe, Genus, Species)
- Add in Rank (level of taxonomic resolution)
- Designate NonTarget taxa (taxa that should be excluded from index calculations, in keeping with MassDEP's existing RBP III guidance – see Section A3). Examples include terrestrials, surface dwellers and crayfish.
- Add in Attributes
    - MA master taxa table - functional feeding group (FFG), tolerance values
    - In addition, obtained master taxa and traits tables from other entities
        - VT DEC, CT DEEP, EPA NRSA, Poff et al. 2016, Viera et al. 2016
        - FFG, tolerance value, habit, life cycle/voltinism and thermal preferences (see Table 2 in main report)
    - Evaluate level of agreement across sources (higher level of agreement = greater confidence in assignment). Used MassDEP assignments as primary, but sometimes revised based on multiple lines of evidence (majority rules). Asked Bob Nuzzo (MassDEP) to review the attribute assignments if there were discrepancies; went with his discretion if he had a strong opinion, otherwise used majority rules.
    - As a secondary check of Tolerance Value assignments, we looked at the distribution of taxa across disturbance categories (see Supplemental materials). If a taxon that is designated as intolerant occurs in a relatively high number of stressed samples, Tt brought this to the attention of MassDEP biologists and we reevaluated the tolerance designation for that taxon
    - Fill in blanks where appropriate (e.g., if one species within a genera was missing an attribute assignment but all the other species within that genera had the same attribute assignment, Tetra Tech assigned the same attribute to that species; if there were discrepancies, Tetra Tech did not make an assignment
    - Sources of FFG, tolerance value and habit attribute assignments were recorded in the Master Taxa table

Step 3 – Assemble Benthic table

- Rarify/randomly subsample data as needed (IBI calibration was based on 100-count samples). MassDEP rarifed some of the samples and Tetra Tech rarified some of the samples. Tetra Tech used the Rarify function in the BioMonTools R package (https://github.com/leppott/BioMonTools) to do the subsampling.
- The rarified samples were added to the database as new BenSampIDs (with .R at the end - e.g., 2014056.R).
- Tetra Tech ran the 'markExcluded' function in the BioMonTools R package (https://github.com/leppott/BioMonTools) to identify redundant taxa on a sample-by-sample basis; taxa marked as 'TRUE' in the Exclude column were not counted in richness calculations. Section A4 contains the decision criteria that was used to exclude taxa.

Step 4 – Select samples for the IBI calibration dataset

- Exclude samples that occurred outside of MA
- If a site had multiple years of data, select one sample per site (randomly)
- Exclude samples if there were fewer than 80 or greater than 120 total individuals (100 ± 20%)
- Exclude samples if taxonomic resolution was too coarse (samples with Chironomids to subfamily or tribe-level were not used)
- Limited IBI calibration and verification dataset to RBP kick net samples collected in July-September, from 2000 onward in the Central Hills and Western Highlands

Step 5 – Assess need for Operational Taxonomic Units (OTUs) (Cuffney et al. 2007)

- Evaluate whether any taxonomic groups need to be collapsed to a higher level of taxonomic resolution due to inconsistencies over time (e.g., should mites be collapsed to Order-level, Chironomids to tribe or subfamily-level or worms to family-level?)
    - OTU1 - lowest practical (available) level. Calculated number of species- and genus-level calls within the major Orders – did these change noticeably over time? (no, not in the IBI cal/verif dataset)
    - Tried an alternate OTU (lowest practical level, except mites were collapsed to Order-level and Chironomids were collapsed to genus level). Differences (if any) were very small between metrics calculated using OTU1 vs. OTU2, so we decided to use OTU1.

Step 6 - Calculate metrics

- A query was set up in the MS Access database to generate an output that was run through the Metric Calculation function in the BioMonTools R package (https://github.com/leppott/BioMonTools). Appendix B contains the list of metrics that were calculated and considered as candidates for inclusion in the IBIs.

A2 Taxonomists over time (MassDEP dataset)

CollYear	ID by
1983	R. Nuzzo
1984	A. Johnson
	Johnson
	R. Nuzzo
1985	B. Prynoski
	R. Nuzzo
	R. Nuzzo, B. Prynoski
	R. Nuzzo, M. Tasse
1986	R. Nuzzo, B. Prynoski
	R. Nuzzo, B. Prysnoski
	R. Nuzzo, M. Tasse
1987	R. Nuzzo
	R. Nuzzo, S. Grubbs
	S. Grubbs
1988	J. Fiorentino, S. Grubbs
	R. Nuzzo, S. Grubbs
1989	R. Nuzzo, S. Grubbs
	S. Grubbs
1990	R. Nuzzo
1991	G. Szal
1992	G. Szal
	R. Nuzzo

1993	J. Fiorentino
	R. Nuzzo
1994	G. Szal
	R. Nuzzo
1995	J. Fiorentino
1996	J. Fiorentino
	R. Nuzzo
1997	J. Fiorentino
	R. Nuzzo
1998	J. Fiorentino
	N/A
	R. Nuzzo
	S. Gaughan
1999	J. Fiorentino
2000	J. Fiorentino
	R. Nuzzo
2001	K. Curry, et al.
	other
	R. Nuzzo
2002	A.R. Williams
	J. Fiorentino
	other
	P. Mitchell
2003	Mitchell
	other
	P. Mitchell
	R. Nuzzo
2004	J. Fiorentino
	other
	P. Mitchell
	P.Mitchell
	R. Nuzzo
2005	J. Fiorentino
	M. Cole
	P. Mitchell
	R. Nuzzo
2006	J. Fiorentino
	M. Cole
	R. Nuzzo
2007	M. Cole
	M.Cole
	R. Nuzzo
2008	M. Cole
2009	M. Cole
2010	EcoAnalysts

	EcoAnaysts
2011	M. Cole
2012	M. Cole
	other
	R. Nuzzo
2013	EcoAnalysts
	M. Cole
	other
2014	M. Cole
2015	M. Cole
2016	M. Cole
2017	M. Cole

A3 Non-target taxa (provided by Bob Nuzzo, MassDEP)

In general terms, sample processing involves separating macroinvertebrates from other materials in the sample. For the purposes of this document "macroinvertebrate" is defined to include:

- *all aquatic Annelida;*
- *all aquatic Mollusca;*
- *aquatic macro Crustacea (except as noted below);*
- *all aquatic Arachnida (Oribatid mites, which I've been told are not truly aquatic; exclusion not yet captured in the SOP revision); and*
- *the aquatic life stages of Insecta except Hemiptera and adult Coleoptera other than Elmidae.*

Those macroinvertebrates excluded from the above list are not used for one of three reasons: either there is insufficient ecological information on them to make them useful for biomonitoring, they are surface film dwellers, or they are capable of escaping the aquatic environment at will to avoid temporarily unfavorable conditions. One further exception is crayfish (Class Crustacea, Family Cambaridae), which often are seen evacuating the immediate area as kick-sampling begins, and even swimming out of the kick-net. Crayfish species are noted when present in the sample but are not counted toward total numbers.

A4 Excluded Taxa Decision Criteria (Tetra Tech 2019)

When calculating metrics for benthic macroinvertebrates there are occasions when certain taxa are not included in taxa richness metrics but the individuals are included for all other metrics. This is done to avoid double counting taxa that may have been identified to a more coarse level when taxa of a finer level are present in the same sample.

These taxa have been referred to by many names – e.g., Excluded Taxa, NonUnique Taxa, or Ambiguous Taxa. This document will use the term Excluded.

When you run the markExcluded function, redundant taxa are excluded based on the following steps:

1. Calculate and find all taxa names that appear in a sample at each taxonomic rank more than once (for an example, see Figure 1). These are the potential "parents" to be excluded.

2. Check if any of the potential "parents" equal a final ID in their respective samples.

3. If you get a match these are marked as "Excluded"

All Excluded decisions are sample-specific and the rules should be reapplied if sample contents change. Also, if the level of effort or operational taxonomic units change, the Excluded taxa designations should be recalculated.

TAXA LIST

BCG Attribute	FinalID	Count	FFG	Thermal	Toler_Sed	Redundant	Excluded
4	*Nais*	7	NA	--	NA	FALSE	FALSE
4	*Atractides*	1	PR	--	NA	FALSE	FALSE
4	*Hygrobates*	3	PR	--	NA	FALSE	FALSE
4	*Lebertia*	6	PR	--	NA	FALSE	FALSE
4	*Sperchon*	2	PR	--	NA	FALSE	FALSE
3	*Torrenticola*	1	PR	--	NA	FALSE	FALSE
4	*Dytiscidae*	3	PR	--	NA	TRUE	FALSE
3	*Oreodytes*	1	PR	--	NA	FALSE	FALSE
3	*Heterlimnius corpulentus*	19	GC	--	5	FALSE	FALSE
3	*Narpus concolor*	2	GC	--	5	FALSE	FALSE
3	*Clinocera*	1	PR	--	NA	FALSE	FALSE
4	*Neoplasta*	1	NA	--	NA	FALSE	FALSE
2	*Glutops*	2	PR	--	NA	FALSE	FALSE
x	*Ceratopogoninae*	2	PR	--	NA	FALSE	FALSE
4	*Thienemannimyia group*	9	PR	--	NA	FALSE	FALSE
4	*Micropsectra*	19	GC	--	NA	FALSE	FALSE

... Data_Taxa_Master | Data_Metrics | Data_Habitat | Data_Taxa_Samps | Samp0001 | Samp0009 | Samp0018

Figure 1. Example - Dytiscidae (family-level) is excluded from the richness metrics in this sample because these organisms could be the same taxon as Oreodytes (genus-level). The exclusion rule is applied on a sample by sample basis.

Below is a more detailed description of the process that the markExcluded function follows. Before starting, it is necessary to have a complete and correct master taxa list (all phylogenetic information and ranks).

Terminology

- Target Rank = intended level of taxonomy for identification, e.g., genus. Typically, specified in the project's SOP but can be adjusted during the OTU process.

- Parent or Parent Taxon = a taxon that occurs in the data in addition to other taxa in the same group that are identified to a more specific level. For example, the family Baetidae may occur in the data in addition to genera within the family Baetidae. In this case the name Baetidae is a parent to the other taxa within the family. Parents do not have to be only a single rank above the child taxon. That is, the class and order ranks are parents of any family ranks within them.
- Child or Children Taxa = a taxa or taxon that occurs in the data in addition to individuals identified to a coarser level. For example, the genera Baetis and Procloeon may occur in addition to the family Baetidae (of which the 2 genera listed are a member). In this case Baetis and Procloen are children of Baetidae.

Rule Development

For each sample:

1. Determine "potential" taxa for exclusion based on rank (or level) names appearing more than once in a sample.
    a. This is done for all ranks present; phylum, class, order, family, tribe, genus, species.
2. Check if any "potential" taxa are equal to a final (unique) ID in the same sample.
3. Stage is combined with taxa names if used in the dataset.

Requirements

1. A sample taxa table or data frame.
    a. All non-count and zero individual taxa have been removed.
    b. Unique sample ID code in a single column.
    c. A column with a final identification that is narrative not numeric. That is, Baetidae is ok but the ITIS number is not.
    d. Phylogenetic rank/level columns.
        i. This can be applied from a master taxa table but needs to be included in this table. One column per rank.
        ii. Names need to be consistently spelled.

Procedures

1. Find all potential Parents (those with a rank coarser than the target rank). This is done by creating a list of taxa rank names that appear more than once in a sample. This is done for each taxonomic rank.
2. The above list is compared to the final identifications for each sample.
    a. Special consideration is made for ranks of finer detail than genus. That is, names that are a combination of more than one field.
3. Any matches are marked as "Excluded".

There is still a need for manual review / QC check of the final list of Excluded designations.

Appendix B

Macroinvertebrate Metrics

Table B1. List of candidate macroinvertebrate metrics that were calculated with the BioMonTools R package (https://github.com/leppott/BioMonTools).

Metric Name	Metric Category	Description
nt_total	RICH	number of taxa - total
pi_Amph	COMP	percent individuals - Order Amphipoda
nt_Chiro	RICH	number of taxa - Family Chironomidae
pi_Chiro	COMP	percent individuals - Family Chironomidae
pt_Chiro	RICH	percent taxa - Family Chironomidae
nt_Coleo	RICH	number of taxa - Order Coleoptera
pi_Coleo	COMP	percent individuals - Order Coleoptera
pt_Coleo	RICH	percent taxa - Order Coleoptera
nt_Dipt	RICH	number of taxa - Order Diptera
pi_Dipt	COMP	percent individuals - Order Diptera
pt_Dipt	RICH	percent taxa - Order Diptera
nt_Ephem	RICH	number of taxa - Order Ephemeroptera
pi_Ephem	COMP	percent individuals - Order Ephemeroptera
pt_Ephem	RICH	percent taxa - Order Ephemeroptera
pi_EphemNoCaeBae	COMP	percent individuals - Order Ephemeroptera, excluding Families Caenidae and Baetidae
nt_EPT	RICH	number of taxa - Orders Ephemeroptera, Plecoptera & Trichoptera (EPT)
pi_EPT	COMP	percent individuals - Orders Ephemeroptera, Plecoptera & Trichoptera (EPT)
pt_EPT	RICH	percent taxa - Orders Ephemeroptera, Plecoptera & Trichoptera (EPT)
pi_Hydro	COMP	percent individuals - Family Hydropsychidae
nt_Insect	RICH	number of taxa - Class Insecta
pi_Insect	COMP	percent individuals - Class Insecta
pt_Insect	RICH	percent taxa - Class Insecta
nt_NonIns	RICH	number of taxa - Class not Insecta
pi_NonIns	COMP	percent individuals - Class not Insecta
pt_NonIns	RICH	percent taxa - Class not Insecta
nt_Oligo	RICH	number of taxa - Class Oligochaeta
pi_Oligo	COMP	percent individuals - Class Oligochaeta
pt_Oligo	RICH	percent taxa - Class Oligochaeta
nt_Pleco	RICH	number of taxa - Order Plecoptera
pi_Pleco	COMP	percent individuals - Order Plecoptera
pt_Pleco	RICH	percent taxa - Order Plecoptera
nt_POET	RICH	number of taxa - Orders Plecoptera, Odonata, Ephemeroptera & Trichoptera (POET)
pt_POET	RICH	percent taxa - Orders Plecoptera, Odonata, Ephemeroptera & Trichoptera (POET)
nt_Trich	RICH	number of taxa - Order Trichoptera
pi_Trich	COMP	percent individuals - Order Trichoptera
pt_Trich	RICH	percent taxa - Order Trichoptera
pi_TricNoHydro	COMP	percent individuals - Order Trichoptera, excluding Family Hydropsychidae

nt_ti_cc	TEMP	number of taxa - thermal indicator - cold/cool
pi_ti_cc	TEMP	percent individuals - thermal indicator - cold/cool
pt_ti_cc	TEMP	percent taxa - thermal indicator - cold/cool
nt_ti_w	TEMP	number of taxa - thermal indicator - warm
pi_ti_w	TEMP	percent individuals - thermal indicator - warm
pt_ti_w	TEMP	percent taxa - thermal indicator - warm
nt_tv_intol	TOLER	number of taxa - tolerance value - intolerant ≤ 3
pi_tv_intol	TOLER	percent individuals - tolerance value - intolerant ≤ 3
pt_tv_intol	TOLER	percent taxa - tolerance value - intolerant ≤ 3
nt_tv_toler	TOLER	number of taxa - tolerance value -tolerant ≥ 7
pi_tv_toler	TOLER	percent individuals - tolerance value -tolerant ≥ 7
pt_tv_toler	TOLER	percent taxa - tolerance value -tolerant ≥ 7
nt_ffg_col	FFG	number of taxa - Functional Feeding Group (FFG) - collector-gatherer (CG)
nt_ffg_filt	FFG	number of taxa - Functional Feeding Group (FFG) - collector-filterer (CF)
nt_ffg_pred	FFG	number of taxa - Functional Feeding Group (FFG) - predator (PR)
nt_ffg_scrap	FFG	number of taxa - Functional Feeding Group (FFG) - scraper (SC)
nt_ffg_shred	FFG	number of taxa - Functional Feeding Group (FFG) - shredder (SH)
pi_ffg_col	FFG	percent individuals - Functional Feeding Group (FFG) - collector-gatherer (CG)
pi_ffg_filt	FFG	percent individuals - Functional Feeding Group (FFG) - collector-filterer (CF)
pi_ffg_pred	FFG	percent individuals - Functional Feeding Group (FFG) - predator (PR)
pi_ffg_scrap	FFG	percent individuals - Functional Feeding Group (FFG) - scraper (SC)
pi_ffg_shred	FFG	percent individuals - Functional Feeding Group (FFG) - shredder (SH)
pt_ffg_col	FFG	percent taxa - Functional Feeding Group (FFG) - collector-gatherer (CG)
pt_ffg_filt	FFG	percent taxa - Functional Feeding Group (FFG) - collector-filterer (CF)
pt_ffg_pred	FFG	percent taxa - Functional Feeding Group (FFG) - predator (PR)
pt_ffg_scrap	FFG	percent taxa - Functional Feeding Group (FFG) - scraper (SC)
pt_ffg_shred	FFG	percent taxa - Functional Feeding Group (FFG) - shredder (SH)
nt_habit_burrow	HABIT	number of taxa - Habit - burrowers (BU)
nt_habit_climb	HABIT	number of taxa - Habit - climbers (CB)
nt_habit_cling	HABIT	number of taxa - Habit - clingers (CN)
nt_habit_sprawl	HABIT	number of taxa - Habit - sprawlers (SP)
nt_habit_swim	HABIT	number of taxa - Habit - swimmers (SW)
pi_habit_burrow	HABIT	percent individuals - Habit - burrowers (BU)
pi_habit_climb	HABIT	percent individuals - Habit - climbers (CB)
pi_habit_cling	HABIT	percent individuals - Habit - clingers (CN)
pi_habit_sprawl	HABIT	percent individuals - Habit - sprawlers (SP)
pi_habit_swim	HABIT	percent individuals - Habit - swimmers (SW)
pt_habit_burrow	HABIT	percent taxa - Habit - burrowers (BU)
pt_habit_climb	HABIT	percent taxa - Habit - climbers (CB)
pt_habit_cling	HABIT	percent taxa - Habit - clingers (CN)
pt_habit_sprawl	HABIT	percent taxa - Habit - sprawlers (SP)

pt_habit_swim	HABIT	percent taxa - Habit - swimmers (SW)
nt_volt_multi	VOLT	number of taxa - multivoltine (MULTI)
nt_volt_semi	VOLT	number of taxa - semivoltine (SEMI)
nt_volt_uni	VOLT	number of taxa - univoltine (UNI)
pi_volt_multi	VOLT	percent individuals - multivoltine (MULTI)
pi_volt_semi	VOLT	percent individuals - semivoltine (SEMI)
pi_volt_uni	VOLT	percent individuals - univoltine (UNI)
pt_volt_multi	VOLT	percent taxa - multivoltine (MULTI)
pt_volt_semi	VOLT	percent taxa - semivoltine (SEMI)
pt_volt_uni	VOLT	percent taxa - univoltine (UNI)
pi_dom01	RICH	percent individuals - most dominant taxon [max(N_TAXA)]
pi_dom02	RICH	percent individuals - two most dominant taxa
pi_dom03	RICH	percent individuals - three most dominant taxa
pi_dom04	RICH	percent individuals - four most dominant taxa
pi_dom05	RICH	percent individuals - five most dominant taxa
x_Becks	TOLER	Becks Biotic Index = 2*[C1Taxa]+[C2Taxa] (see footnote)
x_HBI	TOLER	Hilsenhoff Biotic Index (references the TolVal field)
x_Shan_2	RICH	Shannon Wiener Diversity Index (log base 2) - x_Shan_Num/log(2)
x_D	RICH	Simpson's Index
x_Evenness	RICH	Evenness=x_Shan_e/log(nt_total)

Appendix C

StreamCat & NHDPlusV2 data

Data sources

The National Hydrography Dataset (NHD) Plus Version 2 (NHDPlusV2) (McKay et al. 2012) is a publicly available geospatial framework that depicts a network of approximately 2.65 million streams and rivers within the conterminous U.S. based on digitized lines of U.S. Geological Survey (USGS) topographic quadrangle maps. NHDPlusV2 data were downloaded from this website: http://www.horizon-systems.com/NHDPlus/NHDPlusV2_01.php

In 2016, EPA published the Stream-Catchment (StreamCat) Dataset (Hill et al. 2016), which is an extensive database of natural and anthropogenic landscape metrics that are associated with NHDPlusV2 stream segments. These metrics are statistical summaries of GIS layers (e.g., climate data, land cover data). StreamCat data are available at two spatial scales: local catchment and full upstream watershed. StreamCat data were downloaded from this website: http://www2.epa.gov/nationalaquatic-resource-surveys/streamcat.

Table C1 lists the StreamCat metrics that were considered during development of the disturbance gradient in MA. The StreamCat data include the Indices of Catchment and Watershed Integrity (ICI and IWI). We used version 1 of the ICI and IWI (Thornbrugh et al. 2018) when developing the disturbance gradient because version 2.1 (Johnson et al. 2019) did not become available until after the disturbance gradient had been established. When version 2.1 was released, we did a comparison of the two sets of scores (version 1 vs. version 2.1). Overall, relative rankings of sites were similar; where differences arose, the version 2.1 IWI scores were generally lower. If the difference in scores affected the disturbance category designations, Tt investigated those sites further by looking at aerial imagery and consulting MassDEP staff. Adjustments were made to the final disturbance category designations as deemed appropriate.

In addition to the StreamCat data, the following NHDPlusV2 attribute data were also included in the dataset:

- FTYPE & FCODE
    - Definition: flowline feature attributes (see Table C2 for descriptions)
    - Source: \NHDSnapshot\Hydrography\NHDFlowline (dbf)
- SLOPE
    - Definition: slope of flowline (meters/meters) based on smoothed elevations
    - Source: \NHDPlusAttributes\ElevSlope (dbf)

Table C1. StreamCat metrics (Hill et al. 2016) that were associated with the macroinvertebrate sampling sites and considered during development of the disturbance gradient. Metrics were either taken directly from StreamCat (Source: StreamCat) or calculated based on a combination of StreamCat variables (Source: Derived). More detailed documentation of the StreamCat dataset is available online: http://www2.epa.gov/nationalaquatic-resource-surveys/streamcat

Variable	Description	Source
CatAreaSqKm	Area of local NHDPlus catchment (square km)	StreamCat
WsAreaSqKm	Watershed area (square km) at NHDPlus stream segment outlet, i.e., at the most downstream location of the vector line segment	StreamCat
ICI	Index of catchment integrity	StreamCat
IWI	Index of watershed integrity	StreamCat
CHYD	Hydrologic regulation component score calculated using catchment metrics	StreamCat
CCHEM	Regulation of water chemistry component score calculated using catchment metrics	StreamCat
CSED	Sediment regulation component score calculated using catchment metrics	StreamCat
CCONN	Hydrologic connectivity component score calculated using catchment metrics	StreamCat
CTEMP	Temperature regulation component score calculated using catchment metrics	StreamCat
CHABT	Habitat provision component score calculated using catchment metrics	StreamCat
WHYD	Hydrologic regulation component score calculated using watershed metrics	StreamCat
WCHEM	Regulation of water chemistry component score calculated using watershed metrics	StreamCat
WSED	Sediment regulation component score calculated using watershed metrics	StreamCat
WCONN	Hydrologic connectivity component score calculated using watershed metrics	StreamCat
WTEMP	Temperature regulation component score calculated using watershed metrics	StreamCat
WHABT	Habitat provision component score calculated using watershed metrics	StreamCat
PctUrbLMH2011Cat	% of catchment area classified as developed, high + medium + low-intensity land use (NLCD 2011 class 24+23+22)	Derived
PctUrbLMH2011Ws	% of watershed area classified as developed, high + medium + low-intensity land use (NLCD 2011 class 24+23+22)	Derived
PctUrbLMH2011CatRp100	% of catchment area classified as developed, high + medium + low-intensity land use (NLCD 2011 class 24+23+22) within a 100-m buffer of NHD streams	Derived
PctUrbLMH2011WsRp100	% of watershed area classified as developed, high + medium + low-intensity land use (NLCD 2011 class 24+23+22) within a 100-m buffer of NHD streams	Derived
PctHayCrop2011Cat	% of catchment area classified as hay and crop land use (NLCD 2011 class 82+81)	Derived

PctHayCrop2011Ws	% of watershed area classified as hay and crop land use (NLCD 2011 class 82+81)	Derived
PctHayCrop2011CatRp100	% of catchment area classified as hay and crop land use (NLCD 2011 class 82+81) within a 100-m buffer of NHD streams	Derived
PctHayCrop2011WsRp100	% of watershed area classified as hay and crop land use (NLCD 2011 class 82+81) within a 100-m buffer of NHD streams	Derived
PctFrst2011Cat	% of catchment area classified as deciduous + evergreen + mixed deciduous/evergreen forest land cover (NLCD 2011 class 41+42+43)	Derived
PctFrst2011WS	% of watershed area classified as deciduous + evergreen + mixed deciduous/evergreen forest land cover (NLCD 2011 class 41+42+43)	Derived
PctFrst2011CatRp100	% of catchment area classified as deciduous + evergreen + mixed deciduous/evergreen forest land cover (NLCD 2011 class 41+42+43) within a 100-m buffer of NHD streams	Derived
PctFrst2011WsRp100	% of watershed area classified as deciduous + evergreen + mixed deciduous/evergreen forest land cover (NLCD 2011 class 41+42+43) within a 100-m buffer of NHD streams	Derived
PctWetWat2011Cat	% of catchment area classified as open water + herbaceous wetland + woody wetland land cover (NLCD 2011 class 11+95+90)	Derived
PctWetWat2011Ws	% of watershed area classified as open water + herbaceous wetland + woody wetland land cover (NLCD 2011 class 11+95+90)	Derived
PctWetWat2011CatRp100	% of catchment area classified as open water + herbaceous wetland + woody wetland land cover (NLCD 2011 class 11+95+90) within a 100-m buffer of NHD streams	Derived
PctWetWat2011WsRp100	% of watershed area classified as open water + herbaceous wetland + woody wetland land cover (NLCD 2011 class 11+95+90) within a 100-m buffer of NHD streams	Derived
PctShrb2011Cat	% of catchment area classified as shrub/scrub land cover (NLCD 2011 class 52)	StreamCat
PctShrb2011Ws	% of watershed area classified as shrub/scrub land cover (NLCD 2011 class 52)	StreamCat
PctGrs2011Cat	% of catchment area classified as grassland/herbaceous land cover (NLCD 2011 class 71)	StreamCat
PctGrs2011Ws	% of watershed area classified as grassland/herbaceous land cover (NLCD 2011 class 71)	StreamCat
PctHay2011Cat	% of catchment area classified as hay land use (NLCD 2011 class 81)	StreamCat
PctHay2011Ws	% of watershed area classified as hay land use (NLCD 2011 class 81)	StreamCat
PctHay2011CatRp100	% of catchment area classified as hay land use (NLCD 2011 class 81) within a 100-m buffer of NHD streams	StreamCat

PctHay2011WsRp100	% of watershed area classified as hay land use (NLCD 2011 class 81) within a 100-m buffer of NHD streams	StreamCat
PctCrop2011Cat	% of catchment area classified as crop land use (NLCD 2011 class 82)	StreamCat
PctCrop2011Ws	% of watershed area classified as crop land use (NLCD 2011 class 82)	StreamCat
PctCrop2011CatRp100	% of catchment area classified as crop land use (NLCD 2011 class 82) within a 100-m buffer of NHD streams	StreamCat
PctCrop2011WsRp100	% of watershed area classified as crop land use (NLCD 2011 class 82) within a 100-m buffer of NHD streams	StreamCat
PctImp2011Cat	Mean imperviousness of anthropogenic surfaces (NLCD 2011) within catchment	StreamCat
PctImp2011Ws	Mean imperviousness of anthropogenic surfaces (NLCD 2011) within watershed	StreamCat
PctImp2011CatRp100	Mean imperviousness of anthropogenic surfaces (NLCD 2011) within catchment and within a 100-m buffer of NHD stream lines	StreamCat
PctImp2011WsRp100	Mean imperviousness of anthropogenic surfaces (NLCD 2011) within watershed and within a 100-m buffer of NHD stream lines	StreamCat
HUDen2010Cat	Mean housing unit density (housing units/square km) within catchment and within a 100-m buffer of NHD stream lines	StreamCat
HUDen2010Ws	Mean housing unit density (housing units/square km) within watershed and within a 100-m buffer of NHD stream lines	StreamCat
PopDen2010Cat	Mean populating density (people/square km) within catchment	StreamCat
PopDen2010Ws	Mean populating density (people/square km) within watershed	StreamCat
CBNFCat	Mean rate of biological nitrogen fixation from the cultivation of crops in kg N/ha/yr, within catchment	StreamCat
CBNFWs	Mean rate of biological nitrogen fixation from the cultivation of crops in kg N/ha/yr, within watershed	StreamCat
FertCat	Mean rate of synthetic nitrogen fertilizer application to agricultural land in kg N/ha/yr, within the catchment	StreamCat
FertWs	Mean rate of synthetic nitrogen fertilizer application to agricultural land in kg N/ha/yr, within watershed	StreamCat
ManureCat	Mean rate of manure application to agricultural land from confined animal feeding operations in kg N/ha/yr, within catchment	StreamCat
ManureWs	Mean rate of manure application to agricultural land from confined animal feeding operations in kg N/ha/yr, within watershed	StreamCat
AllAgNCat	[CBNFCat]+[FertCat]+[ManureCat]	Derived
AllAgNWs	[CBNFWs]+[FertWs]+[ManureWs]	Derived
AllMineDensCat	[MineDensCat]+[CoalMineDensCat]; Density of mines + coal mines within catchment (mines/square km)	Derived

AllMineDensWs	[MineDensWs]+[CoalMineDensWs]; Density of mines + coal mines within watershed (mines/square km)	Derived
RdDensCat	Density of roads (2010 Census Tiger Lines) within catchment (km/square km)	StreamCat
RdDensWs	Density of roads (2010 Census Tiger Lines) within watershed (km/square km)	StreamCat
RdDensCatRp100	Density of roads (2010 Census Tiger Lines) within catchment and within a 100-m buffer of NHD stream lines (km/square km)	StreamCat
RdDensWsRp100	Density of roads (2010 Census Tiger Lines) within watershed and within a 100-m buffer of NHD stream lines (km/square km)	StreamCat
RdCrsCat	Density of roads-stream intersections (2010 Census Tiger Lines-NHD stream lines) within catchment (crossings/square km)	StreamCat
RdCrsWs	Density of roads-stream intersections (2010 Census Tiger Lines-NHD stream lines) within watershed (crossings/square km)	StreamCat
RdCrsSlpWtdCat	Density of roads-stream intersections (2010 Census Tiger Lines-NHD stream lines) multiplied by NHDPlusV21 slope within catchment (crossings*slope/square km)	StreamCat
RdCrsSlpWtdWs	Density of roads-stream intersections (2010 Census Tiger Lines-NHD stream lines) multiplied by NHDPlusV21 slope within watershed (crossings*slope/square km)	StreamCat
DamDensCat	Density of georeferenced dams within catchment (dams/ square km)	StreamCat
DamDensWs	Density of georeferenced dams within watershed (dams/ square km)	StreamCat
DamNIDStorCat	Volume all reservoirs (NORM_STORA in NID) per unit area of catchment (cubic meters/square km)	StreamCat
DamNIDStorWs	Volume all reservoirs (NORM_STORA in NID) per unit area of watershed (cubic meters/square km)	StreamCat
DamNrmStorCat	Volume all reservoirs (NID_STORA in NID) per unit area of catchment (cubic meters/square km)	StreamCat
DamNrmStorWs	Volume all reservoirs (NID_STORA in NID) per unit area of watershed (cubic meters/square km)	StreamCat
NABD_DensCat	Density of georeferenced dams within catchment (dams/ square km)	StreamCat
NABD_DensWs	Density of georeferenced dams within watershed (dams/ square km)	StreamCat
NABD_NrmStorCat	Volume all reservoirs (NORM_STORA in NID) per unit area of catchment (cubic meters/square km)	StreamCat
NABD_NrmStorWs	Volume all reservoirs (NORM_STORA in NID) per unit area of watershed (cubic meters/square km)	StreamCat
NABD_NIDStorCat	Volume all reservoirs (NID_STORA in NID) per unit area of catchment (cubic meters/square km)	StreamCat
NABD_NIDStorWs	Volume all reservoirs (NID_STORA in NID) per unit area of watershed (cubic meters/square km)	StreamCat

NPDESDensCat	Density of permitted NPDES (National Pollutant Discharge Elimination System) sites within catchment (sites/square km)	StreamCat
NPDESDensWs	Density of permitted NPDES (National Pollutant Discharge Elimination System) sites within watershed (sites/square km)	StreamCat
NPDESDensCatRp100	Density of permitted NPDES (National Pollutant Discharge Elimination System) sites within catchment and within a 100-m buffer of NHD stream lines (sites/square km)	StreamCat
NPDESDensWsRp100	Density of permitted NPDES (National Pollutant Discharge Elimination System) sites within watershed and within 100-m buffer of NHD stream lines (sites/square km)	StreamCat
SuperfundDensCat	Density of Superfund sites within catchment (sites/square km)	StreamCat
SuperfundDensWs	Density of Superfund sites within watershed (sites/square km)	StreamCat
SuperfundDensCatRp100	Density of Superfund sites within 100-m buffer of NHD stream lines within the catchment (sites/square km)	StreamCat
SuperfundDensWsRp100	Density of Superfund sites within watershed and within a 100-m buffer of NHD stream lines (sites/square km)	StreamCat
TRIDensCat	Density of TRI (Toxic Release Inventory) sites within catchment (sites/square km)	StreamCat
TRIDensWs	Density of TRI (Toxic Release Inventory) sites within watershed (sites/square km)	StreamCat
TRIDensCatRp100	Density of TRI (Toxic Release Inventory) sites within 100-m buffer of NHD stream lines in the catchment (sites/square km)	StreamCat
TRIDensWsRp100	Density of TRI (Toxic Release Inventory) sites within watershed and within a 100-m buffer of NHD stream lines (sites/square km)	StreamCat
ElevCat	Mean catchment elevation (m)	StreamCat
ElevWs	Mean watershed elevation (m)	StreamCat
BFICat	Base flow is the component of streamflow that can be attributed to ground-water discharge into streams. The BFI is the ratio of base flow to total flow, expressed as a percentage, within catchment	StreamCat
BFIWs	Base flow is the component of streamflow that can be attributed to ground-water discharge into streams. The BFI is the ratio of base flow to total flow, expressed as a percentage, within watershed	StreamCat
RunoffCat	Mean runoff (mm) within catchment	StreamCat
RunoffWs	Mean runoff (mm) within watershed	StreamCat
Precip8110Cat	PRISM climate data - 30-year normal mean precipitation (mm): Annual period: 1981-2010 within the catchment	StreamCat
Precip8110Ws	PRISM climate data - 30-year normal mean precipitation (mm): Annual period: 1981-2010 within the watershed	StreamCat

Tmin8110Cat	PRISM climate data - 30-year normal minimum temperature (C°): Annual period: 1981-2010 within the catchment	StreamCat
Tmin8110Ws	PRISM climate data - 30-year normal minimum temperature (C°): Annual period: 1981-2010 within the watershed	StreamCat
Tmean8110Cat	PRISM climate data - 30-year normal mean temperature (C°): Annual period: 1981-2010 within the catchment	StreamCat
Tmean8110Ws	PRISM climate data - 30-year normal mean temperature (C°): Annual period: 1981-2010 within the watershed	StreamCat
Tmax8110Cat	PRISM climate data - 30-year normal maximum temperature (C°): Annual period: 1981-2010 within the catchment	StreamCat
Tmax8110Ws	PRISM climate data - 30-year normal maximum temperature (C°): Annual period: 1981-2010 within the watershed	StreamCat
WetIndexCat	Mean Composite Topographic Index (CTI)[Wetness Index] within catchment	StreamCat
WetIndexWs	Mean Composite Topographic Index (CTI)[Wetness Index] within watershed	StreamCat
AgKffactCat	Mean soil erodibility (Kf) factor (unitless) of soils within catchment on agricultural land. The Kffactor is used in the Universal Soil Loss Equation (USLE) and represents a relative index of susceptibility of bare, cultivated soil to particle detachment and transport by rainfall.	StreamCat
AgKffactWs	Mean soil erodibility (Kf) factor (unitless) of soils within watershed on agricultural land. The Kffactor is used in the Universal Soil Loss Equation (USLE) and represents a relative index of susceptibility of bare, cultivated soil to particle detachment and transport by rainfall.	StreamCat
KffactCat	Mean soil erodibility (Kf) factor (unitless) of soils within catchment. The Kffactor is used in the Universal Soil Loss Equation (USLE) and represents a relative index of susceptibility of bare, cultivated soil to particle detachment and transport by rainfall.	StreamCat
KffactWs	Mean soil erodibility (Kf) factor (unitless) of soils within watershed. The Kffactor is used in the Universal Soil Loss Equation (USLE) and represents a relative index of susceptibility of bare, cultivated soil to particle detachment and transport by rainfall.	StreamCat
Al2O3Cat	Mean % of lithological aluminum oxide (Al2O3) content in surface or near surface geology within catchment	StreamCat
Al2O3Ws	Mean % of lithological aluminum oxide (Al2O3) content in surface or near surface geology within watershed	StreamCat
CaOCat	Mean % of lithological calcium oxide (CaO) content in surface or near surface geology within catchment	StreamCat
CaOWs	Mean % of lithological calcium oxide (CaO) content in surface or near surface geology within watershed	StreamCat
Fe2O3Cat	Mean % of lithological ferric oxide (Fe2O3) content in surface or near surface geology within catchment	StreamCat

Fe2O3Ws	Mean % of lithological ferric oxide (Fe2O3) content in surface or near surface geology within watershed	StreamCat
K2OCat	Mean % of lithological potassium oxide (K2O) content in surface or near surface geology within catchment	StreamCat
K2OWs	Mean % of lithological potassium oxide (K2O) content in surface or near surface geology within watershed	StreamCat
MgOCat	Mean % of lithological magnesium oxide (MgO) content in surface or near surface geology within catchment	StreamCat
MgOWs	Mean % of lithological magnesium oxide (MgO) content in surface or near surface geology within watershed	StreamCat
Na2OCat	Mean % of lithological sodium oxide (Na2O) content in surface or near surface geology within catchment	StreamCat
Na2OWs	Mean % of lithological sodium oxide (Na2O) content in surface or near surface geology within watershed	StreamCat
P2O5Cat	Mean % of lithological phosphorous oxide (P2O5) content in surface or near surface geology within catchment	StreamCat
P2O5Ws	Mean % of lithological phosphorous oxide (P2O5) content in surface or near surface geology within watershed	StreamCat
SCat	Mean % of lithological sulfur (S) content in surface or near surface geology within catchment	StreamCat
SWs	Mean % of lithological sulfur (S) content in surface or near surface geology within watershed	StreamCat
SiO2Cat	Mean % of lithological silicon dioxide (SiO2) content in surface or near surface geology within catchment	StreamCat
SiO2Ws	Mean % of lithological silicon dioxide (SiO2) content in surface or near surface geology within watershed	StreamCat
NCat	Mean % of lithological nitrogen (N) content in surface or near surface geology within catchment	StreamCat
NWs	Mean % of lithological nitrogen (N) content in surface or near surface geology within watershed	StreamCat

Table C2. Descriptions of the flowline feature attributes in the NHDPlusV2 dataset (McKay et al. 2012).

FCODE	Description
55800	Artificial Path
33603	Canal Ditch: Canal Ditch Type = Stormwater
33600	Canal/Ditch
33601	Canal/Ditch: Canal/Ditch Type = Aqueduct
56600	Coastline
33400	Connector
42800	Pipeline
42801	Pipeline: Pipeline Type = Aqueduct; Relationship to Surface = At or Near
42802	Pipeline: Pipeline Type = Aqueduct; Relationship to Surface = Elevated
42803	Pipeline: Pipeline Type = Aqueduct; Relationship to Surface = Underground
42807	Pipeline: Pipeline Type = General Case; Relationship to Surface = Underground
42809	Pipeline: Pipeline Type = Penstock; Relationship to Surface = At or Near
42811	Pipeline: Pipeline Type = Penstock; Relationship to Surface = Underground
42813	Pipeline: Pipeline Type = Siphon
42823	Pipeline: Pipeline Type = Stormwater; Relationship to Surface = Underground
46000	Stream/River
46007	Stream/River: Hydrographic Category = Ephemeral
46003	Stream/River: Hydrographic Category = Intermittent
46006	Stream/River: Hydrographic Category = Perennial

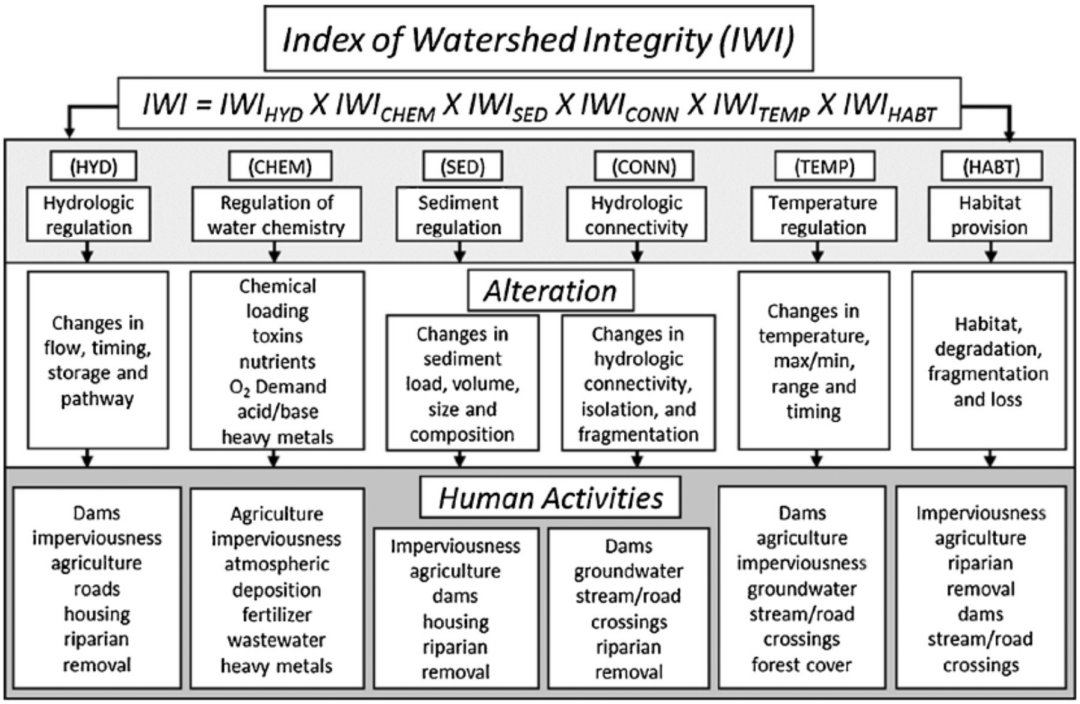


Figure C1. Conceptual model of the calculation of the IWI. Source: Thornbrugh et al. 2018 (Fig. 1).

Lit Cited

Flotemersch, J.E., Leibowitz, S.G., Hill, R.A., Stoddard, J.L., Thomas, M.C., & Tharme, R.E. 2016. A watershed Integrity Definition and Assessment approach to Support Strategic Management of Watersheds. River Research and Applications, 32, 1654–1671.

Hill, R.A., Weber, M.H., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., 2016. The Stream-Catchment (StreamCat) dataset: a database of watershed metrics for the Conterminous United States. J. Am. Water Res. Assoc. 52 (1), 120–128

Johnson, Zachary & G. Leibowitz, Scott & Hill, Ryan. (2018). Revising the index of watershed integrity national maps. Science of The Total Environment. 10.1016/j.scitotenv.2018.10.112.

McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., Reah, A., 2012. NHDPlus Version 2: User Guide. U.S. Environmental Protection Agency

Thornbrugh, D. J., Leibowitz, S.G., Hill, R. A., Weber, M. H., Johnson, Z.C. Olsen, A. R., Flotemersch, J. E., Stoddard, J. L., & Peck, D. V. 2018. Mapping watershed integrity for the conterminous United States. Ecological Indicators, 85, 1133-1148.

Appendix D

Comparison of Human Disturbance Index (HDI) scores with Index of Watershed and Catchment Integrity scores (IWI, ICI)

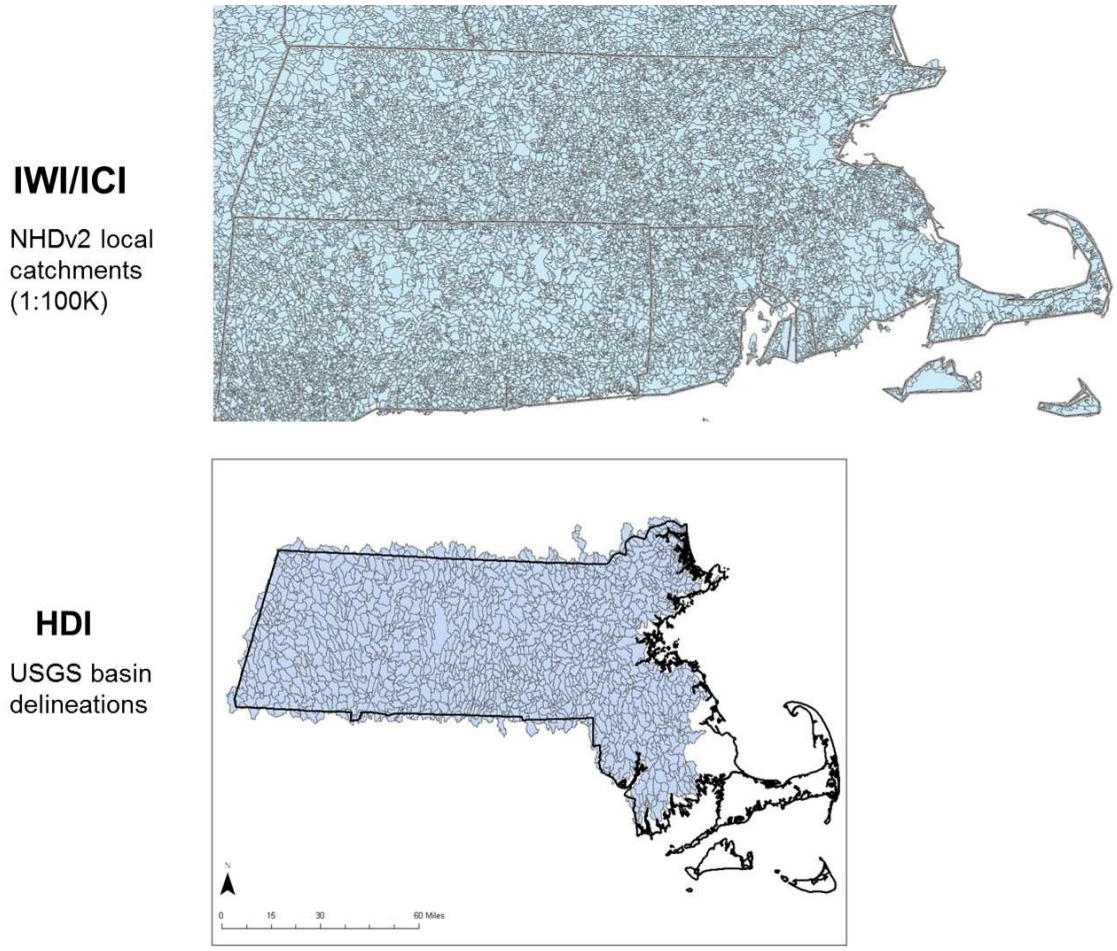


Figure D1. Comparison of IWI/ICI (top) and HDI (bottom) catchment delineations. The NHDPlusV2 geospatial dataset (with 1:100K resolution) is the basis of the IWI/ICI catchment delineations. The NHDPlusV2 catchments have slightly finer spatial resolution than the HDI basin delineations.

Human Disturbance Index (HDI)

Scaled from 1 to 5; 5 is most disturbed

Disturbance indicators used to calculate the human disturbance index (Weiskel 2010).

Indicator	Definition
Streamflow Alteration	
August Flow	Percentage of alteration of the August median flows for period of record
Water-Use Intensity	Ratio of overall water use in a subwatershed (withdrawals + wastewater discharges) to unaffected mean annual subwatershed outflow.
Dam Storage Ratio	Ratio of maximum impounded subwatershed storage to long-term mean annual outflow, in days divided by watershed size
Landscape	
Impervious Cover	Percentage of impervious cover in the cumulative watershed.
Local Impervious Cover	Percentage of impervious cover in a 61 meter stream buffer in the pour point HUC
Agriculture	Percentage of area with agricultural land uses in the cumulative watershed
Local Agriculture	Percentage of impervious cover in a 61 meter stream buffer in the pour point HUC

StreamCat metrics used in the IWI/ICI (version 1)

Metrics in green correspond to metrics in the HDI

Article description	HYD	CHEM	SED	CONN	TEMP	HABT
Presence and volumes of reservoirs (NABD)	x	x	x	x	x	x
Stream channelization and levee construction (NA)	x	x	x	x		
Road/stream intersections (TIGER/NHD) weighted by stream reach slope (NHD)				x		
Percent of watershed composed of agricultural land uses (NLCD)	x				x	x
Total length and density of canals/ditches (NHD)	x					
Percent imperviousness of human-related landscapes (NLCD)	x					
Alteration to and spatial arrangement of riparian vegetation (LANDFIRE)	x	x	x	x	x	x
Boundaries, depths, and flows of aquifers (NA)	x					
Groundwater use (NA)	x			x	x	
Atmospheric deposition of anthropogenic sources of nitrogen and acid rain (NADP)		x				
Percent of watershed composed of urban or agricultural land uses (NLCD)		x				
Fertilizer application rates (FERT)		x				
Presence and density of wastewater treatment facilities (NPDES), industrial facilities (TRI), superfund sites (SUPERFUND), or mines (MINES)		x				
Cattle density (NA)		x				
Chemical constituents of groundwater (NA)		x				
Presence and density of mines (MINES), forest cover loss (GFC), and roads (TIGER)			x			
Agriculture (NLCD) weighted by soil erodibility (CONUS-SOIL)			x			
Density of ditches/canals (NHD)				x		
Presence and density of wastewater discharge sites (NPDES)				x	x	
Percent of riparian zone composed of urban or agricultural land uses (NLCD)				x		
Percent of watershed composed of urban land uses in the riparian zone (NLCD)					x	
Density of housing unit developments within riparian zones (TIGER)						x
Density of road/stream intersections (TIGER/NHD)						x
Density of roads within riparian zones (TIGER)						x

Figure D2. Comparison of input metrics in the HDI (left) (Weiskel et al. 2010) and ICI and IWI version 1 (right) (Thornbrugh et al. 2018).

Preliminary reference designations

Grouped by Level III ecoregion

Number of sites

EPA Level III ecoregion	Prelim Status	HDI									
		1	1.5	2	2.5	3	3.5	4	4.5	5	NA
Northeastern Coastal Zone	Ref	5		13	1	9	1	2			57
Northeastern Coastal Zone					17	82	49	139	113	99	539
Northeastern Highlands	Ref	19	7	62	1	1					34
Northeastern Highlands					19	73	36	13	5	1	159

Mean IWI scores (total watershed) in each HDI category

EPA Level III ecoregion	Prelim Status	HDI									
		1	1.5	2	2.5	3	3.5	4	4.5	5	NA
Northeastern Coastal Zone	Ref	0.91		0.86	0.89	0.83	0.82	0.84			0.85
Northeastern Coastal Zone					0.84	0.80	0.77	0.72	0.69	0.58	0.76
Northeastern Highlands	Ref	0.89	0.87	0.85	0.90	0.84					0.86
Northeastern Highlands					0.85	0.80	0.78	0.75	0.75	0.79	0.80

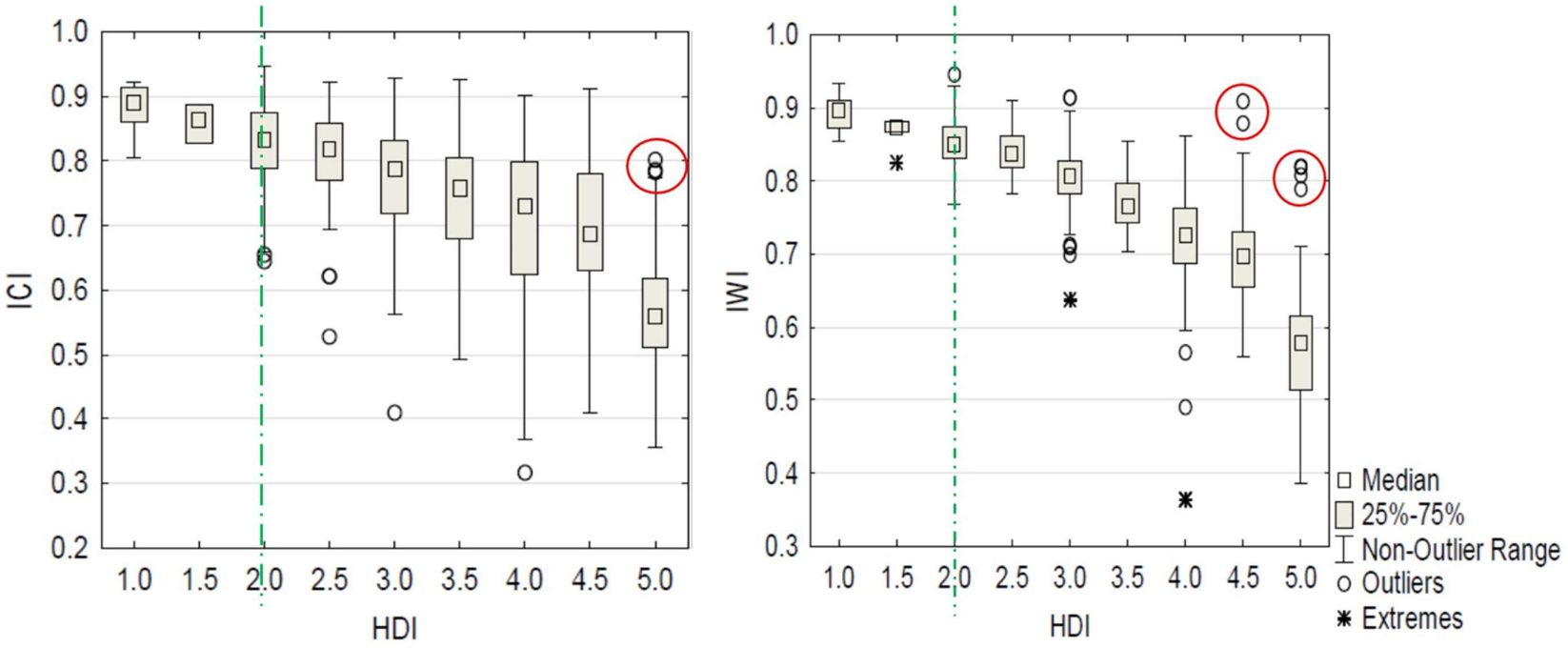
Mean ICI scores (local) in each HDI category

EPA Level III ecoregion	Prelim Status	HDI									
		1	1.5	2	2.5	3	3.5	4	4.5	5	NA
Northeastern Coastal Zone	Ref	0.89		0.85	0.92	0.83	0.81	0.87			0.84
Northeastern Coastal Zone					0.78	0.76	0.75	0.70	0.69	0.57	0.74
Northeastern Highlands	Ref	0.89	0.86	0.82	0.90	0.79					0.85
Northeastern Highlands					0.82	0.77	0.74	0.71	0.68	0.74	0.79

Figure D3. Relationships between preliminary reference designations and HDI and IWI/ICI (version 1) scores, grouped by Level III ecoregions. HDI scores are scaled from 1 to 5, with 5 being the most disturbed. ICI and IWI scores are scaled from 0 to 1, with higher values having greater integrity/better watershed condition.

Box plots - HDI vs ICI & IWI

Overall, the pattern was as expected (with some exceptions/outliers); the lower HDI scores (=less disturbed conditions) generally corresponded with higher ICI and IWI scores (higher values = greater integrity/better watershed condition).
The green vertical dotted line at HDI = 2 corresponds with a preliminary threshold that was used to designate reference sites in MA.



Source data: all sites with HDI scores in the two main Level III ecoregions
(Northeastern Highlands (code 58) + Northeastern Coastal Zone (code 59))

Figure D4. Box plots showing the distribution of ICI and IWI scores across the HDI score categories, with all sites (from both Level III ecoregions) combined.

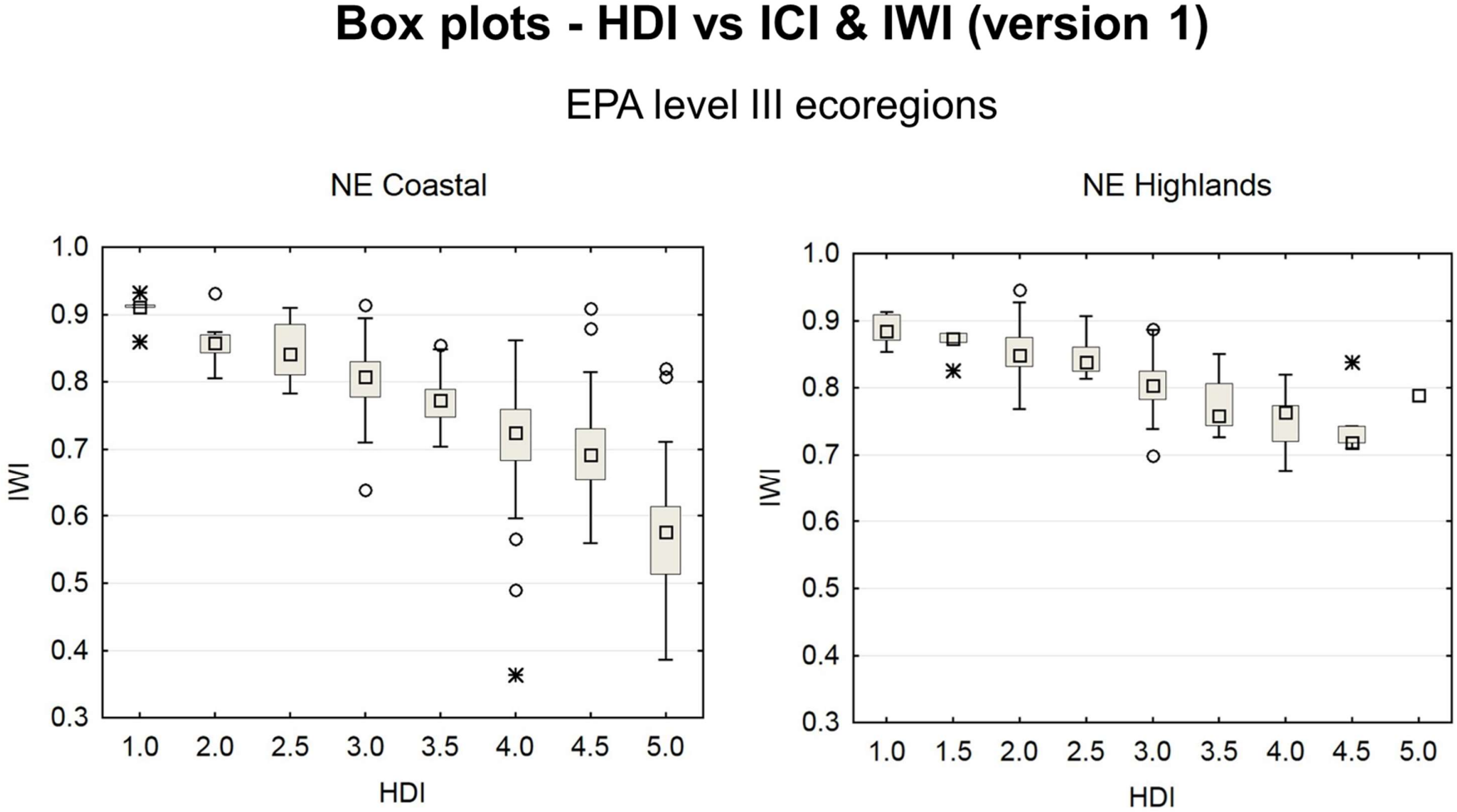


Figure D5. Box plots showing the distribution of IWI scores across the HDI score categories in each Level III ecoregion.

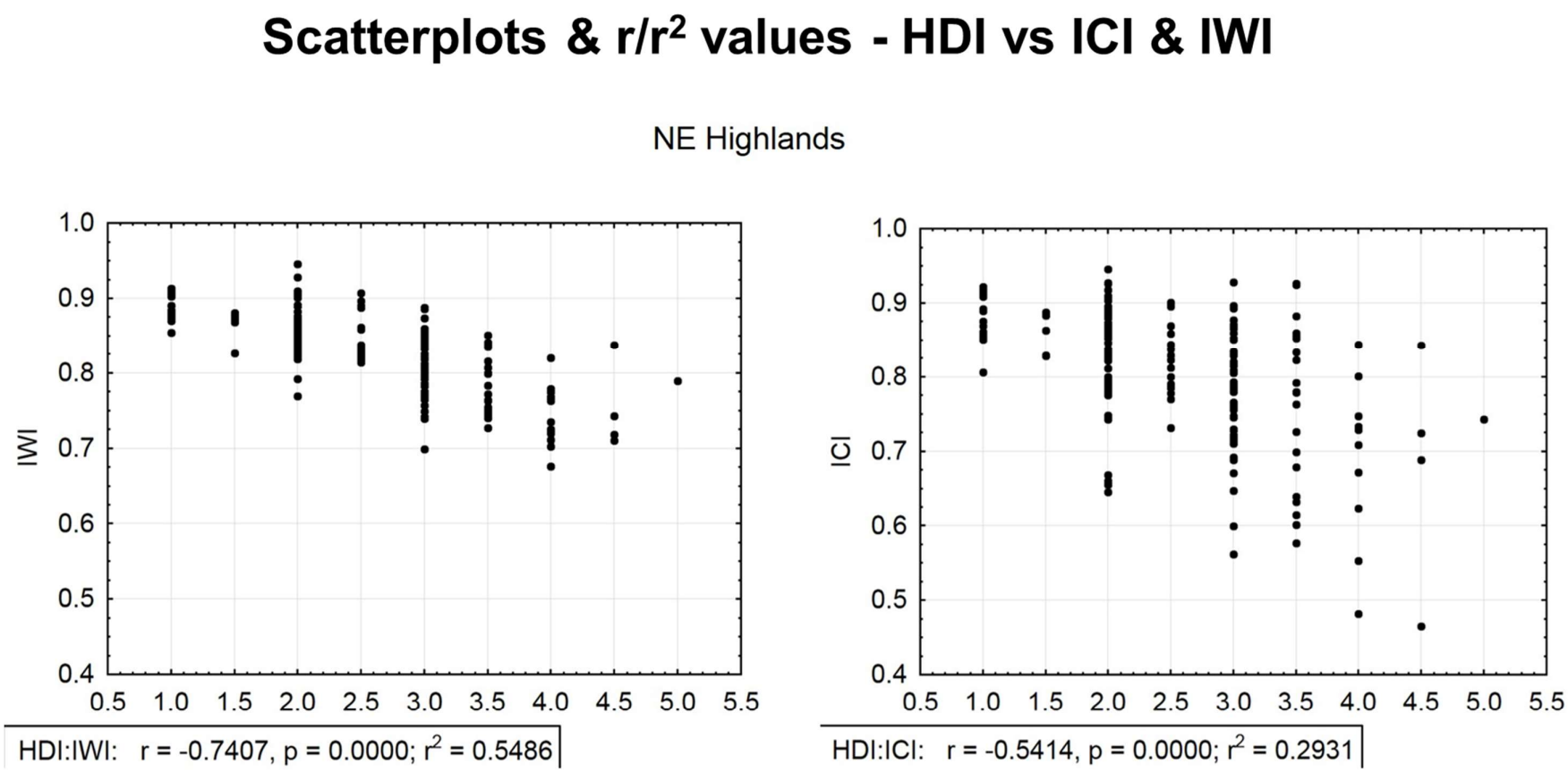


Figure D6. Scatterplots showing IWI and ICI scores vs HDI scores in the Northeastern Highlands Level III ecoregion. Correlation coefficients and r2 values are also included.

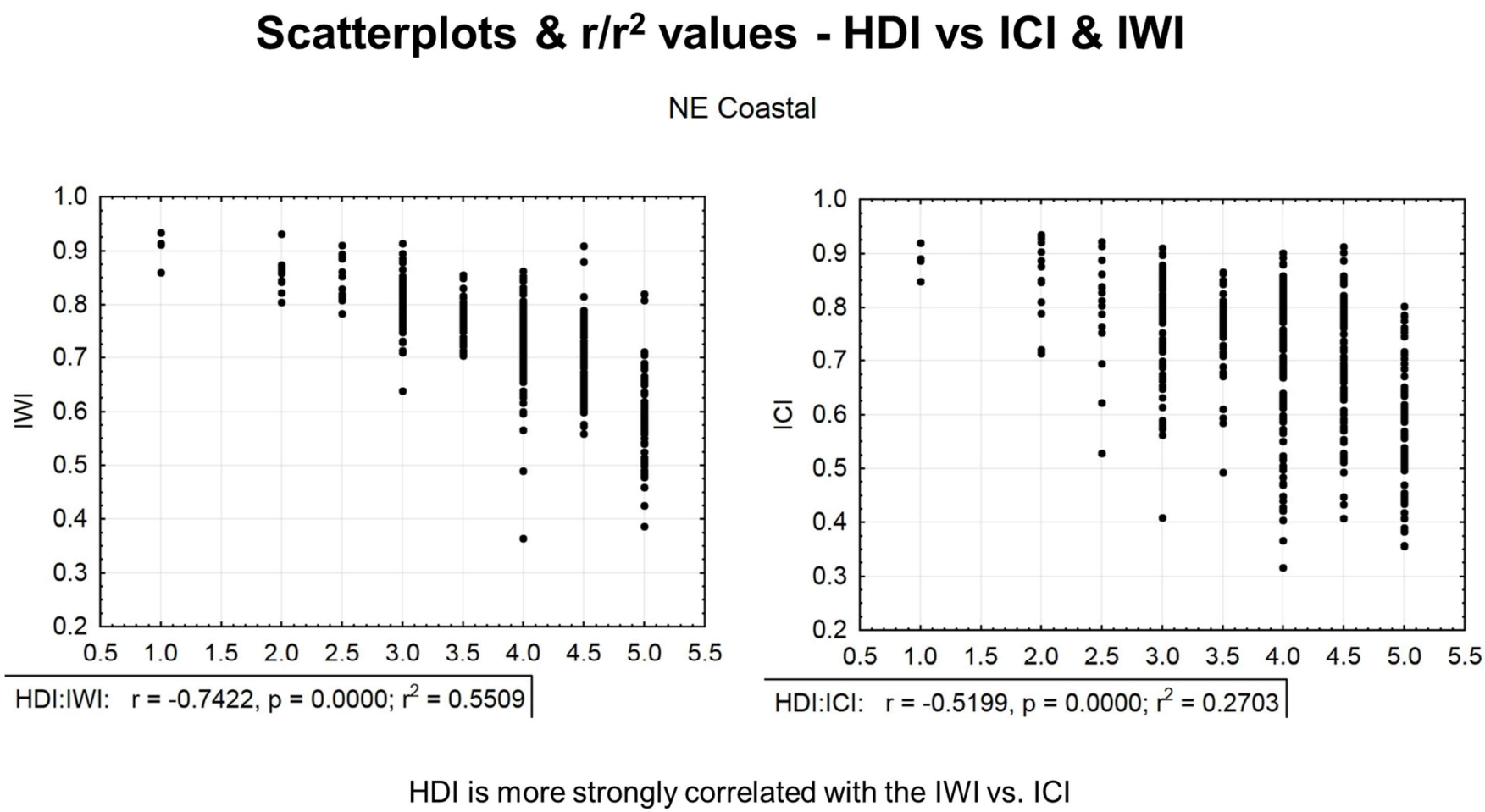


Figure D7. Scatterplots showing IWI and ICI scores vs HDI scores in the Northeastern Coastal Level III ecoregion. Correlation coefficients and r2 values are also included.

Appendix E

List of reference and stressed samples that were used for IBI calibration and verification in the Central Hills and Western Highlands

Central Hills IBI *calibration* samples (reference and stressed)

Table E1. Reference and stressed samples that were used to calibrate the Central Hills IBI calibration. The samples were collected with the RBP kicknet method from 2000 onward.

Class	Entity	Unique_ID	BenSampID	Collection Method	Collection Date	Status	Disturbance category	Longitude	Latitude	Level IV ecoregion
Central Hills	MA_DEP	B0900	2014033.R	RBP kicknet	2014-07-24	Ref	1_BestRef	-72.373514	42.387326	58g
Central Hills	MA_DEP	B0699	2015041.R	RBP kicknet	2015-07-28	Ref	1_BestRef	-72.357188	42.587089	58g
Central Hills	MA_DEP	B0526	2004035	RBP kicknet	2004-08-27	Ref	1_BestRef	-71.846692	42.062144	59c
Central Hills	MA_DEP	B0548	2004015	RBP kicknet	2004-08-24	Ref	2_Ref	-72.13195	42.0391	59b
Central Hills	MA_DEP	B0891	2014012.R	RBP kicknet	2014-07-10	Ref	2_Ref	-72.368693	42.442071	58g
Central Hills	MA_DEP	B0944	2016007.AR	RBP kicknet	2016-07-25	Ref	2_Ref	-72.26846	42.161712	59b
Central Hills	MA_DEP	B0653	2008044	RBP kicknet	2008-09-04	Ref	2_Ref	-72.224755	42.479448	59b
Central Hills	MA_DEP	B0867	2014005.R	RBP kicknet	2014-07-07	Ref	2_Ref	-71.751964	42.706137	59h
Central Hills	MA_DEP	B0736	2015040.R	RBP kicknet	2015-07-28	Ref	2_Ref	-72.384545	42.464706	58g
Central Hills	MA_DEP	B0821	2014060.R	RBP kicknet	2014-08-12	Ref	2_Ref	-72.227184	42.695393	58g
Central Hills	MA_DEP	B0654	2008045	RBP kicknet	2008-09-04	Ref	2_Ref	-72.187273	42.443434	59b
Central Hills	MA_DEP	B0706	2011021	RBP kicknet	2011-07-12	Ref	2_Ref	-72.195029	42.62532	58g
Central Hills	MA_DEP	B0445	2000015	RBP kicknet	2000-09-11	Ref	2_Ref	-72.466338	42.57641	58g
Central Hills	MA_DEP	B0648	2008055	RBP kicknet	2008-09-10	Ref	3_SubRef	-72.259335	42.153512	59b
Central Hills	MA_DEP	B0870	2014032.R	RBP kicknet	2014-07-24	Ref	3_SubRef	-72.431373	42.349683	58g
Central Hills	MA_DEP	B0715	2011015	RBP kicknet	2011-07-07	Ref	3_SubRef	-72.40184	42.61498	58g
Central Hills	MA_DEP	B0600	2006025	RBP kicknet	2006-07-25	Ref	3_SubRef	-70.779752	42.191834	59f
Central Hills	MA_DEP	B0713	2011031	RBP kicknet	2011-07-18	Ref	3_SubRef	-72.187518	42.057726	59b
Central Hills	MA_DEP	B0720	2011016	RBP kicknet	2011-07-07	Ref	3_SubRef	-72.231027	42.556236	59b
Central Hills	MA_DEP	B0530	2004017	RBP kicknet	2004-08-24	Ref	3_SubRef	-72.16363	42.060875	59b
Central Hills	MA_DEP	B0901	2014011.R	RBP kicknet	2014-07-10	Ref	3_SubRef	-72.322599	42.469994	58g
Central Hills	MA_DEP	B0539	2004016	RBP kicknet	2004-08-24	Ref	3_SubRef	-72.157307	42.033794	59b
Central Hills	MA_DEP	B0737	2014018.R	RBP kicknet	2014-07-15	Ref	3_SubRef	-72.161586	42.034815	59b
Central Hills	MA_DEP	B0743	2013059.R	RBP kicknet	2013-08-08	Ref	3_SubRef	-71.837751	42.383816	59h
Central Hills	MA_DEP	B0450	2005088.1	RBP kicknet	2005-09-13	Ref	3_SubRef	-72.121627	42.595083	58g

Table E1 continued…

Class	Entity	Unique_ID	BenSampID	Collection Method	Collection Date	Status	Disturbance category	Longitude	Latitude	Level IV ecoregion
Central Hills	MA_DEP	B0557	2005091	RBP kicknet	2005-09-13	Ref	3_SubRef	-72.123136	42.592346	58g
Central Hills	MA_DEP	B0514	2008023	RBP kicknet	2008-07-21	Ref	3_SubRef	-72.481277	42.378563	59a
Central Hills	MA_DEP	B0449	2000024A.1	RBP kicknet	2000-09-13	Ref	3_SubRef	-72.179348	42.68846	58g
Central Hills	MA_DEP	B0508	2003009	RBP kicknet	2003-07-22	Ref	3_SubRef	-72.512579	42.415533	59a
Central Hills	MA_DEP	B0446	2000017	RBP kicknet	2000-09-11	Ref	3_SubRef	-72.437622	42.598139	58g
Central Hills	MA_DEP	B0694	2010018	RBP kicknet	2010-07-19	Strs	7_HighStrs	-71.252721	42.546252	59h
Central Hills	MA_DEP	B0199	2001005	RBP kicknet	2001-07-05	Strs	7_HighStrs	-71.343895	42.359101	59h
Central Hills	MA_DEP	B0636	2008022	RBP kicknet	2008-07-21	Strs	7_HighStrs	-72.427215	42.281853	59b
Central Hills	MA_DEP	B0614	2007011	RBP kicknet	2007-07-17	Strs	7_HighStrs	-71.360766	42.187033	59c
Central Hills	MA_DEP	B0704	2011027	RBP kicknet	2011-07-14	Strs	7_HighStrs	-71.856173	42.202911	59c
Central Hills	MA_DEP	B0744	2009030	RBP kicknet	2009-07-20	Strs	7_HighStrs	-71.011578	42.257022	59d
Central Hills	MA_DEP	B0474	2001034	RBP kicknet	2001-09-04	Strs	7_HighStrs	-72.751831	42.14923	59a
Central Hills	MA_DEP	B0745	2009031	RBP kicknet	2009-07-20	Strs	7_HighStrs	-70.997025	42.248219	59c
Central Hills	MA_DEP	B0198	2001003	RBP kicknet	2001-07-03	Strs	7_HighStrs	-71.530367	42.385487	59h
Central Hills	MA_DEP	B0917	2015007.R	RBP kicknet	2015-07-13	Strs	7_HighStrs	-71.196718	42.390098	59d
Central Hills	MA_DEP	B0757	2009042	RBP kicknet	2009-08-07	Strs	7_HighStrs	-71.172253	42.394225	59d
Central Hills	MA_DEP	B0681	2010013.A	RBP kicknet	2010-07-14	Strs	7_HighStrs	-70.920712	42.52482	59d
Central Hills	MA_DEP	B0703	2011014	RBP kicknet	2011-07-06	Strs	7_HighStrs	-71.756915	42.526618	59h
Central Hills	MA_DEP	B0665	2008034	RBP kicknet	2008-08-05	Strs	7_HighStrs	-71.882046	42.365314	59h
Central Hills	MA_DEP	B0667	2008035	RBP kicknet	2008-08-05	Strs	7_HighStrs	-71.751737	42.529673	59h
Central Hills	MA_DEP	B0097	2008014	RBP kicknet	2008-07-09	Strs	7_HighStrs	-71.803135	42.241737	59h
Central Hills	MA_DEP	B0754	2009038	RBP kicknet	2009-08-06	Strs	7_HighStrs	-71.140341	42.457763	59d
Central Hills	MA_DEP	B0763	2009025	RBP kicknet	2009-07-15	Strs	7_HighStrs	-71.253172	42.148564	59c
Central Hills	MA_DEP	B0098	2008016.A	RBP kicknet	2008-07-09	Strs	7_HighStrs	-71.838378	42.234408	59h
Central Hills	MA_DEP	B0762	2009024	RBP kicknet	2009-07-14	Strs	7_HighStrs	-71.187894	42.179096	59d
Central Hills	MA_DEP	B0130	2009043	RBP kicknet	2009-08-07	Strs	7_HighStrs	-71.158546	42.417835	59d

Table E1 continued…

Class	Entity	Unique_ID	BenSampID	Collection Method	Collection Date	Status	Disturbance category	Longitude	Latitude	Level IV ecoregion
Central Hills	MA_DEP	B0859	2013040.R	RBP kicknet	2013-07-29	Strs	7_HighStrs	-71.169094	42.421301	59d
Central Hills	MA_DEP	B0509	2003015	RBP kicknet	2003-07-23	Strs	7_HighStrs	-72.654506	42.319697	59a
Central Hills	MA_DEP	B0721	2011022	RBP kicknet	2011-07-12	Strs	7_HighStrs	-71.829957	42.578712	59h
Central Hills	MA_DEP	B0076	2003017.1	RBP kicknet	2003-09-03	Strs	7_HighStrs	-71.768544	42.561889	59h
Central Hills	MA_DEP	B0484	2001012	RBP kicknet	2001-07-19	Strs	7_HighStrs	-71.439527	42.289836	59h
Central Hills	MA_DEP	B0711	2011024	RBP kicknet	2011-07-13	Strs	7_HighStrs	-72.02541	42.076742	59c
Central Hills	MA_DEP	B0059	2007018	RBP kicknet	2007-07-19	Strs	7_HighStrs	-71.188419	42.365257	59d
Central Hills	MA_DEP	B0587	2006024	RBP kicknet	2006-07-10	Strs	7_HighStrs	-71.302439	42.633438	59h
Central Hills	MA_DEP	B0086	2008042	RBP kicknet	2008-08-26	Strs	7_HighStrs	-71.551328	42.715824	59h

Central Hills IBI *verification* samples (reference and stressed)

Table E2. Reference and stressed samples that were used to verify the Central Hills IBI.

Class	Entity	Unique_ID	BenSampID	Collection Method	Collection Date	Status	Disturbance category	Longitude	Latitude	Level IV ecoregion
Central Hills	MA_DEP	B0527	2004034	RBP kicknet	2004-08-27	Ref	1_BestRef	-71.848399	42.062281	59c
Central Hills	MA_DEP	B0659	2008009	RBP kicknet	2008-07-08	Ref	2_Ref	-71.719714	42.016789	59c
Central Hills	MA_DEP	B0829	2013047.AR	RBP kicknet	2013-08-01	Ref	2_Ref	-71.70506	42.00792	59c
Central Hills	MA_DEP	B0448	2000022	RBP kicknet	2000-09-12	Ref	2_Ref	-72.1145	42.681872	58g
Central Hills	MA_DEP	B0440	2000010	RBP kicknet	2000-07-19	Ref	3_SubRef	-70.903712	42.660748	59f
Central Hills	MA_DEP	B0542	2004030	RBP kicknet	2004-08-27	Ref	3_SubRef	-71.904709	42.204192	59b
Central Hills	MA_DEP	B0670	2005099	RBP kicknet	2005-09-14	Ref	3_SubRef	-72.27686	42.6715	58g
Central Hills	MA_DEP	B0819	2014016.R	RBP kicknet	2014-07-15	Ref	3_SubRef	-71.630154	42.046385	59c
Central Hills	MA_DEP	B0537	2004028	RBP kicknet	2004-08-26	Ref	3_SubRef	-72.014417	42.061813	59c
Central Hills	MA_DEP	B0644	2008057	RBP kicknet	2008-09-17	Ref	3_SubRef	-72.401061	42.148359	59b
Central Hills	MA_DEP	B0718	2011017	RBP kicknet	2011-07-07	Ref	3_SubRef	-72.257365	42.642539	58g
Central Hills	MA_DEP	B0447	2000020	RBP kicknet	2000-09-12	Ref	3_SubRef	-72.198629	42.651487	58g
Central Hills	MA_DEP	B0613	2007016	RBP kicknet	2007-07-19	Strs	7_HighStrs	-71.406671	42.215173	59c
Central Hills	MA_DEP	B0202	2001007	RBP kicknet	2001-07-05	Strs	7_HighStrs	-71.500851	42.258166	59h
Central Hills	MA_DEP	B0551	2005055	RBP kicknet	2005-07-26	Strs	7_HighStrs	-71.261478	42.487231	59h
Central Hills	MA_DEP	B0755	2009040.A	RBP kicknet	2009-08-06	Strs	7_HighStrs	-71.120552	42.485469	59d
Central Hills	MA_DEP	B0439	2000009	RBP kicknet	2000-07-19	Strs	7_HighStrs	-70.844894	42.661044	59f
Central Hills	MA_DEP	B0635	2008020	RBP kicknet	2008-07-21	Strs	7_HighStrs	-72.580391	42.247038	59a
Central Hills	MA_DEP	B0507	2003003	RBP kicknet	2003-07-22	Strs	7_HighStrs	-72.581951	42.245228	59a
Central Hills	MA_DEP	B0698	2010024	RBP kicknet	2010-07-22	Strs	7_HighStrs	-71.489025	42.304872	59h
Central Hills	MA_DEP	B0746	2009032	RBP kicknet	2009-07-20	Strs	7_HighStrs	-70.980182	42.220963	59c
Central Hills	MA_DEP	B0131	2009039	RBP kicknet	2009-08-06	Strs	7_HighStrs	-71.138747	42.446835	59d
Central Hills	MA_DEP	B0724	2011029	RBP kicknet	2011-07-14	Strs	7_HighStrs	-71.825842	42.239095	59h
Central Hills	MA_DEP	B0913	2015008.R	RBP kicknet	2015-07-13	Strs	7_HighStrs	-71.233834	42.304938	59d
Central Hills	MA_DEP	B0078	2008043.A	RBP kicknet	2008-08-26	Strs	7_HighStrs	-71.612278	42.556972	59h
Central Hills	MA_DEP	B0655	2008048	RBP kicknet	2008-09-04	Strs	7_HighStrs	-72.510199	42.16063	59a

Western Highlands IBI *calibration* samples (reference and stressed)

Table E3. Reference and stressed samples that were used to calibrate the Western Highlands (WH) IBI calibration. The samples were collected with the RBP kicknet method from 2000 onward.

Class	Entity	Unique_ID	BenSampID	Collection Method	Collection Date	Status	Disturbance category	Longitude	Latitude	Level IV ecoregion
WH	MA_DEP	B0216	2001026	RBP kicknet	2001-08-13	Ref	1_BestRef	-72.90985	42.08154	58e
WH	MA_DEP	B0035	2017011.Ar	RBP kicknet	2017-07-20	Ref	1_BestRef	-73.22432	42.62731	58a
WH	DRWA	DUBUQUE05	DUBUQUE05-11-1-1.R	RBP kicknet	2011-08-01	Ref	1_BestRef	-72.9184	42.5719	58c
WH	MA_DEP	B0789	2012021	RBP kicknet	2012-07-11	Ref	1_BestRef	-73.28035	42.24776	58d
WH	DRWA	CATAMOUNT08	CATAMOUNT08-11-1-1.R	RBP kicknet	2011-08-02	Ref	1_BestRef	-72.7408	42.6365	58f
WH	MA_DEP	B0881	2014021.R	RBP kicknet	2014-07-17	Ref	2_Ref	-72.81804	42.26076	58e
WH	MA_DEP	B0817	2014058.R	RBP kicknet	2014-08-11	Ref	2_Ref	-72.66912	42.71614	58f
WH	MA_DEP	B0787	2012016	RBP kicknet	2012-07-10	Ref	2_Ref	-73.017	42.0476	58d
WH	MA_DEP	B0498	2007025	RBP kicknet	2007-08-07	Ref	2_Ref	-73.14013	42.6232	58b
WH	MA_DEP	B0794	2012024	RBP kicknet	2012-07-12	Ref	2_Ref	-72.97368	42.16952	58d
WH	MA_DEP	B0700	2014054.R	RBP kicknet	2014-08-07	Ref	2_Ref	-72.9487	42.25418	58e
WH	MA_DEP	B0581	2006073	RBP kicknet	2006-08-16	Ref	2_Ref	-73.07692	42.19562	58d
WH	DRWA	CHRM03	CHRM03-08-1-1.R	RBP kicknet	2008-09-22	Ref	2_Ref	-72.94725	42.57756	58c
WH	MA_DEP	B0799	2012033.BR	RBP kicknet	2012-07-18	Ref	2_Ref	-72.97959	42.40227	58c
WH	MA_DEP	B0585	2006066	RBP kicknet	2006-08-15	Ref	2_Ref	-73.13228	42.04032	58d
WH	MA_DEP	B0215	2006065	RBP kicknet	2006-08-15	Ref	2_Ref	-72.96653	42.06405	58d
WH	MA_DEP	B0816	2012018	RBP kicknet	2012-07-11	Ref	2_Ref	-73.48299	42.10614	58a
WH	MA_DEP	B0806	2012041	RBP kicknet	2012-07-24	Ref	2_Ref	-73.01461	42.55161	58c
WH	MA_DEP	B0948	2017019.r	RBP kicknet	2017-07-27	Ref	2_Ref	-72.9408	42.03859	58d
WH	MA_DEP	B0788	2012056.R	RBP kicknet	2012-09-04	Ref	2_Ref	-72.93808	42.6392	58c
WH	DRWA	CDRM01	CDRM01-08-1-1.R	RBP kicknet	2008-09-18	Ref	2_Ref	-72.96105	42.64143	58c
WH	MA_DEP	B0458	2000034	RBP kicknet	2000-09-25	Ref	2_Ref	-72.93323	42.63497	58c
WH	MA_DEP	B0943	2015020.R	RBP kicknet	2015-07-20	Ref	2_Ref	-72.95334	42.34377	58e
WH	MA_DEP	B0560	2013054.R	RBP kicknet	2013-08-06	Ref	2_Ref	-72.66953	42.71644	58f

Table E3 continued…

Class	Entity	Unique_ID	BenSampID	Collection Method	Collection Date	Status	Disturbance category	Longitude	Latitude	Level IV ecoregion
WH	MA_DEP	B0947	2017016.r	RBP kicknet	2017-07-26	Ref	2_Ref	-72.9263	42.32292	58e
WH	MA_DEP	B0791	2012029	RBP kicknet	2012-07-17	Ref	2_Ref	-73.01147	42.30598	58e
WH	MA_DEP	B0800	2012052	RBP kicknet	2012-08-13	Strs	6_Strs	-72.61264	42.67328	58f
WH	MA_DEP	B0577	2006088	RBP kicknet	2006-09-06	Strs	6_Strs	-72.92779	42.39631	58e
WH	DRWA	CMBM01	CMBM01-06-1-1.R	RBP kicknet	2006-09-23	Strs	6_Strs	-72.7729	42.50907	58f
WH	MA_DEP	B0578	2006089	RBP kicknet	2006-09-06	Strs	6_Strs	-72.88977	42.46146	58c
WH	MA_DEP	B0882	2014050.AR	RBP kicknet	2014-08-05	Strs	6_Strs	-72.76956	42.42258	58e
WH	MA_DEP	B0489	2002022	RBP kicknet	2002-08-12	Strs	6_Strs	-73.10278	42.6146	58b
WH	MA_DEP	B0790	2012039	RBP kicknet	2012-07-24	Strs	6_Strs	-72.80105	42.51153	58f
WH	MA_DEP	B0036	2007030	RBP kicknet	2007-08-08	Strs	6_Strs	-73.25305	42.65197	58b
WH	MA_DEP	B0492	2002024	RBP kicknet	2002-08-13	Strs	6_Strs	-73.2074	42.72554	58b
WH	MA_DEP	B0641	2008025	RBP kicknet	2008-07-22	Strs	6_Strs	-72.73391	42.2922	58e
WH	MA_DEP	B0501	2002039	RBP kicknet	2002-09-10	Strs	6_Strs	-73.137	42.48612	58b
WH	DRWA	CLBM01	CLBM01-08-1-1.R	RBP kicknet	2008-09-23	Strs	6_Strs	-72.76904	42.61247	58f
WH	MA_DEP	B0880	2014022.R	RBP kicknet	2014-07-17	Strs	6_Strs	-72.7294	42.28135	58e
WH	MA_DEP	B0633	2007043	RBP kicknet	2007-08-29	Strs	6_Strs	-73.14837	42.48448	58b
WH	MA_DEP	B0622	2007033	RBP kicknet	2007-08-09	Strs	6_Strs	-73.32006	42.54163	58a
WH	MA_DEP	B0793	2012043	RBP kicknet	2012-07-25	Strs	6_Strs	-73.31275	42.54503	58a
WH	MA_DEP	B0022	2002041	RBP kicknet	2002-09-10	Strs	6_Strs	-73.27121	42.44013	58b
WH	MA_DEP	B0502	2002040	RBP kicknet	2002-09-10	Strs	6_Strs	-73.12842	42.45186	58c
WH	MA_DEP	B0795	2012032	RBP kicknet	2012-07-18	Strs	6_Strs	-73.14121	42.47392	58b
WH	MA_DEP	B0632	2007041	RBP kicknet	2007-08-29	Strs	6_Strs	-73.12976	42.45725	58c
WH	MA_DEP	B0782	2012050.A	RBP kicknet	2012-08-08	Strs	6_Strs	-72.73353	42.67417	58f
WH	MA_DEP	B0032	2002033	RBP kicknet	2002-08-14	Strs	6_Strs	-73.23126	42.67649	58b

Table E3 continued…

Class	Entity	Unique_ID	BenSampID	Collection Method	Collection Date	Status	Disturbance category	Longitude	Latitude	Level IV ecoregion
WH	MA_DEP	B0021	2007040	RBP kicknet	2007-08-29	Strs	6_Strs	-73.26013	42.44115	58b
WH	MA_DEP	B0034	2007032	RBP kicknet	2007-08-09	Strs	6_Strs	-73.19805	42.70902	58b
WH	MA_DEP	B0802	2012045	RBP kicknet	2012-07-25	Strs	6_Strs	-73.2002	42.7029	58b
WH	MA_DEP	B0017	2002038	RBP kicknet	2002-09-09	Strs	6_Strs	-73.36368	42.22686	58b
WH	MA_DEP	B0497	2002034	RBP kicknet	2002-09-09	Strs	6_Strs	-73.37813	42.17856	58b
WH	MA_DEP	B0039	2007023	RBP kicknet	2007-08-07	Strs	6_Strs	-73.10756	42.64375	58b
WH	MA_DEP	B0040	2002028	RBP kicknet	2002-08-13	Strs	6_Strs	-73.10845	42.64765	58b
WH	MA_DEP	B0798	2012046	RBP kicknet	2012-07-25	Strs	6_Strs	-73.10371	42.66954	58b
WH	DRWA	NORM01	NORM01-07-1-1.R	RBP kicknet	2007-09-27	Strs	6_Strs	-72.73609	42.62784	58f
WH	MA_DEP	B0504	2002045	RBP kicknet	2002-09-11	Strs	6_Strs	-73.24524	42.34415	58b
WH	MA_DEP	B0041	2007029.1	RBP kicknet	2007-08-08	Strs	6_Strs	-73.20922	42.72932	58b
WH	MA_DEP	B0626	2007027	RBP kicknet	2007-08-08	Strs	6_Strs	-73.21848	42.73468	58b
WH	MA_DEP	B0505	2007045	RBP kicknet	2007-08-30	Strs	6_Strs	-73.24088	42.28335	58b
WH	DRWA	SORM08	SORM08-06-1-1.R	RBP kicknet	2006-09-10	Strs	7_HighStrs	-72.79185	42.53296	58f
WH	DRWA	PHBM01	PHBM01-06-1-1.R	RBP kicknet	2006-09-24	Strs	7_HighStrs	-72.69787	42.50811	58f
WH	MA_DEP	B0809	2012030	RBP kicknet	2012-07-18	Strs	7_HighStrs	-73.35414	42.35721	58b
WH	DRWA	SORM06	SORM06-06-1-1.R	RBP kicknet	2006-09-23	Strs	7_HighStrs	-72.75991	42.50973	58f
WH	MA_DEP	B0490	2002021	RBP kicknet	2002-08-12	Strs	7_HighStrs	-73.11367	42.591	58b
WH	MA_DEP	B0624	2007034	RBP kicknet	2007-08-28	Strs	7_HighStrs	-73.33421	42.30161	58b
WH	MA_DEP	B0233	2000045	RBP kicknet	2000-09-27	Strs	7_HighStrs	-72.69393	42.5419	58f

Western Highlands IBI *verification* samples (reference and stressed)

Table E4. Reference and stressed samples that were used to verify the Western Highlands IBI.

Class	Entity	Unique_ID	BenSampID	Collection Method	Collection Date	Status	Disturbance category	Longitude	Latitude	Level IV ecoregion
WH	MA_DEP	B0623	2007031	RBP kicknet	2007-08-09	Ref	1_BestRef	-73.22661	42.71702	58b
WH	MA_DEP	B0820	2015047.R	RBP kicknet	2015-07-30	Ref	1_BestRef	-72.95888	42.70347	58c
WH	MA_DEP	B0815	2012044	RBP kicknet	2012-07-25	Ref	2_Ref	-73.20167	42.6579	58b
WH	DRWA	CDRM03	CDRM03-08-1-1.R	RBP kicknet	2008-09-18	Ref	2_Ref	-73.03029	42.66645	58c
WH	DRWA	TNBM01	TNBM01-08-1-1.R	RBP kicknet	2008-09-18	Ref	2_Ref	-72.99827	42.62974	58c
WH	MA_DEP	B0949	2017020.r	RBP kicknet	2017-07-27	Ref	2_Ref	-72.9175	42.04234	58e
WH	MA_DEP	B0803	2012022	RBP kicknet	2012-07-11	Ref	2_Ref	-73.27398	42.27211	58b
WH	MA_DEP	B0740	2013049.R	RBP kicknet	2013-08-05	Ref	2_Ref	-72.94935	42.16783	58e
WH	MA_DEP	B0818	2014053.R	RBP kicknet	2014-08-07	Ref	2_Ref	-73.02775	42.32001	58c
WH	DRWA	SDBM01	SDBM01-07-1-1.R	RBP kicknet	2007-09-22	Ref	2_Ref	-72.78436	42.70371	58c
WH	DRWA	CDRM02	CDRM02-08-1-1.R	RBP kicknet	2008-09-18	Ref	2_Ref	-72.93443	42.63635	58c
WH	MA_DEP	B0037	2007026	RBP kicknet	2007-08-07	Ref	2_Ref	-73.14844	42.59686	58b
WH	MA_DEP	B0783	2012031	RBP kicknet	2012-07-18	Strs	6_Strs	-73.30162	42.45191	58b
WH	MA_DEP	B0812	2012053	RBP kicknet	2012-08-13	Strs	6_Strs	-72.64486	42.62827	58f
WH	MA_DEP	B0630	2007039	RBP kicknet	2007-08-29	Strs	6_Strs	-73.22528	42.42448	58b
WH	MA_DEP	B0506	2002046	RBP kicknet	2002-09-11	Strs	6_Strs	-73.22719	42.2943	58b
WH	DRWA	CLBM02	CLBM02-08-1-1.R	RBP kicknet	2008-09-23	Strs	6_Strs	-72.80022	42.57976	58f
WH	DRWA	WBNM01	WBNM01-07-1-1.R	RBP kicknet	2007-09-22	Strs	6_Strs	-72.72384	42.66643	58f
WH	MA_DEP	B0807	2012028	RBP kicknet	2012-07-17	Strs	6_Strs	-72.87868	42.26069	58e
WH	DRWA	EBNM01	EBNM01-07-1-1.R	RBP kicknet	2007-09-23	Strs	6_Strs	-72.71802	42.66989	58f
WH	MA_DEP	B0503	2007042	RBP kicknet	2007-08-29	Strs	6_Strs	-73.19547	42.46955	58b
WH	MA_DEP	B0500	2002036	RBP kicknet	2002-09-09	Strs	6_Strs	-73.31183	42.04741	58b
WH	MA_DEP	B0042	2002026	RBP kicknet	2002-08-13	Strs	6_Strs	-73.21566	42.7302	58b
WH	DRWA	SORM07	SORM07-06-1-1.R	RBP kicknet	2006-09-23	Strs	7_HighStrs	-72.7808	42.5218	58f
WH	DRWA	SORM03	SORM03-06-1-1.R	RBP kicknet	2006-09-22	Strs	7_HighStrs	-72.69678	42.51748	58f

Table E4 continued…

Class	Entity	Unique_ID	BenSampID	Collection Method	Collection Date	Status	Disturbance category	Longitude	Latitude	Level IV ecoregion
WH	MA_DEP	B0627	2007044.A	RBP kicknet	2007-08-29	Strs	7_HighStrs	-73.25651	4.46968	58b

Appendix F

Selection of an IBI scoring scheme

1 Background

A number of different scoring schemes can be used when scoring metrics for IBIs, each with the objective of assigning relative values to reflect the degree to which each metric departs from expected conditions. During development of the MA RBP kick net IBIs, we initially used a 0 to 100-point, non-normalized scoring scheme for both 'increaser' and 'decreaser' metrics based on the 5th and 95th percentiles. Using the 5th and 95th percentiles instead of the maximum and minimum reduces the influence of outliers (Barbour et al. 1999). The original scoring equations were as follows:

Metrics that ***decrease with increasing stress*** (e.g., number of intolerant taxa):

Metric Score = 100*(Metric Value – 5th Percentile) / (95th Percentile-5th Percentile)

Metrics that ***increase with stress*** (like number of tolerant taxa):

Metric Score = 100*(95th Percentile - Metric Value) / (95th Percentile-5th Percentile)

The original scoring formulas for each component metric in the Western Highlands (WH) and Central Hills (CH) IBIs are shown in Tables 1 & 2, respectively.

After IBI development, we examined the distributions of IBI scores across the two regions and found that differences in the reference distributions were affecting metric scoring in a way that could cause confusion when IBI results are being interpreted. As an example, as shown in Figure 1, the median IBI score for reference calibration samples was ~70 in the CH dataset vs. ~55 in the WH. Someone might mistakenly interpret this to mean that reference sites in the CH are in better biological condition than reference sites in the WH, but this is not the case (it's actually the opposite). The IBI scales for the two stream classes are independent. An IBI score in one site class does not imply the same biological condition in the other.

We took a closer look at what was causing the differences and found the primary driver to be the 'decreaser' metrics, which comprise most of the IBI input metrics (Tables 1 & 2). When scoring the 'decreaser' metrics with the original formulas, the 5th percentile was subtracted from the metric value. If you compare the 5th percentile value for the taxa richness metric (which is used in both the CH and WH IBIs), the value is 21 in the WH dataset (Table 1) versus 11 in the CH dataset (Table 2). This difference caused the metric scores in the WH to be lower. For example, if you have a taxa richness value of 30 and you run it through the metric scoring formulas for both classes, the metric score for the CH sample will be 79, compared to a metric score of 50 for the WH sample.

We tried several alternate scoring schemes to see if we could eliminate or lessen this issue. The goal was to find an IBI scoring scheme that retained the performance characteristics of the original IBIs (especially discrimination efficiency (DE)), had similar reference distribution statistics among site classes (especially at the 25th and lower percentiles), was easier to communicate and, if numeric biocriteria were someday established, could potentially allow for the use of one set of impairment thresholds across both regions. In the end, we selected a scoring scheme that used the minimum possible value instead of the 5th percentile when scoring 'decreaser' metrics (referred to as the 'minimum floor' (MinFloor) option). The formula for the 'increaser' metrics was left unchanged. In this appendix, we describe how we came to select the MinFloor scoring formula for 'decreaser' metrics.

Central Hills		
Dataset	**Number of sites**	
	Ref	**Strs**
cal	30	30
verif	12	14
verif pre-2000	3	19

Western Highlands		
Dataset	**Number of sites**	
	Ref	**Strs**
cal	26	42
verif	12	14
verif pre-2000	3	14

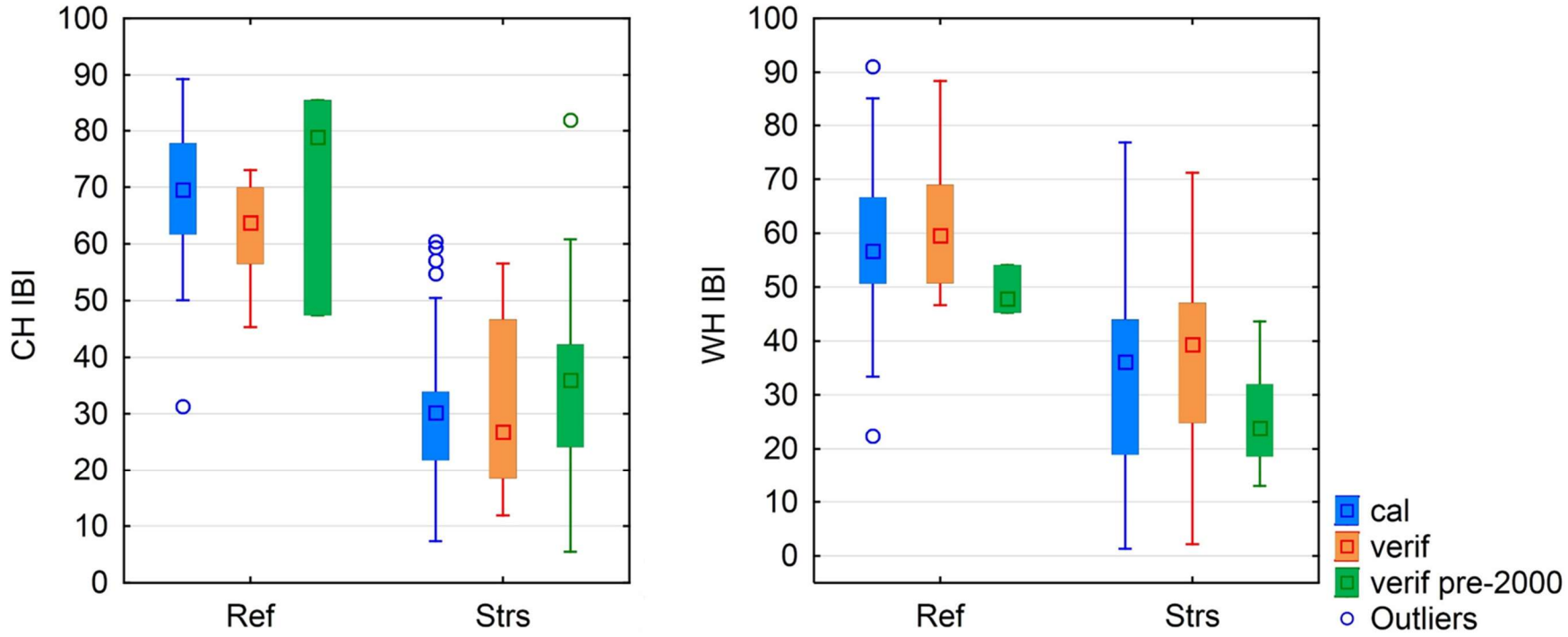


Figure 1. Box plots showing the distribution of Central Hills (CH) IBI (left) and Western Highland (WH) IBI (right) scores in the reference and stressed calibration and verification datasets (2000-2017) and the pre-2000 verification kick net datasets.

Table 1. Metrics in the Western Highlands index, with original and final (MinFloor) scoring formulas, Discrimination Efficiency (DE) scores and trend. Formulas that differ between the two scoring schemes are in red text. The MinFloor uses the minimum possible value (0) instead of the 5th percentile.

Metric abbrev	Metric	Category	Trend	DE	5th	95th	95th-5th	Scoring	Scoring formula
nt_total	Number of taxa - total	RICH	Dec.	52.4	21	38.8	17.8	original	100*(metric-21)/ 17.8
								MinFloor	100*(metric-0)/ 38.8
pi_Pleco	Percent individuals - Order Plecoptera	COMP	Dec.	66.7	0	18.3	18.3	original	100*(metric-0)/ 18.3
								MinFloor	same formula (5th = 0)
pi_ffg_filt	Percent individuals - Functional Feeding Group (FFG) - collector-filterer (CF)	FFG	Inc.	50	9.76	50.5	40.7	original	100*(50.5-metric)/ 40.7
								MinFloor	same formula (increaser)
pi_ffg_shred	Percent individuals - FFG - shredder (SH)	FFG	Dec.	61.9	1.17	23	21.8	original	100*(metric-1.17)/ 21.8
								MinFloor	100*(metric-0)/ 23
pi_tv_intol	Percent individuals - tolerance value - intolerant ≤ 3	TOLER	Dec.	59.5	6.09	51.5	45.4	original	100*(metric-6.09)/ 45.4
								MinFloor	100*(metric-0)/ 51.5
x_Becks	Becks Biotic Index*	TOLER	Dec.	57.1	12	36.8	24.8	original	100*(metric-12)/ 24.8
								MinFloor	100*(metric-0)/ 36.8

Table 2. Metrics in the Central Hills index, with original and final (MinFloor) scoring formulas, Discrimination Efficiency (DE) scores and trend. Formulas that differ between the two scoring schemes are in red text. The MinFloor uses the minimum possible value (0) instead of the 5th percentile.

Metric abbrev	Metric	Category	Trend	DE	5th	95th	95th-5th	Scoring	Scoring formula
nt_total	Number of taxa - total	RICH	Dec.	66.7	11	34.9	23.9	original	100*(metric-11)/ 23.9
								MinFloor	100*(metric-0)/ 34.9
pt_EPT	Percent taxa - Orders Ephemeroptera, Plecoptera & Trichoptera (EPT)	RICH	Dec.	76.7	10.6	54.5	43.9	original	100*(metric-10.6)/ 43.9
								MinFloor	100*(metric-0)/ 54.5
pi_Ephem NoCaeBae	Percent individuals - Order Ephemeroptera, excluding Families Caenidae and Baetidae	COMP	Dec.	66.7	0	13.9	13.9	original	100*(metric-0)/ 13.9
								MinFloor	same formula (5th = 0)
pi_ffg_filt	Percent individuals - Functional Feeding Group (FFG) - collector-filterer (CF)	FFG	Inc.	76.7	13	79.9	66.9	original	100*(79.9-metric)/ 66.9
								MinFloor	same formula (increaser)
pt_ffg_pred	Percent taxa - Functional Feeding Group (FFG) - predator (PR)	FFG	Dec.	90	0	28.5	28.5	original	100*(metric-0)/ 28.5
								MinFloor	same formula (5th = 0)
pt_tv_intol	Percent taxa - tolerance value - intolerant ≤ 3	TOLER	Dec.	100	0	39.1	39.1	original	100*(metric-0)/ 39.1
								MinFloor	same formula (5th = 0)

2 Methods

We tried eight alternate scoring methods, which are described in Table 3. These included the MinFloor option, which utilizes the minimum possible value instead of the 5th percentile when scoring 'decreaser' metrics, as well as changing from a 100-point scale to a normalized scale based on the mean and standard deviation of the index calibration reference distribution. We tested the alternate scoring methods by:

- Calculating DE and Z-scores and evaluating whether the alternates performed worse, better or the same as the original IBI scores.
- Generating box plots to evaluate IBI score distributions and discriminatory ability across disturbance categories (reference, other, and stressed).
- Calculating reference distribution statistics for the alternative index formulations, including mean minus standard deviation statistics (X-SD) to illustrate the relationship between percentiles and standard deviations.
- Performing a t-test on IBI reference distributions among site classes.
- Generating Cumulative Distribution Function (CDF) plots to illustrate comparative distributions of reference, other, and stressed sites among classes.

We considered doing regressions of one IBI upon the another but decided against it due to a lack of predictive rationale (there is no cause for one index to be related to the other in independent samples).

Table 3. Descriptions of the IBI scoring alternatives.

Short Name	Description
Indx_Orig	Uses the 5th and 95th percentiles to establish the scoring range for all metrics. Dataset: post-1999, cal + verif, all disturbance categories (ref+other+strs).
Indx_MinFloor	Uses the minimum possible value (zero) and the 95th percentile for all decreasing metrics. Metrics that increase with stress (e.g., % filterers) use the 5th and 95th percentiles. Dataset: post-1999, cal + verif, all disturbance categories (ref+other+strs).
Indx_ALL595_cal	Uses the 5th and 95th percentiles of all data (including pre-2000) to establish the scoring range for all metrics. Dataset: pre- and post-2000, cal + verif, all disturbance categories (ref+other+strs). Used calibration data (post-1999) to evaluate comparability of index characteristics to other schemes.
Indx_ALL595_all	Uses the 5th and 95th percentiles of all data (including pre-2000) to establish the scoring range for all metrics. Dataset: pre- and post-2000, cal + verif, all disturbance categories (ref+other+strs). Evaluated the index using all data (pre- and post-2000) to understand overall sensitivity.
Indx_OrigMSD	Indx_Orig standardized to the mean and standard deviation of the index calibration reference distribution (mean of reference now = 0). Standardization dataset: post-1999, cal only, ref only.
IndxMF_MSD	Indx_MinFloor standardized to the mean and standard deviation of the index calibration reference distribution. Standardization dataset: post-1999, cal only, ref only.
Indx_AllYrs_cal.MSD	Indx_ALL595_cal standardized to the mean and standard deviation of the index reference calibration data. Standardization dataset: all years, cal only, ref only. Evaluates index characteristics using only calibration (post-1999) for comparability to other schemes using calibration data.
Indx_AllYrs_all.MSD	Indx_ALL595_all standardized to the mean and standard deviation of the reference distribution of all data. Standardization dataset: all years, cal only, ref only. Evaluates index characteristics using all data (calibration, pre- and post-2000) to understand overall sensitivity.

3 Results

3.1 Discrimination efficiency (DE) and Z-scores

Most index formulations resulted in identical or similar performance characteristics in the Central Hills (Table 4). In comparison to the original index in calibration samples, the indices evaluated using all samples performed slightly poorer. The highest Z-score was evident in the index alternatives that use the minimum floor for scoring metrics.

In the Western Highlands, the DE of all alternatives were at least as sensitive as the original scoring alternative. The most sensitive index alternatives when measured by DE included those that use the minimum floor for scoring metrics. Those also had the worst Z-scores. The indices evaluated using all samples had the best Z-score in the Western Highlands. The index performances after normalizing to the means and standard deviations are the same as performances of the un-normalized alternatives.

Table 4. Discrimination efficiency (DE) and Z-scores for the eight index alternatives in two site classes.

	Central Hills		Western Highlands	
Index Alternatives	**DE**	**Z**	**DE**	**Z**
Indx_Orig	100	2.93	85.7	1.47
Indx_MinFloor	100	3.02	88.1	1.40
Indx_ALL595_cal	100	2.94	85.7	1.48
Indx_ALL595_all	98	2.72	87.5	1.57
Indx_OrigMSD	100	2.93	85.7	1.47
IndxMF_MSD	100	3.02	88.1	1.40
Indx_AllYrs_cal.M.SD	100	2.94	85.7	1.48
Indx_AllYrs_all.M.SD	98	2.72	87.1	1.61

3.2 Box plots

The index score distributions by disturbance category and site class are illustrated in Figures 2 through 9. All the index alternatives show a distinction between reference and stressed values. The index alternatives that are evaluated with all samples (pre- and post-2000) show an evenly stepped decrease in index scores from Reference to Other to Stressed disturbance categories (based on visual inspection of the box plots). In alternatives with only calibration data, the Other category in the Western Highlands is skewed towards the Reference distribution. It seems that Other earlier samples had generally lower biological condition than the calibration samples alone. The effect of non-calibration samples was not as obvious in the Reference and Stressed categories or in the Central Hills.

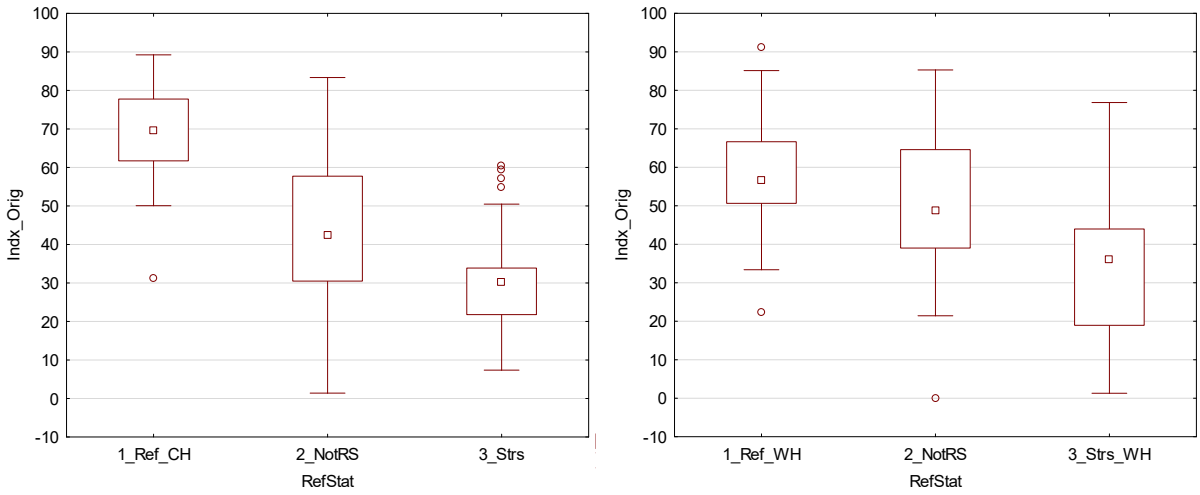


Figure 2. Illustration of Indx_Orig index value distributions in reference (Ref_CH), other (NotRS), and stressed (Strs) sites for the Central Hills (left) and the Western Highlands (right). This includes calibration data only (post-1999).

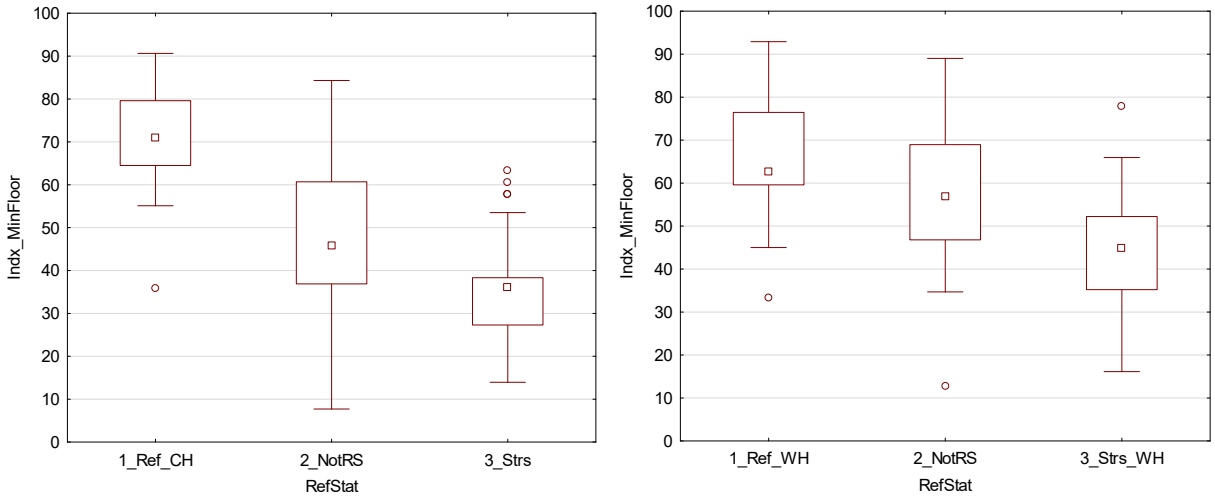


Figure 3. Illustration of Indx_MinFloor index value distributions in reference (Ref_CH), other (NotRS), and stressed (Strs) sites for the Central Hills (left) and the Western Highlands (right). This includes calibration data only (post-1999).

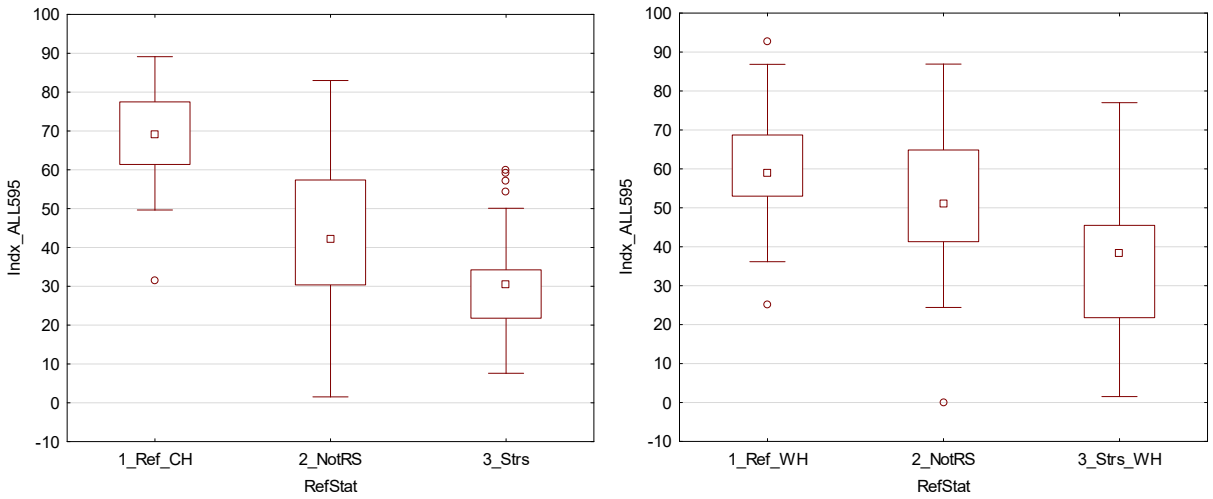


Figure 4. Illustration of Indx_ALL595_cal index value distributions in reference (Ref_CH), other (NotRS), and stressed (Strs) sites for the Central Hills (left) and the Western Highlands (right). This illustrates calibration data only (calibration, post-1999).

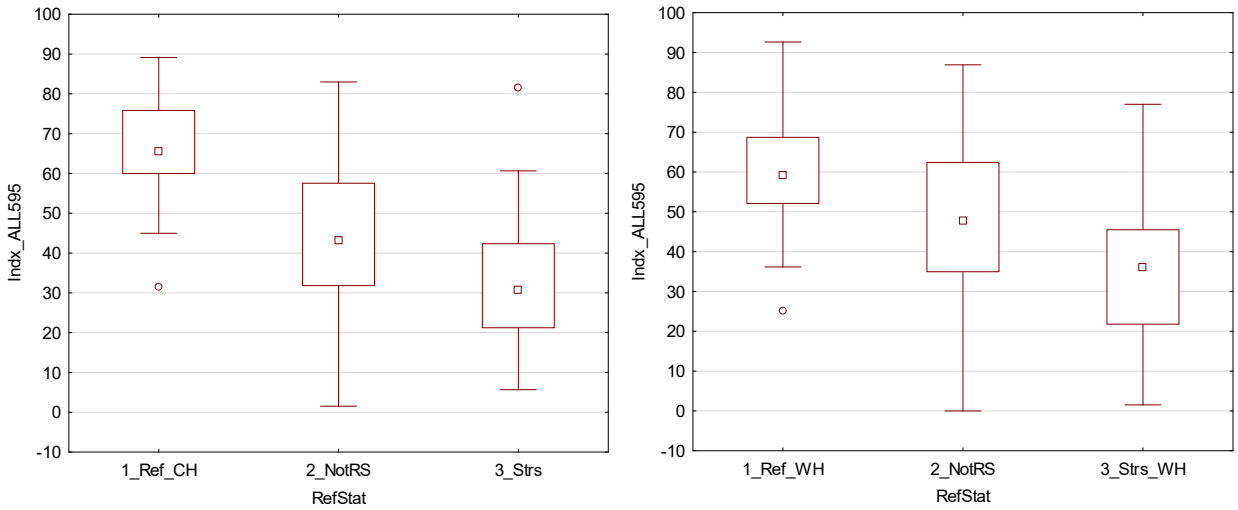


Figure 5. Illustration of Indx_ALL595_all index value distributions in reference (Ref_CH), other (NotRS), and stressed (Strs) sites for the Central Hills (left) and the Western Highlands (right). This illustrates all data (calibration, pre-, and post-2000).

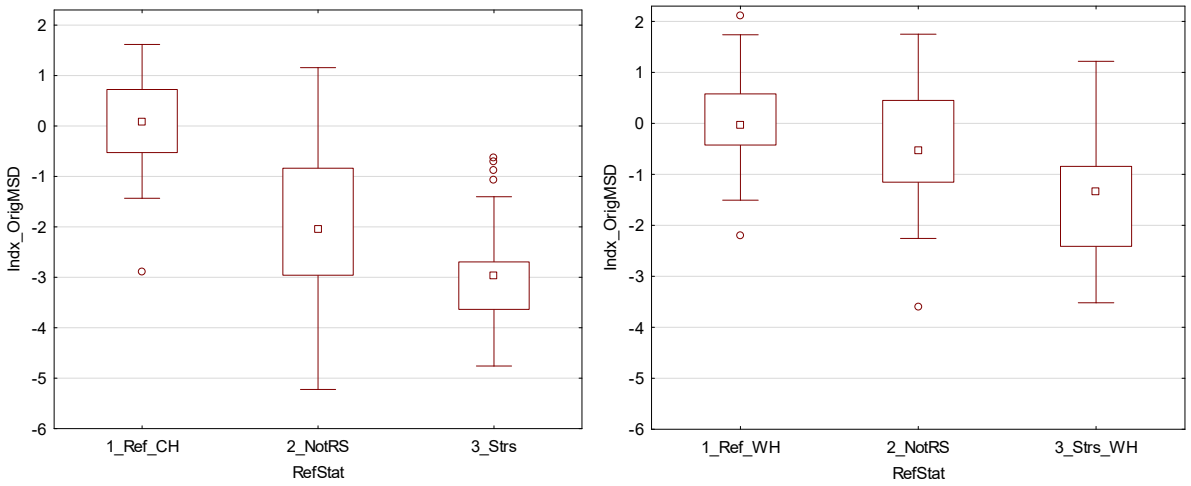


Figure 6. Illustration of Indx_OrigMSD index value distributions in reference (Ref_CH), other (NotRS), and stressed (Strs) sites for the Central Hills (left) and the Western Highlands (right). This includes calibration data only (post-1999).

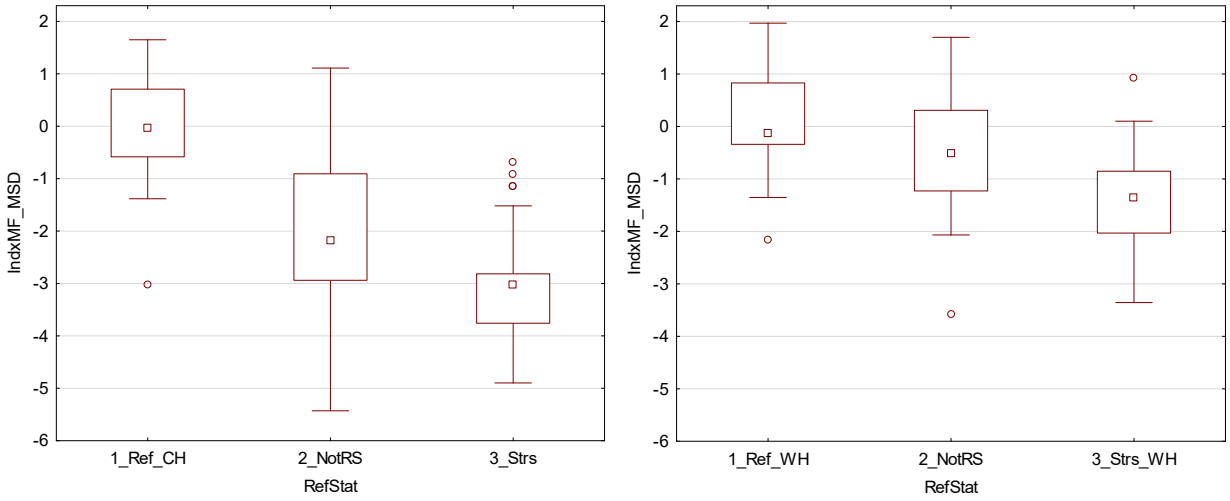


Figure 7. Illustration of IndxMF_MSD index value distributions in reference (Ref_CH), other (NotRS), and stressed (Strs) for the Central Hills (left) and the Western Highlands (right). This includes calibration data only (post-1999).

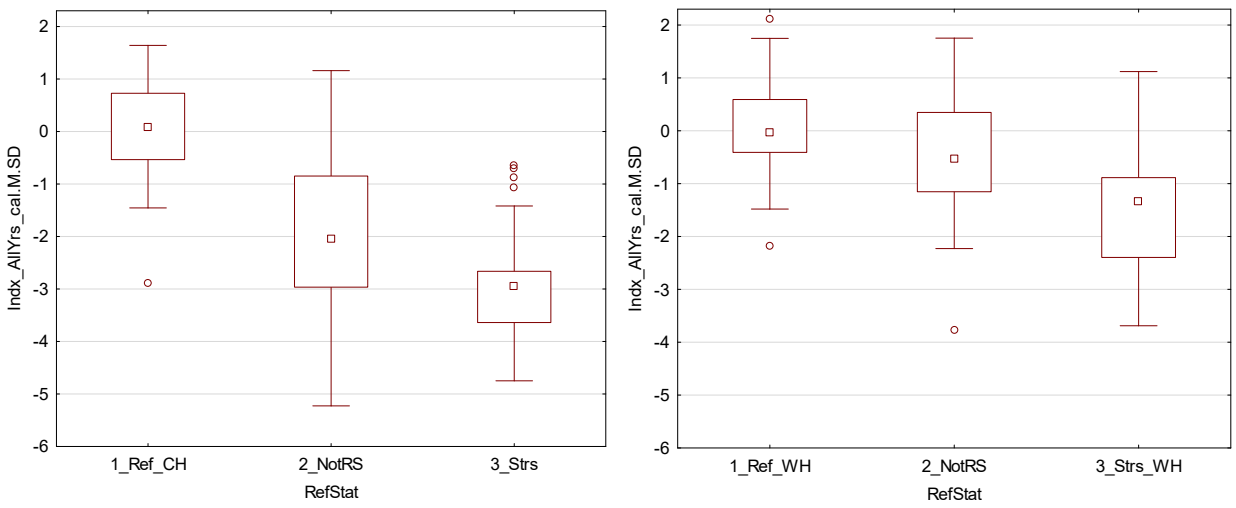


Figure 8. Illustration of Indx_AllYrs_cal.MSD index value distributions in reference (Ref_CH), other (NotRS), and stressed (Strs) sites for the Central Hills (left) and the Western Highlands (right). This includes calibration data only (post-1999).

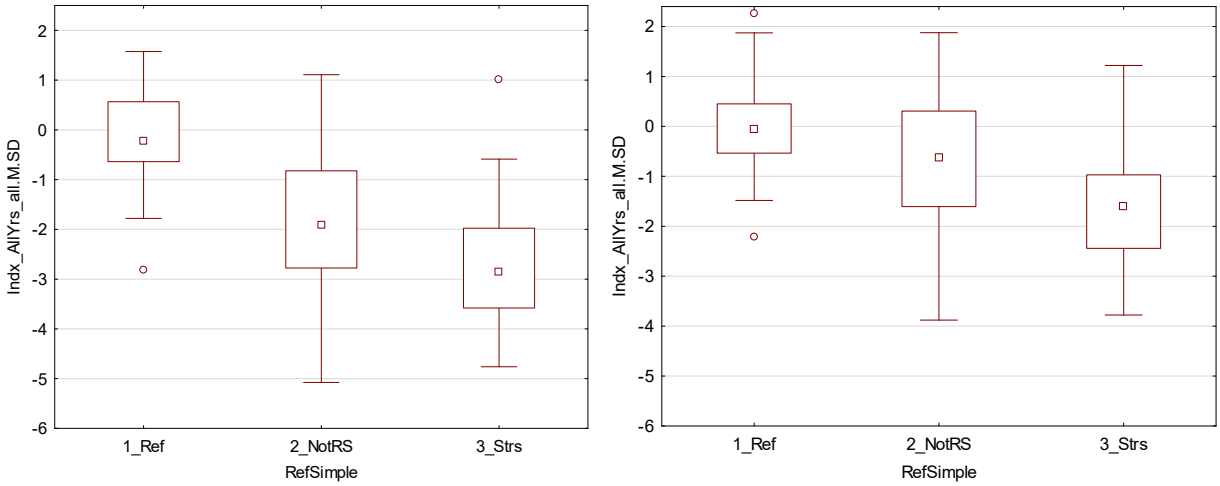


Figure 9. Illustration of Indx_AllYrs_all.MSD index value distributions in reference (Ref_CH), other (NotRS), and stressed (Strs) sites for the Central Hills (left) and the Western Highlands (right). This illustrates all data (calibration, pre-, and post-2000).

3.3 Reference distribution statistics

The distributions of index values were compared among site classes, with special attention to the lower end of the reference distribution because this is typically the part of the distribution that informs biological condition thresholds. Although the intention of some scoring alternatives was to standardize the index values among site classes, it seems that the reference distribution statistics in the Central Hills are consistently different than the same statistics in the Western Highlands (Table 5). For index values on the 100-point scale, the Central Hills percentiles were consistently higher than the Western Highlands percentiles by about 10-15 points. Conversely, the percentiles for mean-normalized indices were consistently lower in the Central Hills compared to the Western Highlands, though only for the lower percentiles and not for the medians, which are near zero by design for both classes. On that scale, which is generally between -5 and 2 standard deviations, the differences between percentiles among site classes was about 0.5 and 0.20 standard deviations.

The mean minus standard deviations for the metrics on a 100-point scale indicate that one SD from the reference mean is similar to the reference 10th percentile in the Central Hills. In the Western Highlands, the 10th percentile is somewhat more than one SD from the mean. These differences will be important to note if thresholds are based on standardizing the SDs. It appears that the SDs are somewhat smaller in the Central Hills, indicating a more consistent measurement of the reference condition among sites. This does not imply that the deviation from the mean is more or less ecologically important in the Central Hills compared to the Western Highlands.

Table 5. Reference distribution statistics for the alternative index formulations in the two site classes. Index values are shown for the median and percentile statistics (based on the calibration and verification samples collected from 2000 onward). The mean minus standard deviation statistics (X-SD) are shown to illustrate the relationship between percentiles and standard deviations. They are the percentages of reference sites that are < -0.5, -1.0, and -1.5 SD.

Central Hills	Median	25th	10th	5th	X-0.5SD	X-1SD	X-1.5SD
Indx_Orig	69.6	61.7	54.6	50.2	36.7%	13.3%	3.3%
Indx_MinFloor	71	64.7	58.6	56.2	30.0%	13.3%	3.3%
Indx_ALL595_cal	68.2	60.5	53.8	50.5	30.0%	13.3%	3.3%
Indx_ALL595_all	68.3	60.2	51.3	49.0	33.3%	15.2%	6.1%
Indx_OrigMSD	0.09	-0.53	-1.08	-1.42	36.7%	13.3%	3.3%
IndxMF_MSD	-0.03	-0.56	-1.08	-1.38	30.0%	13.3%	3.3%
Indx_AllYrs_cal.MSD	0.04	-0.57	-1.10	-1.36	30.0%	13.3%	3.3%
Indx_AllYrs_all.MSD	0.03	-0.59	-1.27	-1.45	33.3%	15.2%	6.1%
Western Highlands	**Median**	**25th**	**10th**	**5th**	**X-0.5SD**	**X-1SD**	**X-1.5SD**
Indx_Orig	56.7	50.7	36.7	34.1	23.1%	15.4%	7.7%
Indx_MinFloor	62.5	59.7	46.2	45.2	23.1%	19.2%	3.9%
Indx_ALL595_cal	58.7	53.0	39.2	36.8	23.1%	19.2%	3.9%
Indx_ALL595_all	59.0	52.1	43.7	38.9	27.6%	17.2%	6.9%
Indx_OrigMSD	-0.05	-0.42	-1.30	-1.46	23.1%	15.4%	7.7%
IndxMF_MSD	-0.14	-0.33	-1.27	-1.34	23.1%	19.2%	3.9%
Indx_AllYrs_cal.MSD	-0.02	-0.46	-1.00	-1.31	23.1%	19.2%	3.9%
Indx_AllYrs_all.MSD	-0.07	-0.55	-1.14	-1.47	27.6%	17.2%	6.9%

3.4 t-test of index reference distributions

A t-test of index reference distributions among site classes indicated that the most similar mean index values (highest p-values) were for the normalized index scores evaluated in calibration data (Table 6). Of course, these are standardized on the reference mean, so the means must be identical (= zero). Of the non-normalized index scores, the Indx_MinFloor means were the most similar among site classes and were not statistically different (p>0.05).

Table 6. Descriptive statistics for the alternative indices and t-test results among site classes. Entries in red text are statistically different (p<0.05).

	Western Highlands			Central Hills			t-test	
Index Alternative	Mean	Valid N	StdDev	Mean	Valid N	StdDev	t-value	p
Indx_Orig	57.4	26	15.9	68.5	30	12.8	-2.88	0.01
Indx_MinFloor	64.5	26	14.4	71.3	30	11.7	-1.95	0.06
Indx_ALL595_cal	59.4	26	15.7	68.2	30	12.8	-2.31	0.02
Indx_ALL595_all	60.0	41	14.4	66.6	45	12.4	-2.28	0.02
Indx_OrigMSD	0.0	26	1.0	0.0	30	1.0	0.00	1.00
IndxMF_MSD	0.0	26	1.0	0.0	30	1.0	0.00	1.00
Indx_AllYrs_cal.MSD	0.0	26	1.0	0.0	30	1.0	0.00	1.00
Indx_AllYrs_all.MSD	0.1	41	1.0	-0.1	45	0.9	1.13	0.26

3.5 Cumulative Distribution Function (CDF) plots

CDF plots can be used to illustrate comparative distributions of reference, other, and stressed sites among site classes. As an example, Figure 10 shows CDF plots for the IndxMF_MSD index. The reference distribution is similar for most of the range of values, including the maximum and minimum. The 'other' and 'stressed' distributions illustrate that the Central Hills values are generally lower than the Western Highlands values.

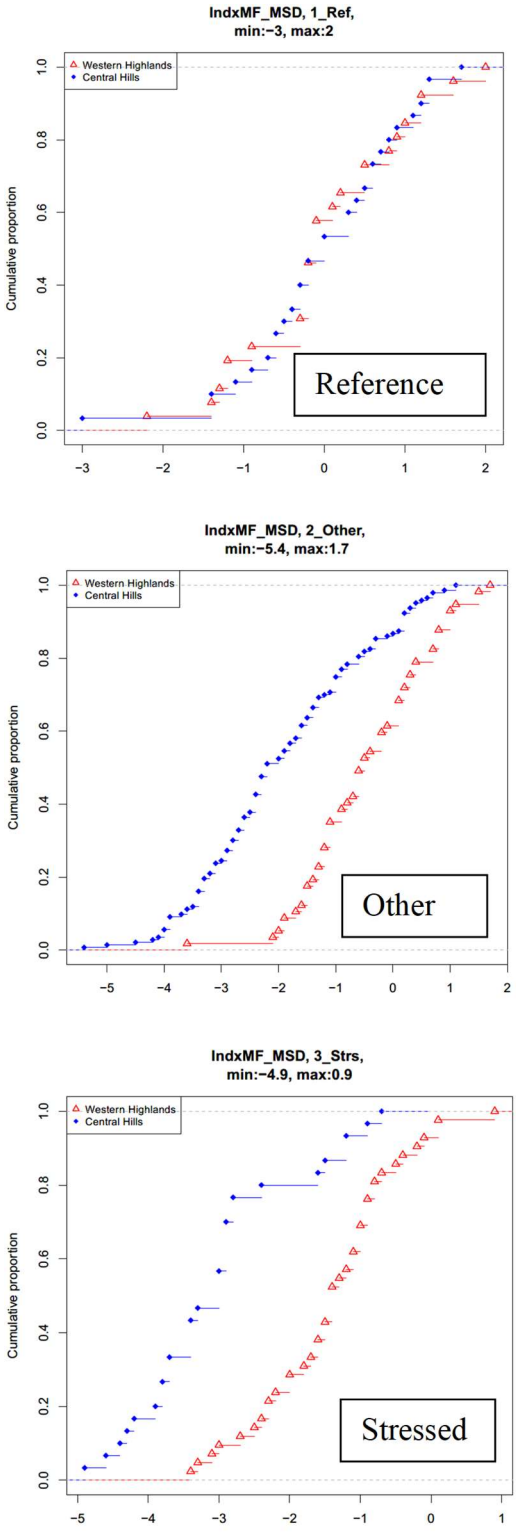


Figure 10. Cumulative Distribution Function (CDF) plots of reference (top), other (middle), and stressed (bottom) disturbance categories among site classes for the IndxMF_MSD index. Note differences in the x-axis for each disturbance category.

4 Discussion

After evaluating the eight alternate scoring methods, we concluded that the best 100-point scale alternative for 'decreaser' metrics is the Indx_MinFloor. Its DE is slightly better than the other options (Table 4) and it makes the distributions of reference site IBI values more similar across the two regions than the original scoring scheme (5th and 95th percentiles for 'decreaser' metrics). This type of scoring scheme (100-point scoring scale ranging from the minimum to the 95th percentile of all data) is also used in Connecticut (Gerritsen and Jessup 2007) and Rhode Island (Jessup et al. 2012) and has been shown to perform well in an independent study by Blocksom (2003).

The normalized version of Indx_MinFloor (IndxMF_MSD, which is standardized to the mean and standard deviation of the index calibration reference distribution performs equally well as the Indx_MinFloor in respect to DE and Z-scores, and brings the reference distribution metrics between the two classes closer together than the 100-point scale (but still not equal; Table 5). For the IndxMF_MSD index, 30% of reference sites had values < the mean minus 0.5 SD in the Central Hills. The same statistic in the Western Highlands was about 23%. To simplify and generalize for comparisons among site classes, the mean minus 0.5 SD was near the 26th percentile of reference, the mean minus 1.0 SD was near the 16th percentile of reference, and the mean minus 1.5 SD was near the 4th percentile of reference. In regard to stressed site index values, the IndxMF_MSD scores in the Central Hills were further departed from reference site values than in the Western Highlands. This held true for the other scoring schemes normalized to the mean of reference as well, as shown in the box plots (Figures 6 through 9) and Z-scores (Table 4). The greater departure is partly due to the smaller SD in the Central Hills reference sites (SD ~ 12) compared to the Western Highlands (SD ~ 15). Another explanation of the greater departure might be more diverse disturbance conditions in the Central Hills. In the Western Highlands, the stressors might not be intensive due to generally less intensive development.
A normalized scoring scheme like the IndxMF_MSD is similar in concept to the Observed/Expected (O/E) index, which is used for bioassessments in several states (e.g., Oregon - Hubler 2008; Wyoming - Hargett et al. 2007). The O/E is centered around an index value of 1.0, which represents the best possible condition, in which all expected reference taxa are represented. Deviation from this optimal score represents loss or gain of taxa relative to the ideal reference conditions. The O/E index is evaluated based on the standard deviation of reference scores, which allows rapid translation of deviation from the optimal reference score (Hawkins et al. 2000).

The 100-point and normalized scoring schemes each have advantages and disadvantages (Wiley et al. 2002). The 100-point scale is easily communicated, whereas a normalized scale has intrinsic central tendency and variability and brings the reference distribution metrics between the two classes closer together. Because both have advantages, we carried the Indx_MinFloor and the associated IndxMF_MSD forward in analyses of biological condition thresholds, which are described in Appendix I.

5 References

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

Blocksom, K. 2003. A performance comparison of metric scoring methods for a multimetric index for mid-Atlantic highlands streams. Environmental Management 31(5):670 –682.

Gerritsen, J. and B.K. Jessup. 2007. Calibration of the Biological Condition Gradient for High-gradient Streams of Connecticut. Prepared for U.S. EPA Office of Science and Technology, and Connecticut DEP.
Hawkins, C.P., Norris, R.H., Hogue, J.N. and Feminella, J.W., 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecological Applications 10(5):1456-1477.

Hargett, E.G., J.R. ZumBerge, C.P. Hawkins and J.R. Olson. 2007. Development of a RIVPACS-type predictive model for bioassessment of wadeable streams in Wyoming. Ecological Indicators 7(4): 807-826.

Hawkins, C., Norris, R., Hogue, J., & J. Feminella. 2000. Development and Evaluation of Predictive Models for Measuring the Biological Integrity of Streams. Ecological Applications 10(5): 1456-1477. doi:10.2307/2641298

Hubler, S. 2008. PREDATOR: Development and use of RIVPACS-type macroinvertebrate models to assess the biotic condition of wadeable Oregon streams. DEQ08-LAB-0048-TR. https://www.oregon.gov/deq/FilterDocs/PredatorTechRep.pdf

Jessup, B., J. Stamp, J. Gerritsen, C. Carey, K. DeGoosh, S. Kiernan, and D. MacDonald. 2012. A Multimetric Biological Condition Index for Rhode Island Streams. Prepared for Rhode Island Department of Environmental Management, Office of Water Resources.

Jessup, B. and J. Stamp. 2019. Development of Indices of Biotic Integrity for Assessing Macroinvertebrate Assemblages in Massachusetts Freshwater Wadeable Streams. Prepared for the Massachusetts Department of Environmental Protection.

Wiley, M.J., Seelbach, P.W., Wehrly, K. and Martin, J.S., 2002. Regional ecological normalization using linear models: a meta-method for scaling stream assessment indicators. Biological response signatures: patterns in biological integrity for assessment of freshwater aquatic assemblages. CRC Press, Boca Raton, Florida, pp.201-224.