Development of an Index of Biotic Integrity for Macroinvertebrates in Freshwater Low Gradient Wadeable Streams in Southeast New England

FINAL REPORT

***Prepared for*:**

NEIWPCC
Maryann Dugan, Project Manager

Restore America's Estuaries Southeast New England Program
Tom Ardito, Grant Program Director

***Prepared by*:**

Ben Jessup, Ben Block, and Jen Stamp
Tetra Tech

April 2, 2021

Executive Summary

Under the Clean Water Act, state environmental agencies are charged with monitoring and assessment of streams and rivers. Currently, the Massachusetts Department of Environmental Protection (MassDEP) and the Rhode Island Department of Environmental Management (RIDEM) collect water chemistry data and sample biological communities to characterize the condition of streams. Where available, these data are compared against water quality standards and biological criteria that have been developed to quantify water quality conditions. Along the coast of southeast New England, non-tidal, low gradient, slow-moving streams that either lack or have infrequent riffle habitat are fairly prevalent. Yet, until recently, stream assessment efforts in New England have largely focused on moderate to high gradient, rocky-bottom streams with riffle habitats.

MassDEP and RI DEM have collected macroinvertebrate samples from riffle habitats for many years and have developed riffle habitat multimetric indices to assess the effects of anthropogenic stress on macroinvertebrate assemblages (Jessup et al. 2012, Jessup and Stamp 2020). Multimetric indices (also referred to as Indices of Biotic Integrity (IBIs)) are numeric representations of biological condition based on the combined signals of several different assemblage measurements (Karr 1981). 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. The index scores provide a measure of how far conditions at a site have deviated from the expected state of the macroinvertebrate community, which is based on comparisons with reference sites.

Because there are natural differences in the structure and function of macroinvertebrate assemblages in low gradient versus faster-moving, rocky-bottom streams, the collection methods and riffle habitat multimetric indices that MassDEP and RI DEM have developed cannot be effectively applied in the low gradient, slow-moving streams that occur along the coast of southeast New England. To address this, MassDEP developed a low gradient, multihabitat collection method for macroinvertebrates in 2013. The multihabitat method allowed for effective sampling of snags, root wads, leaf packs, aquatic macrophytes, undercut banks, overhanging vegetation, fine sediments, and hard substrates. In 2019, with funding from the U.S. Environmental Protection Agency (U.S. EPA) Southern New England Program (SNEP), the multihabitat collection method was used to sample over 50 sites in low gradient, non-tidal, wadeable streams in MA and RI. The sites were located in the SNEP region, which consists of coastal watersheds in Cape Cod, Narragansett Bay, Buzzards Bay, and the Islands. The intent of collecting these data was to obtain a dataset that could be used to calibrate a low gradient IBI for macroinvertebrate assemblages in the SNEP region. The Low-Gradient Coastal Index of Biotic Integrity (IBI) for Wadeable Waters in Southern New England project is supported by the Southeast New England Program (SNEP) Watershed Grants. SNEP Watershed Grants are funded by the U.S. Environmental Protection Agency (EPA) through a collaboration with Restore America's Estuaries (RAE) and awarded to the NEIWPCC. For more on SNEP Watershed Grants, see www.snepgrants.org.

In this report, we describe the development of a low gradient multihabitat IBI for the SNEP region. The IBI calibration dataset included data from 109 sites in Massachusetts (MA) and Rhode Island (RI). This work was done concurrently with the development of a statewide low gradient IBI for Massachusetts, which utilized data from an additional 69 low gradient sites located outside the SNEP region. There was overlap across the MassDEP and SNEP datasets and several staff members from MassDEP participated in both projects. Thus, the two projects were not completely independent and often were informing one another.

When developing the IBI, steps included compiling and preparing data, defining site disturbance categories and criteria, performing classification analyses, scoring and selecting metrics, compiling index alternatives, evaluating performance, and selecting and validating the final IBI. The top candidate IBIs had high discrimination efficiency (minimal error when discriminating between reference and stressed sites) and metrics that were familiar to the workgroup members, ecologically meaningful, and diverse in response mechanisms. The workgroup also wanted an IBI that performed well with different subsample sizes (300-, 200-, and 100-organism samples) to simplify application across the region.

The input metrics for the final IBI are listed in Table ES-1. The IBI had low error in the separation of index values in least-disturbed reference and most disturbed stressed sites (Index DE: 97.6%; higher discrimination efficiency indicates that a greater percentage of stressed index values are outside of the reference inter-quartile range) (Figure ES-1). As an alternate measure of performance, the relationship between IBI scores and four measures of disturbance (overall watershed condition at local and total watershed-scales, percent urban, and percent agriculture) were also evaluated. Associations with all but the percent agriculture metric were fairly strong (Spearman correlation coefficients ≥ |0.53|) and in keeping with the expected direction of response. Most sites had low percent agriculture, which likely accounts for the weak correlation between the IBI and percent agriculture.

To validate the IBI, relationships between IBI scores and stressor indicators that were not used in defining the IBI calibration stressor gradient were evaluated. The independent stressor variables included habitat scores, dissolved oxygen (DO), conductivity, and percent forest cover in the watershed. Some natural (non-stressor) variables were also compared, including acidity (pH), substrate, and temperature. Results confirmed that the IBI was indeed responsive along the stressor gradient.

As a final step, exploratory analyses were performed to inform potential numeric thresholds for four biological condition categories (Exceptional Condition, Satisfactory Condition, Moderately Degraded, and Severely Degraded). The thresholds proposed in this report are preliminary and subject to further review, refinement, and approval by MassDEP and RI DEM before they are applicable in biological assessment programs. The new low gradient IBI and preliminary thresholds improve the ability of MassDEP and RI DEM to identify degradation in biological integrity and water quality and will be re-evaluated in coming years as they obtain and analyze more low gradient samples.

Table ES-1. Metrics included in the low gradient IBI.

Metric (abbrev)	Response to increasing stress	Scoring formula
% Plecoptera, Odonata, Ephemeroptera, and Trichoptera (POET) taxa (pt_POET)	Decrease	100*(metric)/40
% Predator taxa (pt_ffg_pred)	Decrease	100*(metric)/32
% Non-insect taxa (pt_NonIns)	Increase	100*(46-metric)/42
% Odonata, Ephemeroptera, and Trichoptera (OET) individuals (pi_OET)	Decrease	100*(metric)/49
% Tolerant taxa (pt_tv_toler)	Increase	100*(36-metric)/33
% Semivoltine taxa (pt_volt_semi)	Decrease	100*(metric)/12

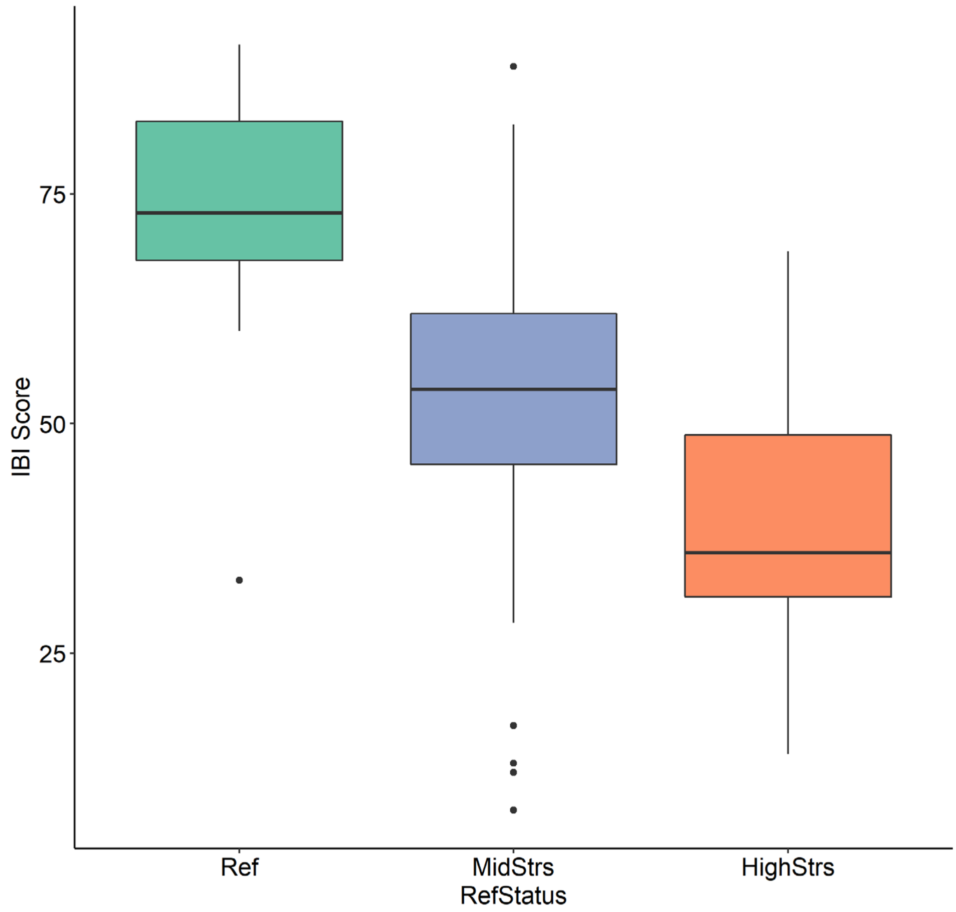


Figure ES-1. Distributions of low gradient IBI values in reference (Ref), intermediate (MidStrs), and stressed (HighStrs) sites.

Acknowledgments

The index development process was funded by a grant from the U.S. Environmental Protection Agency Southeast New England Program (SNEP). The contract to Tetra Tech was facilitated by Maryann Dugan of NEIWPCC. The index development team consisted of biologists from the Massachusetts Department of Environmental Protection (MassDEP) (James Meek, Allyson Yarra, Robert Nuzzo) and the Rhode Island Department of Environmental Management (RIDEM) (Katie DeGoosh, Jane Sawyers). Michael Cole of Cole Ecological, Inc. provided feedback on taxa attribute assignments and nomenclature questions, and worked with Kelly Nolan and Natalie Ohl (Watershed Assessment Associates) on the taxonomic identification quality control process. In addition, several state partners from the Northeast provided low gradient data that were used in the regional taxa tolerance analyses. Regional partners included Ansel Aarrestad and Chris Bellucci from the Connecticut Department of Energy & Environmental Protection, Aaron Moore from the Vermont Agency of Agriculture Food and Markets (AAFM), Steve Fiske (retired, formerly of the Vermont Department of Environmental Conservation), and Gavin Lemley, Brian Duffy and Zachary Smith from New York Department of Environmental Conservation. We are very grateful for the hard work and enthusiasm of all the project participants.

Project authors and analysts included Ben Jessup, Ben Block, and Jen Stamp of Tetra Tech. An appropriate citation for this report is as follows:

Jessup, B., B. D. Block, and J, Stamp. 2021. Development of an Index of Biotic Integrity for Macroinvertebrates in Freshwater Low Gradient Wadeable Streams in Southern New England Prepared for NEIWPCC, Worcester, MA. Prepared by Tetra Tech, Montpelier, VT.

Acronyms

Acronym	Description
BFI	Base Flow Index
BPJ	Best Professional Judgement
Cat	Catchment
CT DEEP	Connecticut Department of Energy & Environmental Protection
CV	Coefficient of Variation
CWA	Clean Water Act
DE	Discrimination Efficiency
DO	Dissolved Oxygen
FFG	Functional Feeding Group
GIS	Geographic Information System
IBI	Index of Biotic Integrity
ICI	Indices of Catchment Integrity
IWI	Indices of Watershed Integrity
MassDEP	Massachusetts Department of Environmental Protection
MSST	Mean Summer Stream Temperature
NBL	Narragansett-Bristol Lowlands
NLCD	National Land Cover Database
NMS	Non-metric multidimensional scaling
NPDES	National Pollutant Discharge Elimination System
NPL	Superfund National Priority List
NRSA	U.S EPA National Rivers and Streams Assessment
NYSDEC	New York State Department of Environmental Conservation
PCA	Principle components analysis
PRISM	PRISM Data Explorer
QAPP	Quality Assurance Project Plan
QC	Quality Control
RBP	Rapid Bioassessment Protocol
RI DEM	Rhode Island Department of Environmental Management
RMN	Regional Monitoring Network
SNECPAH	Southern New England Coastal Plains and Hills
SNEP	Southeast New England Program
SWQS	Massachusetts Surface Water Quality Standards
U.S. EPA	United States Environmental Protection Agency
VT DEC	Vermont Department of Environmental Conservation
Ws	Watershed

Table of Contents

List of Tables

List of Figures

1 Background

Under the Clean Water Act, state environmental agencies are charged with monitoring and assessment of streams and rivers in all areas of southern New England. Currently, the Massachusetts Department of Environmental Protection (MassDEP) and the Rhode Island Department of Environmental Management (RIDEM) collect water chemistry data and sample biological communities to characterize the condition of streams. Where available, these data are compared against water quality standards and biological criteria that have been developed to quantify water quality conditions. Monitoring biological communities, especially macroinvertebrates, in low gradient streams along the coast of southeast New England provides important information about the water quality and the health of aquatic ecosystems.

The MassDEP and RI DEM biomonitoring programs have collected macroinvertebrates from riffle habitats in moderate to high gradient, rocky-bottom streams for many years. Both states have developed riffle habitat multimetric indices to assess the effects of anthropogenic stress on macroinvertebrate assemblages (Jessup et al. 2012, Jessup and Stamp 2020). Multimetric indices (also referred to as Indices of Biotic Integrity (IBIs)) are numeric representations of biological condition based on the combined signals of several different assemblage measurements (Karr 1981). 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. The index scores provide a measure of how far conditions at a site have deviated from the expected state of the macroinvertebrate community.

In southern Massachusetts (MA) and Rhode Island (RI), low gradient, slow-moving streams that either lack or have infrequent riffle habitat are fairly prevalent. Because there are natural differences in the structure and function of macroinvertebrate assemblages in low gradient versus faster-moving, rocky-bottom streams, the collection methods and bioassessment indices that were developed for riffle habitats cannot be effectively applied in these streams. To address this, MassDEP developed a multihabitat collection method for macroinvertebrates in low gradient, slow-moving streams in 2013. The multihabitat method allowed for effective sampling of snags, root wads, leaf packs, aquatic macrophytes, undercut banks, overhanging vegetation, fine sediments, and hard substrates. In 2019, with funding from the U.S. Environmental Protection Agency (U.S. EPA) Southern New England program (SNEP), the multihabitat collection method was used to sample over 50 sites in low gradient, non-tidal, wadeable streams in MA and RI. The sites were located in the SNEP region, which consists of watersheds draining into Narragansett Bay and Buzzards Bay and south from Cape Cod , including the Islands. The intent of collecting these data was to obtain a dataset that could be used to calibrate a low gradient IBI for macroinvertebrate assemblages in the SNEP region.

In this report, we describe the development of a low gradient IBI for macroinvertebrate assemblages in non-tidal, wadeable streams in the SNEP region. Data collection and index development was done concurrently with the development of a statewide low gradient IBI for MassDEP. Data collected by MassDEP using multihabitat methods in the SNEP region were included in the SNEP analysis. Data from an additional 69 low gradient sites located in Massachusetts and outside the SNEP region were used in certain steps of the SNEP analysis, and are described at those steps.

Steps in the IBI development process included data compilation and preparation, definition of site disturbance categories and criteria, classification analyses, metric selection and scoring, index compilations, performance evaluation, selection of the final IBI, and IBI validation. The report concludes with an evaluation of potential IBI thresholds for four levels of biological condition and a discussion on potential applications. The creation of an IBI for coastal low gradient streams in the SNEP region will improve resource managers' ability to identify degradation in biological integrity and water quality and help inform prioritization of streams for protection and restoration.

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 Microsoft (MS) Access relational database.

2.1 Macroinvertebrates

2.1.1 Dataset

The low gradient IBI dataset spanned seven years (2013-2019) and included a total of 114 samples from 109 unique sites in the SNEP study area in RI and MA. Twenty-two sites were located in RI and 87 in MA (Figure 1, Table 1). MassDEP collected 60 of the samples over the seven year period and Tetra Tech (under contract to SNEP) collected 54 samples in 2019. The distribution of sites across Level 4 ecoregions is summarized in Table 1. Most sites were located in the Narragansett/Bristol Lowland and Southern New England Coastal Plains and Hills (SNECPAH) Level 4 ecoregions. Seven were located in the Cape Cod Level 4 ecoregion. For some analyses, we utilized low gradient data from an additional 69 low gradient sites in MA that were located outside the SNEP region (Figure 1).

Table 1. Distribution of the 109 sites across states and Level 3 and 4 ecoregions (U.S. EPA 2011).

L3 ecoregion	**Level 4 ecoregion name**	**Level 4 code**	**Number of sites**	
			MA	**RI**
Atlantic Coastal Pine Barrens	Cape Cod/Long Island	84a	7	0
Northeastern Coastal Zone	Gulf of Maine Coastal Plain	59h	2	0
	Long Island Sound Coastal Lowland	59g	0	3
	Narragansett/Bristol Lowland	59e	66	5
	Southern New England Coastal Plains and Hills	59c	12	14
		Total	***87***	**22**

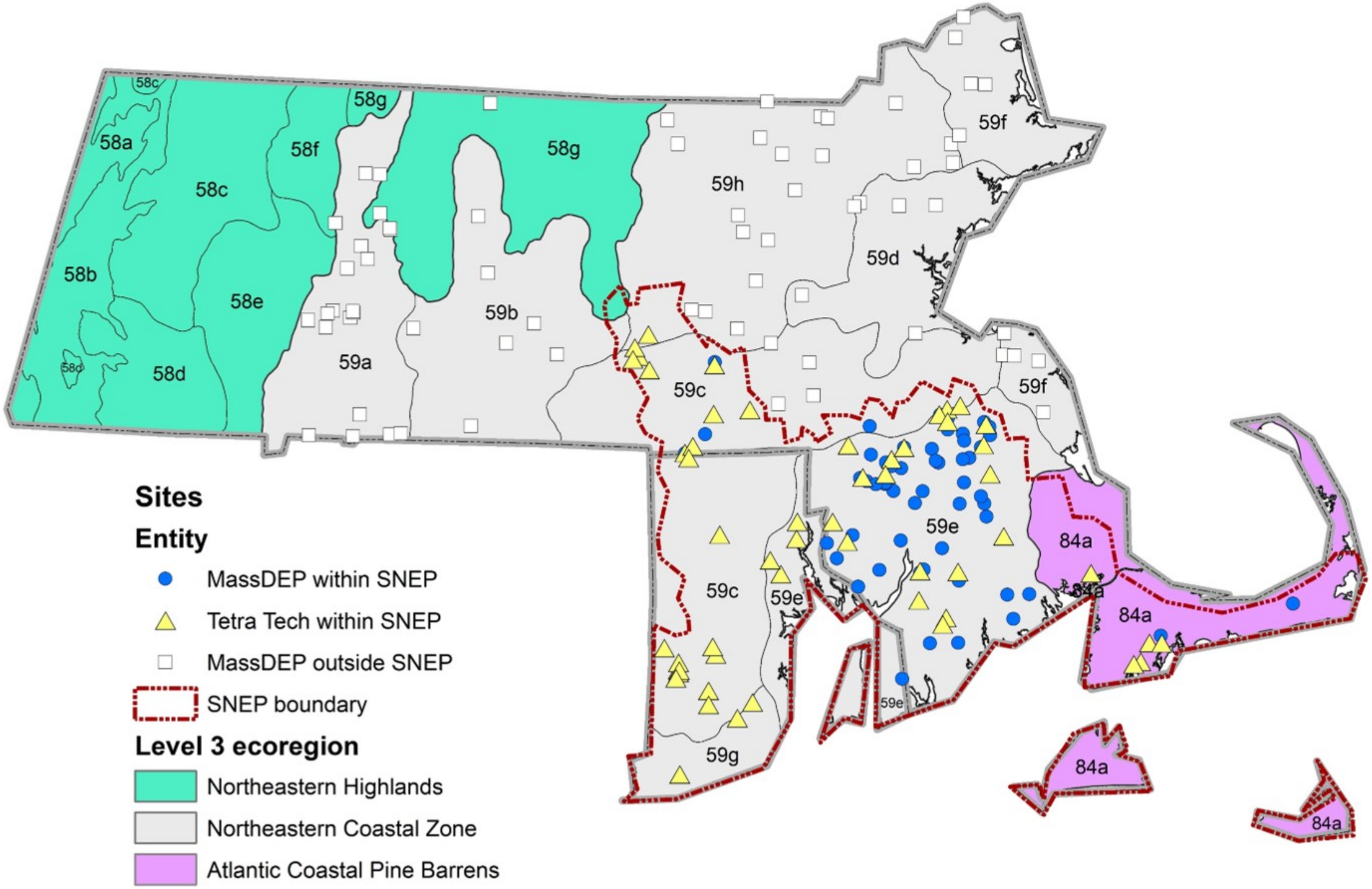


Figure 1. Locations of sites in the SNEP low gradient IBI development dataset (n=109 unique sites), coded by sampling entity (MassDEP or Tetra Tech), with Level 3 ecoregions as the backdrop. An additional 69 low gradient sites in MA that are outside the SNEP boundary were included in some of the analyses.

2.1.2 Collection method

Macroinvertebrate data were collected by MassDEP and Tetra Tech field crews. The MassDEP samples were collected in accordance with MassDEP's standard operating procedures (Nuzzo 2003) and Quality Assurance Project Plan (QAPP) (MassDEP 2004) and the SNEP samples were collected following the SNEP IBI Sampling Analysis Plan (Tetra Tech 2019). Samples consisted of a composite of 10 jabs, sweeps, or kicks from multiple habitats within a 100-meter reach. Samples were collected from July 1 through September 30 when baseflows are typically at the lowest of the year and levels of stress to aquatic organisms are presumed to be greatest. Major habitat types included submerged wood, submerged vegetation, undercut banks, overhanging vegetation, and hard substrate. Habitats were sampled in rough proportion to their occurrence within the reach. For example, if the habitat was 50% submerged wood, 30% submerged vegetation and 20% vegetated margins/banks, then five jabs were taken from submerged wood, three from submerged vegetation, and two from vegetated margins/banks. Field crews used a kick-net with 500 to 600-µm mesh. Table 2 summarizes the MassDEP and SNEP low gradient protocols. The main differences between the protocols were that MassDEP used a brush on woody debris and Tetra Tech field crews used a net with a smaller frame size (28-cm wide opening vs. 46-cm for MassDEP). The SNEP protocols also specify a time limit on each jab (between 30 to 45 seconds), while MassDEP protocols do not. However, MassDEP uses a comparable level of effort (James Meek (MassDEP), personal communication).

Samples were labeled and preserved in the field with denatured 95% ethanol, then brought to the 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 approximately 300 organisms were extracted. Specimens were identified to genus or species as allowed by available keys, specimen condition, and specimen maturity. Cole Ecological, Inc. processed and identified the samples. As a quality control (QC) measure, ten randomly selected samples from the 2019 dataset were independently identified and enumerated both by Cole Ecological, Inc. and Watershed Assessment Associates. The results, which are provided in Attachment A, met the data quality objectives in the MassDEP and SNEP sampling plans.

Table 2. Summary of the protocol elements for the Massachusetts Department of Environmental Protection (MassDEP) and Southeast New England Program (SNEP) low gradient macroinvertebrate methods.

Method	Habitat	Effort	Gear	Reach length	Index period	Target # organisms	Taxonomic resolution
MassDEP multihabitat	Snags and root wads, leaf packs, aquatic macrophytes, undercut banks and overhanging vegetation, hard bottom (riffle/cobble/boulder)	Any combination of 10 kicks, sweeps, and/or jabs, which are then combined into a single composite sample. Sampling is proportional to the relative makeup of the reach by the listed habitat types*	Kick-net with 500-µm mesh, 46-cm wide opening. Brushes are used on woody debris	100-m	July 1 – September 30	300	Lowest practical level
SNEP multihabitat	Submerged wood (including leaf packs wedged in the wood), submerged vegetation, undercut banks/overhanging vegetation, hard bottom/rocky substrates	Composite of 10 jabs, sweeps, or kicks; each jab/sweep/kick lasted for a minimum of 30 seconds and a maximum of 45 seconds. The goal is to dislodge and capture as many organisms as possible in that area. The habitats will be sampled in rough proportion to their occurrence within the reach*	Kick-net with 500-µm mesh and ~28-cm wide opening; brushes are *not* used on woody debris				

*For example, if the habitat is 50% submerged wood, 30% submerged vegetation and 20% vegetated margins/banks, then 5 jabs will be taken from submerged wood, 3 from submerged vegetation, and 2 from vegetated margins/banks. A comparison of habitat types defined by each agency is in Appendix A.

2.1.3 Taxa attributes

We compiled the MassDEP and Tetra Tech macroinvertebrate data into an MS Access relational database. For trait assignments, we used the attribute table that had been created during the calibration of the MassDEP riffle habitat IBI as a starting point (Jessup and Stamp 2020). The table included five sets of traits: functional feeding group (FFG), tolerance value, habit, life cycle/voltinism, and thermal preferences (Table 3). Based on guidance from Cole Ecological, Inc., we updated some of the phylogeny and taxa names to reflect the most current nomenclature and keys and re-checked the attribute assignment based on the sources listed in Table 3.

To help inform tolerance value assignments (which could differ in low vs. higher gradient streams), we ran taxa tolerance analyses on the regional low gradient dataset to explore the distribution of taxa across four generalized disturbance measures: the Indices of Catchment and Watershed Integrity (ICI and IWI, respectively), percent urban and percent agricultural land cover (Thornbrugh et al. 2018, Johnson et al. 2018). Taxa that occurred at fewer than 10 sites were excluded from the analysis because low numbers of occurrences gave unreliable results. Tolerance analyses allow for visualization of the shape of the taxon-stressor relationship across a continuous numerical scale and can be used to identify optima (the point at which the taxon has the highest probability of occurrence) as well as tolerance limits (the range of conditions in which the taxon can persist) (Yuan 2006). To increase the sample size and improve the robustness of the analysis, the analyses were also run on a larger regional dataset that included low gradient data from outside of the SNEP region in Massachusetts, Connecticut, Vermont, and New York. Biologists from MassDEP reviewed results from the analyses and assigned taxa to three tolerance categories: intolerant, intermediate, and highly tolerant (Table 3). More detailed information on the tolerance analyses can be found in Appendix B.

The taxa attribute table is provided in Attachment B. Table 3 shows what percentage of the 542 taxa in the SNEP IBI calibration dataset had attribute assignments for each trait group. FFG was the most complete (97%) while voltinism had the lowest number of assignments (46%). Metrics were calculated with the BioMonTools R package (Leppo et al. 2021). Appendix C contains the list of metrics that were calculated and considered as candidates for inclusion in the IBIs. When developing the list of candidate metrics, we researched metrics being used in other existing low gradient IBIs. Results of that exercise are provided in Appendix C. When making metric calculations, non-target taxa (e.g., Hemiptera, crayfish) were excluded from all metrics and redundant/non-distinct taxa were excluded from the richness metrics (for more information, see Appendix C).

Table 3. Five sets of traits were included in the taxa attribute table for the low gradient SNEP dataset.

Attribute	Description	Categories	Sources*	Number of taxa with attribute assignments (out of 542)	Percent of total
Functional feeding group (FFG)	Refers to the primary process for acquiring food resources	PR = predator, CG = collector-gatherer, SH = shredder, SC = scraper, CF = collector-filterer	MassDEP, CT DEEP, VT DEC, NRSA*	526	97.0%
Tolerance values (TolVal)	Relative sensitivity to pollution, disturbance	Three categories: intolerant (numeric value = 2), intermediate (numeric value = 5) and highly tolerant (numeric value = 8)	Primary: taxa tolerance analyses on the MA/SNEP and regional low gradient datasets. Secondary: riffle habitat assignments from MassDEP, VT DEC, CT DEEP	406	74.9%
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	247	45.6%
Habit	Distinguishes the primary mechanism a particular species utilizes for maintaining position and moving in the aquatic environment (Merritt and Cummins 1996)	SP = sprawler, SW = swimmer, CN = clinger, CB = climber, BU = burrower	NRSA, VT DEC, Poff et al. 2006	479	88.4%
Thermal preference	Thermal preference/optima	Cold_cool or warm	U.S. EPA 2012, U.S. EPA 2016	75**	NA**

*Source abbreviations: Connecticut Department of Energy & Environmental Protection (CT DEEP), Vermont Department of Environmental Conservation (VT DEC), New York State Department of Environmental Conservation (NYSDEC), and EPA National Rivers and Streams Assessment (NRSA)

**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

Habitat and water quality data were collected by field crews at the time of the biological sampling events. Table 4 lists parameters that were collected by both MassDEP and Tetra Tech. These data were used in classification analyses and, where appropriate, in site disturbance characterizations. At the 2019 SNEP sites, Tetra Tech collected additional exploratory parameters such as counts of woody debris and flow velocity measurements (for more information, see Appendices D and F in the SNEP IBI Sampling Analysis Plan; Tetra Tech 2019).

Habitat surveys were performed in accordance with the RBP Rapid Habitat Assessment protocols for low gradient, glide-pool (GP) streams (Barbour et al. 1999). The riffle/run (RR) assessment, which is slightly different, was also performed at a few sites that had characteristics of both RR and GP stream types. The RBP-GP assessment includes ten input metrics: epifaunal substrate/available cover, pool substrate characterization, pool variability (size/depth), sediment deposition, channel flow status, channel alteration, channel sinuosity, bank stability, bank vegetative protection, and riparian vegetative zone width. 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). Habitat scores are estimated by the field crews and are subject to variable interpretations of the scoring scales. However, the crews undergo training and inter-crew calibration during each sampling season to improve estimates of habitat conditions.

Other habitat measures included visual estimates of substrate composition (clay, sand, gravel, cobble, boulder, bedrock), the number of jabs from each major habitat group (submerged wood, submerged vegetation, vegetated margins/undercut banks, and hard bottom), visual estimates of percent canopy cover and mean width, maximum depth and the high water mark (Table 4). Field crews also collected *in situ* water quality data (temperature, conductivity, dissolved oxygen, and pH), and qualitative assessments of color, odor, surface oils, turbidity, where available[1]. Field crews also took photographs of the sites. The photos show the diversity of low gradient sites represented in the IBI calibration dataset, ranging from slow winding, soft bottom streams to slow moving streams with rocky substrates (Figure 2). Stream color ranged from colorless to dark and substrate size and major habitat types varied across sites. Overall, the highest proportion of jabs were taken from submerged wood, (median = 5 out of the 10 jabs) (Figure 3). More detailed information on habitat types can be found in Appendix A.

[1] MassDEP 2019 *in situ* data had not been QC'd in time to use in the analyses. Some of the other sites were missing data due to equipment malfunctions.

Table 4. Habitat variables that were collected by MassDEP and Tetra Tech field crews at the time of the biological sampling events.

Habitat variables	Description
Number of jabs from each major habitat group (10 jabs total)	Four major habitat groups: submerged wood, submerged vegetation, vegetated margins/undercut banks, and hard bottom, sampled in proportion to their occurrence*.
Rapid Habitat Assessment (Barbour et al. 1999)	Visual assessment of the sampling reach. Ten input metrics: epifaunal substrate/available cover, pool substrate characterization, pool variability, degree and type(s) of channel alteration, sediment deposition, channel sinuosity, channel flow status, bank vegetative protection, bank stability, and riparian vegetation zone width.
Substrate composition (%)	A visual estimate of the percentage of inorganic substrates (clay, silt, sand, gravel, cobble, boulder, bedrock) (should sum to 100%) and organic substrates (detritus, muck-mud, marl) (does not need to sum to 100%) throughout the sampling reach.
Canopy cover (%)	A visual estimate of the percent of the wetted area of the sampling reach that is shaded by overhanging vegetation or other structures.
Width (m)	Wetted distance from bank to bank, either based on a single measurement from the portion of the reach that is the most representative of the natural channel, or, if width varies throughout the reach, based on the average from three locations (upstream end, downstream end, and mid-point).
Maximum Depth (m)	Maximum depth in the sampling reach.
High water mark (m)	The vertical distance from bankfull (at base flow) to the high water level indicator (e.g., debris hanging in riparian or floodplain vegetation, deposition of silt or soil).

*MassDEP enters slightly different habitat categories into their database than the ones used by Tetra Tech. Appendix C contains the crosswalk table that was used to align the categories.

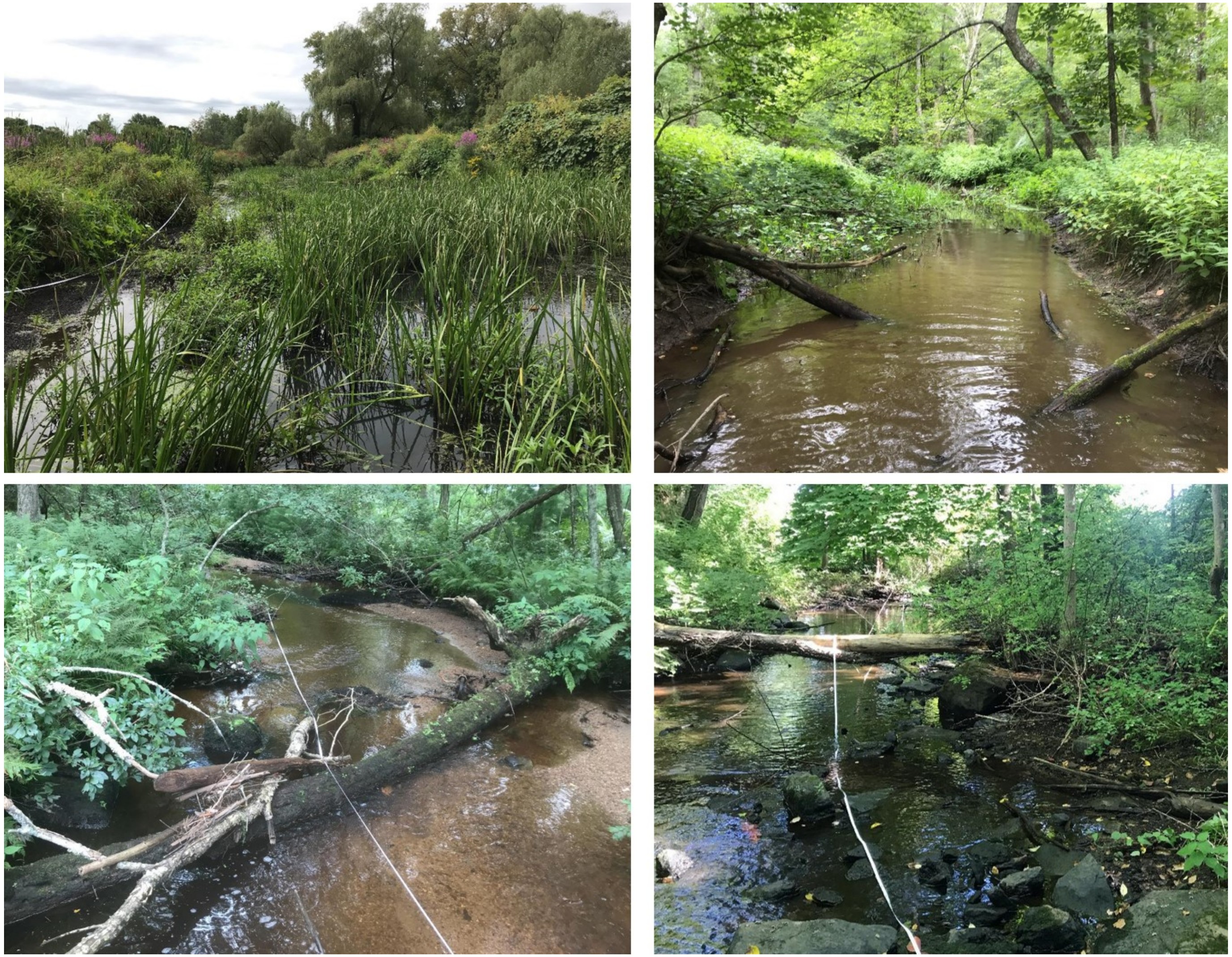


Figure 2. A diverse group of low gradient sites are represented in the IBI calibration dataset, ranging from slow winding, soft bottom streams to slow-moving streams with gravel or cobble substrate.

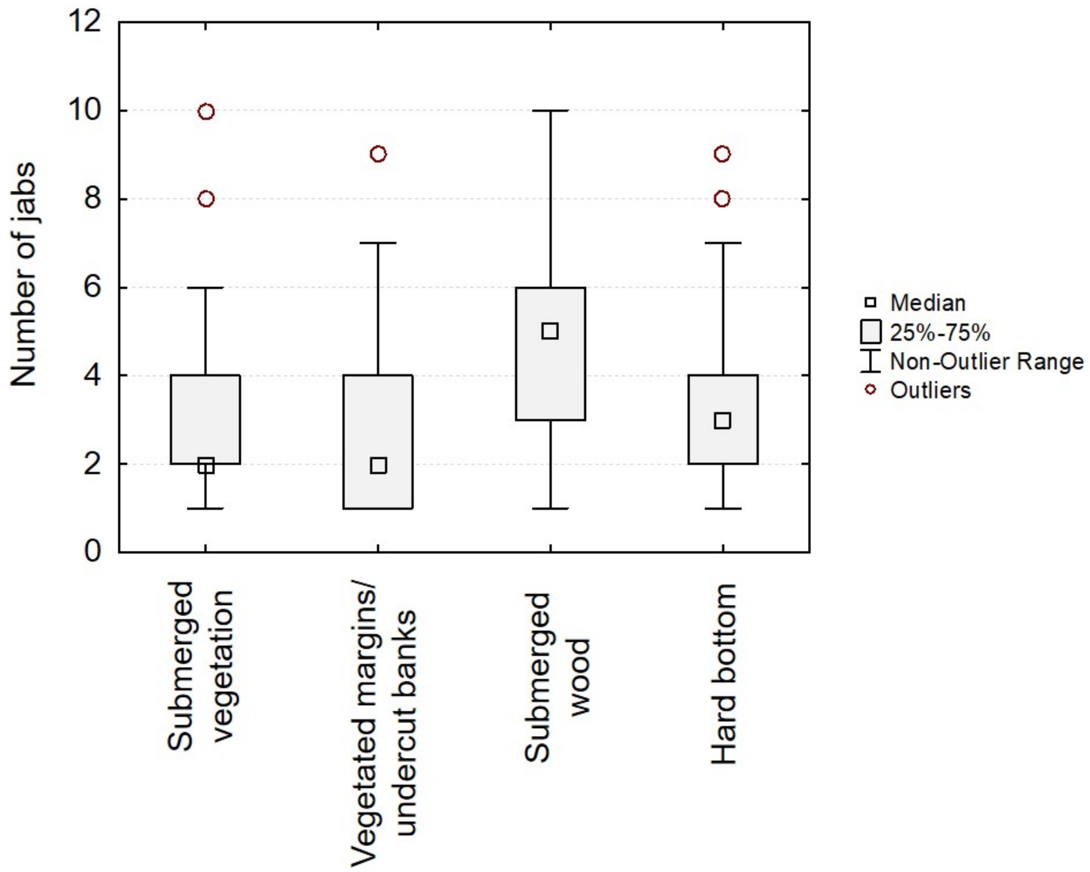


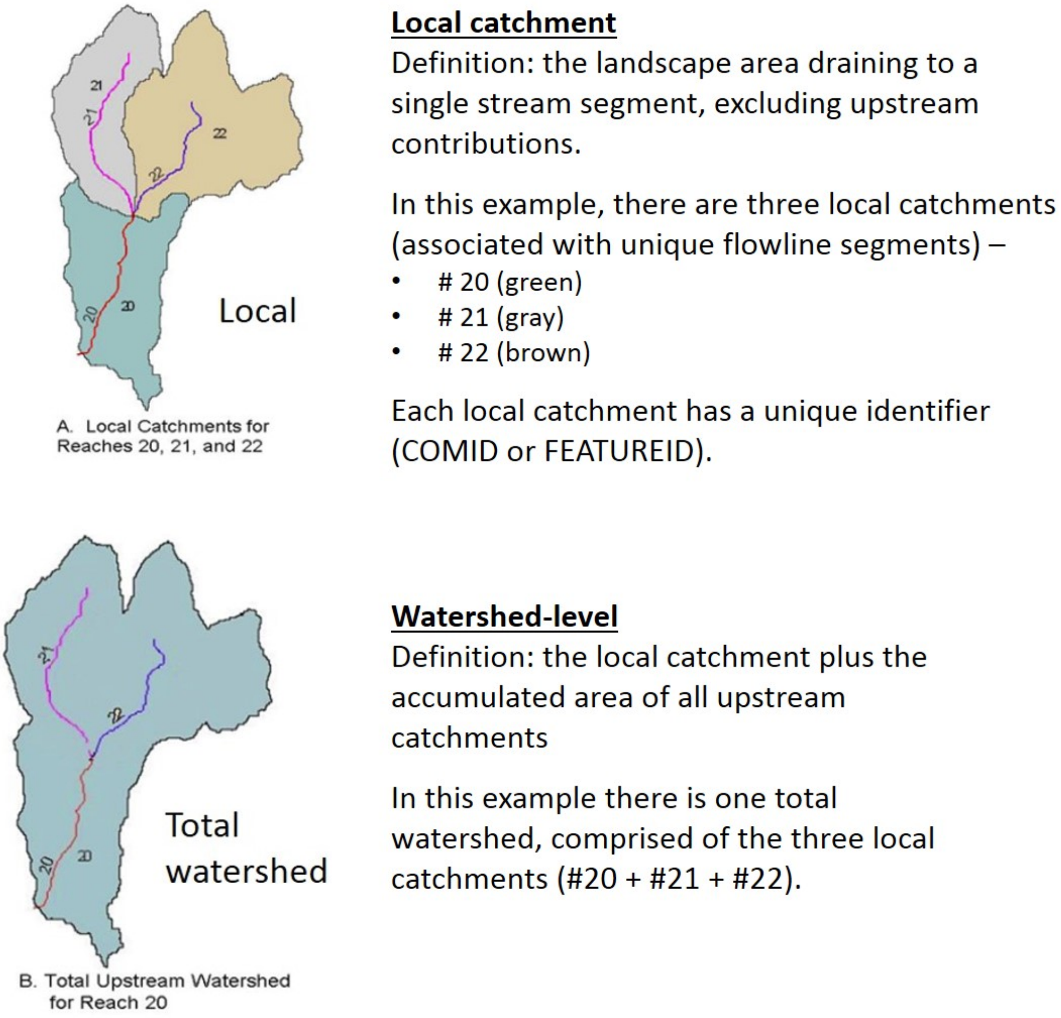
Figure 3. Distribution of jabs per site across the four major habitat types. A total of 10 jabs were taken per site. For more information on the habitat types, see Appendix C.

2.3 Landscape-scale Information (GIS-based)

Landscape-scale metrics were obtained for site disturbance characterization (Section 3) and classification (Section 4). A primary data source was the USEPA 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., overall watershed condition (ICI and IWI), percent agricultural cover, percent 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, mean annual temperature, and precipitation, among others. In addition, NHDPlusV2 attribute data for flowline type (stream/river, canals/ditches, coastline, and artificial pathway) and slope were associated with biological sampling sites, as were EPA level III and IV ecoregions.

To associate the biological sampling sites with the StreamCat dataset, an intersect procedure was performed with Geographic Information System software (ArcGIS 10.7.1), 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[2]. The data were uploaded to MS Access and queries were created to generate tables with the desired StreamCat metrics.

The StreamCat 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. To obtain more accurate, site-specific data, we used USGS StreamStats[3] to delineate exact watersheds for each site, and then used the Regional Monitoring Network (RMN) GIS ArcMap tools (Gibbs and Bierwagen 2017) to generate land cover statistics, drainage area, sinuosity, flowline slope, watershed slope, and baseflow. The land cover statistics were based on the 2016 National Land Cover Database (NLCD). We used land cover data from two spatial scales (1-km upstream and total watershed) in our site disturbance analyses. For sinuosity and flowline slope, we traced flowlines and used the RMN GIS tools to calculate values for 500 and 1000-meter stream lengths. In addition, we screened for dams, mines, National Pollutant Discharge Elimination System (NPDES) major discharge permits, and Superfund National Priority List (NPL) sites within the 1-km upstream watershed.



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).

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.

[2] https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset-0

[3] https://streamstats.usgs.gov/ss/

3 Site Disturbance Characterization

3.1 Purpose

Bioassessment is based on a comparison of conditions in assessable waterbodies to sites with relatively natural environmental conditions, which are 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.

3.2 Approach

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 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).

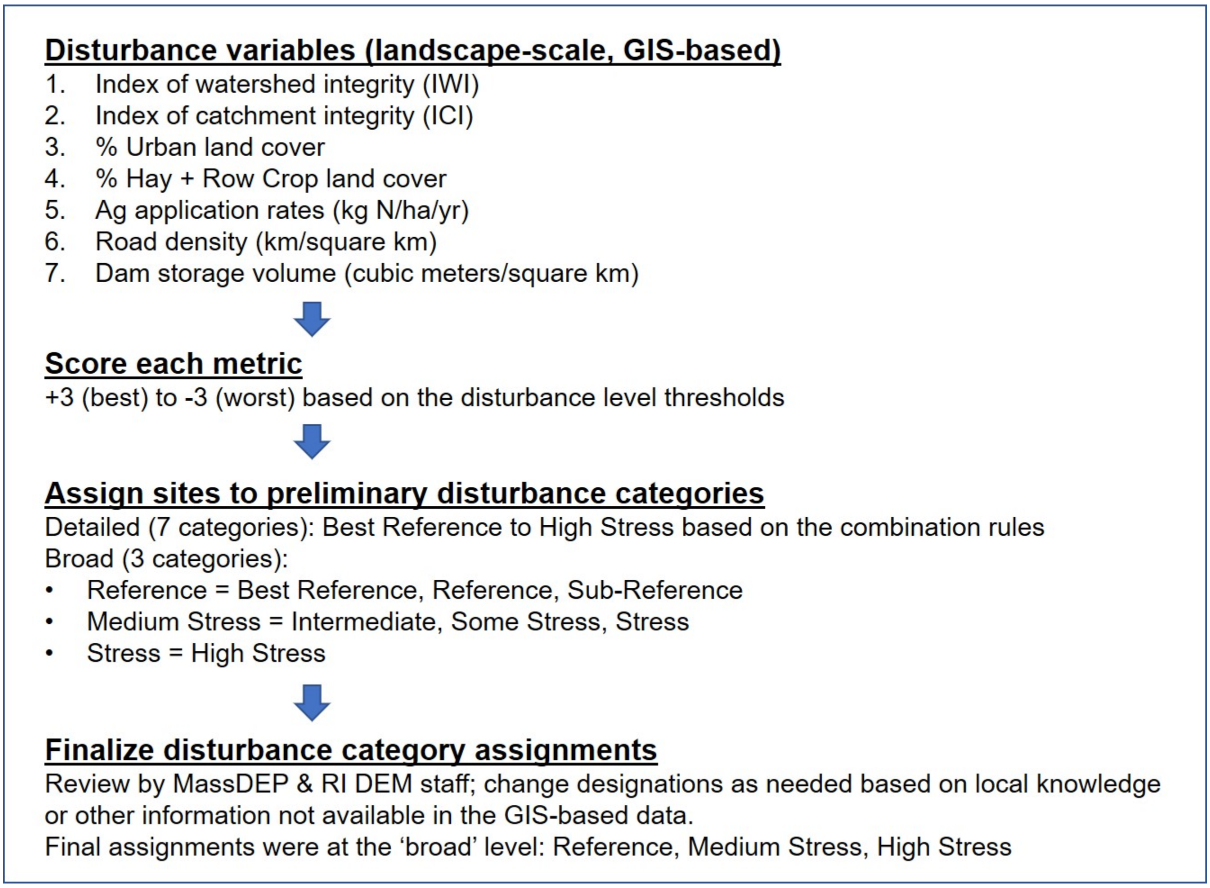
For this exercise, we used a modified version of the disturbance index that was developed during calibration of the MassDEP 100-count riffle habitat IBI (Jessup and Stamp 2020). We used the same seven metrics: ICI, IWI, percent urban land cover, percent agricultural land cover (local catchment), density of roads, dam storage volume, and modeled mean rate of fertilizer application + biological nitrogen fixation + manure application (Table 5). The low gradient disturbance index differed from the one used for the MassDEP riffle habitat IBI in that:

- We switched to version 2.1 of the ICI and IWI (in place of version 1) and adjusted the ICI and IWI metric thresholds to account for this change
- We switched to the 2016 NLCD land cover metrics (in place of NLCD 2011)
- We used two spatial scales (local and total watershed) instead of one
- Land cover statistics were based on exact watershed delineations

Table 5. Seven disturbance variables were used to assign sites to preliminary disturbance categories. Information on variable selection can be found in the MassDEP 100-count riffle habitat IBI report (Jessup and Stamp 2020).

Disturbance variable	Spatial scale	Source	Units	Description
Index of catchment integrity (ICI 2.1)	Local catchment (Cat)	Version 2.1	0 (worst) -1 (best)	A measure of overall watershed condition, based on six components: hydrologic regulation, regulation of water chemistry, sediment regulation, hydrologic connectivity, temperature regulation, and habitat provision
Index of watershed integrity (IWI 2.1)	Upstream watershed (Ws)			
Percent Urban land cover	Maximum value across two scales (1-km upstream, total watershed)	NLCD 2016	percent (0-100)	Percent of area classified as developed, high + medium + low-intensity land use (NLCD classes 24+23+22)
Road density	Maximum value across two scales (Cat, Ws)	Road layer = 2010 Census Tiger Lines	km/km^2	The density of roads within the area
Percent Agricultural (hay/crop) land cover	Maximum value across two scales (1-km upstream, total watershed)	2016 NLCD	percent (0-100)	Percent of the area classified as hay and crop land use (NLCD classes 82+81)
Mean rate of fertilizer application + biological nitrogen fixation + manure application	Maximum value across two scales (Cat, Ws)	EnviroAtlas	mean rate kg N/ ha/yr	[Mean rate of biological nitrogen fixation from the cultivation of crops (CBNF)] + [Mean rate of synthetic nitrogen fertilizer application to agricultural land within area (Fert)] + [Mean rate of manure application to agricultural land from confined animal feeding operations within area (Manure)]
Dam storage volume	Maximum value across two scales (Cat, Ws)	Army Corps of Engineers (ACOE)	m^3/km^2	Volume all reservoirs per unit area. Based on typical volumes stored within reservoirs (NORM_STORA in NID)

We followed the process outlined in Figure 5 to assign sites to disturbance categories. Each of the seven metrics was scored based on their value in relation to the thresholds in Table 6. For example, if a site had an IWI of 0.9, it received an IWI score of +3; or if it had an IWI score of 0.55, it received an IWI score of -1. The metric scores were then considered in combination, using the 'combination rules' described in Table 6. Sites were assigned to one of seven preliminary disturbance categories, ranging from Best Reference to Highly Stressed, which were then collapsed into three broader categories (reference, medium stress, and stressed). The preliminary designations were then reviewed by staff from MassDEP and RI DEM, who either confirmed or changed the designations. Sites were then mapped and color-coded by disturbance category to ensure that their spatial distribution matched with expectations (Figures 6 and 7). Of the 109 sites, 26 were designated as reference sites, 23 as stressed sites, and 60 as medium stress sites. Figure 8 shows the range of disturbance represented in the reference and stressed dataset, as measured by the ICI, IWI, percent urban, and percent agricultural land cover. Appendix D contains additional box plots with disturbance variables as well as natural variables (such as drainage area, slope, temperature, and elevation). Appendix E has additional maps of natural variables. Attachment C contains the site list with preliminary and final disturbance category assignments.



Disturbance variables (landscape-scale, GIS-based)

1. Index of watershed integrity (IWI)
2. Index of catchment integrity (ICI)
3. % Urban land cover
4. % Hay + Row Crop land cover
5. Ag application rates (kg N/ha/yr)
6. Road density (km/square km)
7. Dam storage volume (cubic meters/square km)

Score each metric

+3 (best) to -3 (worst) based on the disturbance level thresholds

Assign sites to preliminary disturbance categories

Detailed (7 categories): Best Reference to High Stress based on the combination rules
Broad (3 categories):

- Reference = Best Reference, Reference, Sub-Reference
- Medium Stress = Intermediate, Some Stress, Stress
- Stress = High Stress

Finalize disturbance category assignments

Review by MassDEP & RI DEM staff; change designations as needed based on local knowledge or other information not available in the GIS-based data.
Final assignments were at the 'broad' level: Reference, Medium Stress, High Stress

Figure 5. Process for assigning sites to disturbance categories. Information on variable selection and development of the disturbance gradient can be found in Jessup and Stamp (2020).

Table 6. Metric scoring thresholds and combination rules that were used to assign sites to preliminary disturbance categories. More detailed information on how metrics and scoring thresholds were selected can be found in the MassDEP riffle habitat IBI report (Jessup and Stamp 2020). Metrics scores of +3 represent least disturbed conditions, while -3 represents the most highly disturbed conditions.

Metric Scores	**IWI (2.1)**	**ICI (2.1)**	**% Urban**	**% Hay/Crop**	**Fertilizer application**	**Road density**	**Dam storage volume**
+3	≥ 0.85	≥ 0.85	≤ 1	≤ 1	≤ 0.5	≤ 1.5	≤ 0.1
+2	< 0.85 and ≥ 0.80	< 0.85 and ≥ 0.80	> 1 and ≤ 2	> 1 and ≤ 2	> 0.5 and ≤ 1	> 1.5 and ≤ 2	> 0.1 and ≤1,000
1	< 0.80 and ≥ 0.70	< 0.80 and ≥ 0.70	> 2 and ≤ 5	> 2 and ≤ 5	> 1 and ≤ 2.5	> 2 and ≤ 3	> 1000 and ≤ 10,000
0	< 0.70 and > 0.60	< 0.70 and > 0.60	> 5 and < 10	> 5 and < 10	> 2.5 and < 5	> 3 and < 5	> 10,000 and < 50,000
-1	≤ 0.60 and > 0.50	≤ 0.60 and > 0.50	≥ 10 and < 40	≥ 10 and < 15	≥ 5 and < 7.5	≥ 5 and < 7.5	≥ 50,000 and < 100,000
-2	≤ 0.50 and > 0.40	≤ 0.50 and > 0.40	≥ 40 and < 60	≥ 15 and < 20	≥ 7.5 and < 10	≥ 7.5 and < 10	≥ 100,000 and < 200,000
-3	≤ 0.40	≤ 0.40	≥ 60	≥ 20	≥ 10	≥ 10	≥ 200,000

Combination rules for assigning sites to preliminary disturbance categories
Best Reference: all metrics meet the +2 scoring thresholds or better
Reference: all metrics meet the +1 scoring thresholds or better
Sub Reference: All metrics meet the 0 scoring thresholds and at least five metrics receive positive scores (> 0)
Intermediate: All metrics meet the 0 scoring thresholds and ≤ four metrics receive positive scores
Some Stress: One or two metrics receive a score of -1 and the rest (at least five) receive positive scores or scores of 0; OR One metric receives a score of -2, another receives a score of -1, and the rest receive scores of 0 or higher
Stressed: Three or more metrics receive scores of -1 or -2; OR At least one metric receives a score of -3, and no more than three metrics receive negative scores
High Stress: At least one metric receives a score of -3, and at least four other metrics receive negative scores

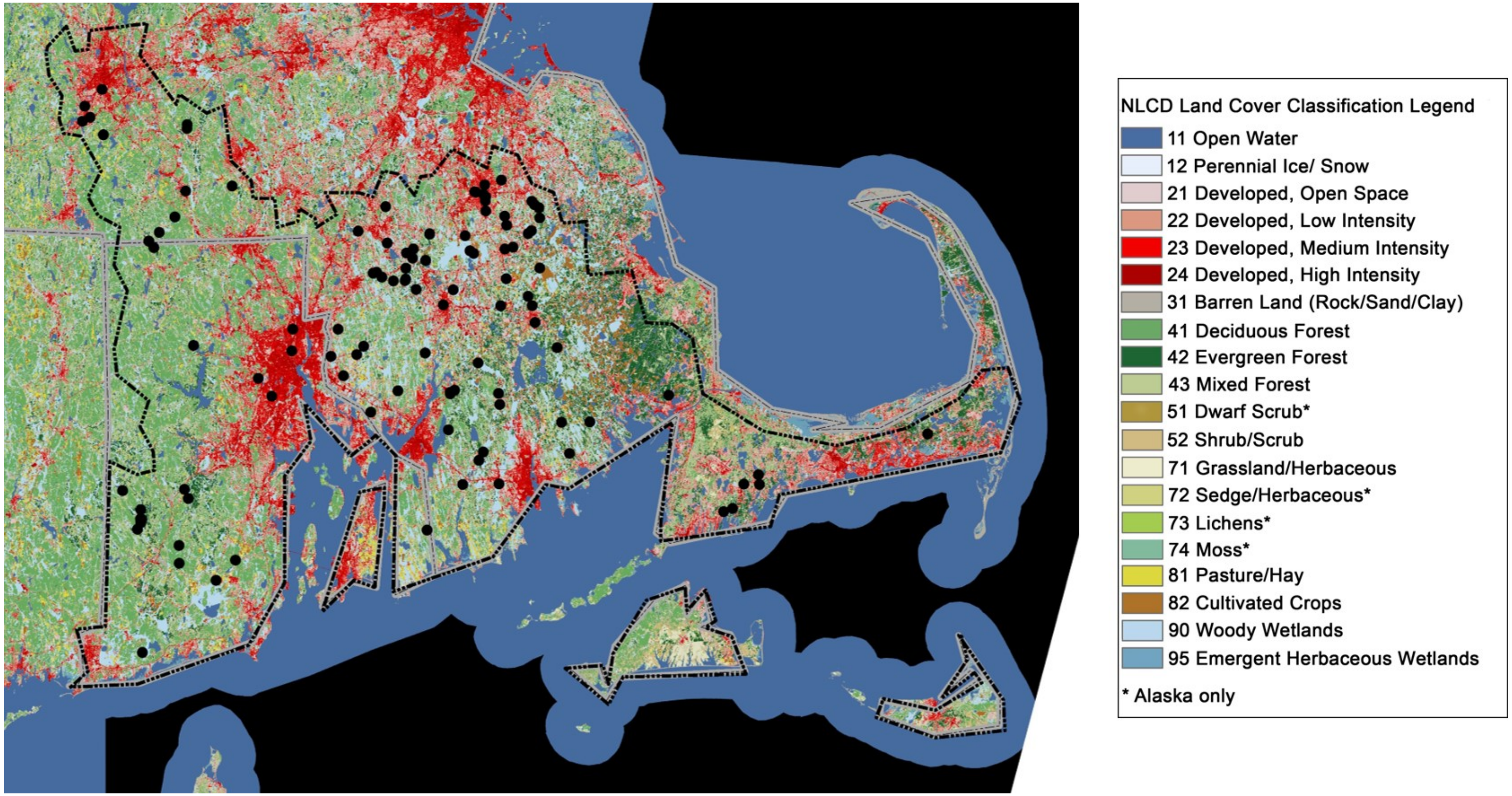


Figure 6. Several urban areas, including Providence, RI, are located in the U.S. EPA Southern New England program (SNEP) region, as well as agricultural areas (including cranberry bogs), forest and wetlands (source: NLCD 2016). Sample sites are shown as black dots.

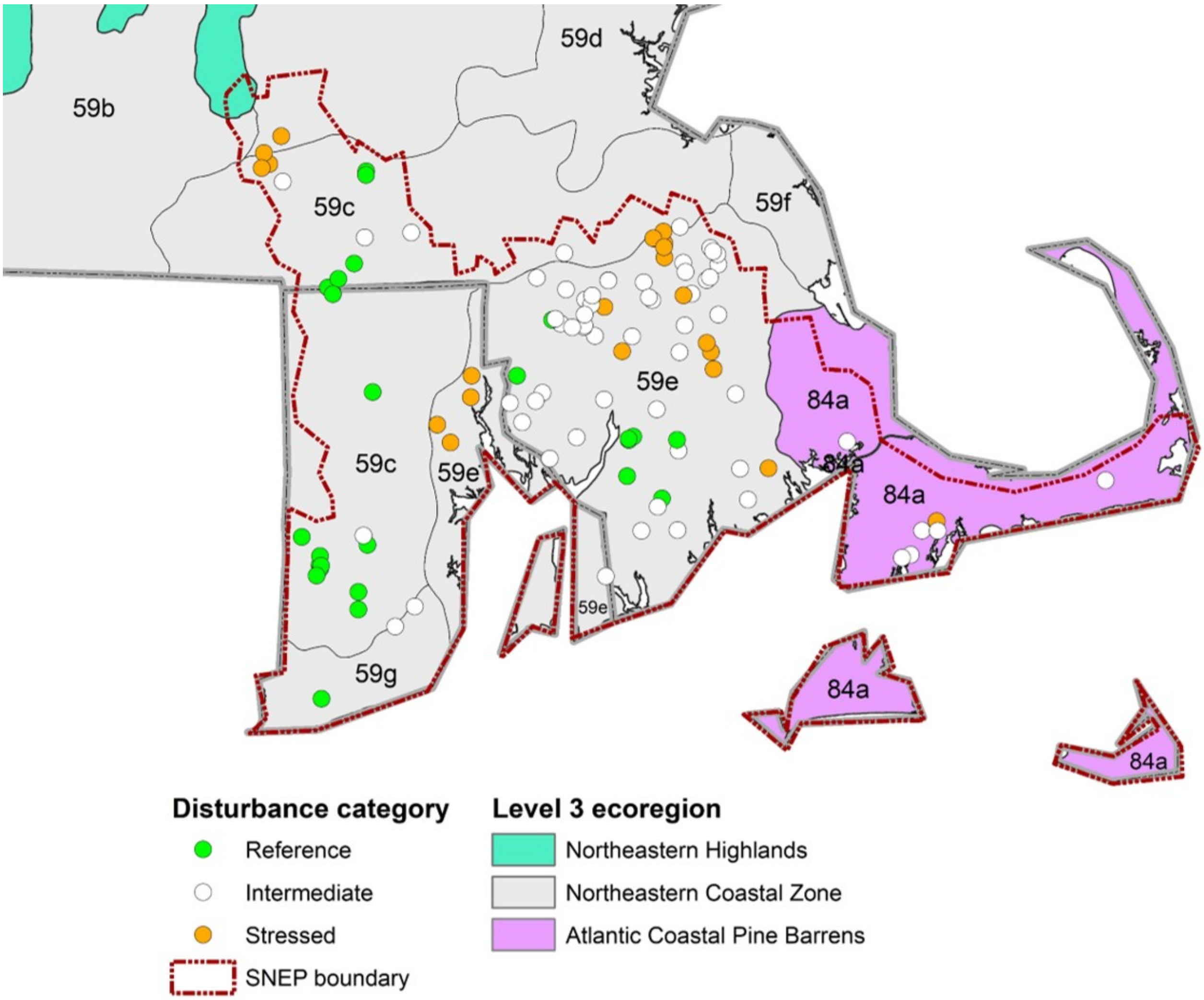


Figure 7. Spatial distribution of SNEP sites color-coded by disturbance category and overlaid on Level 3 and 4 ecoregions.

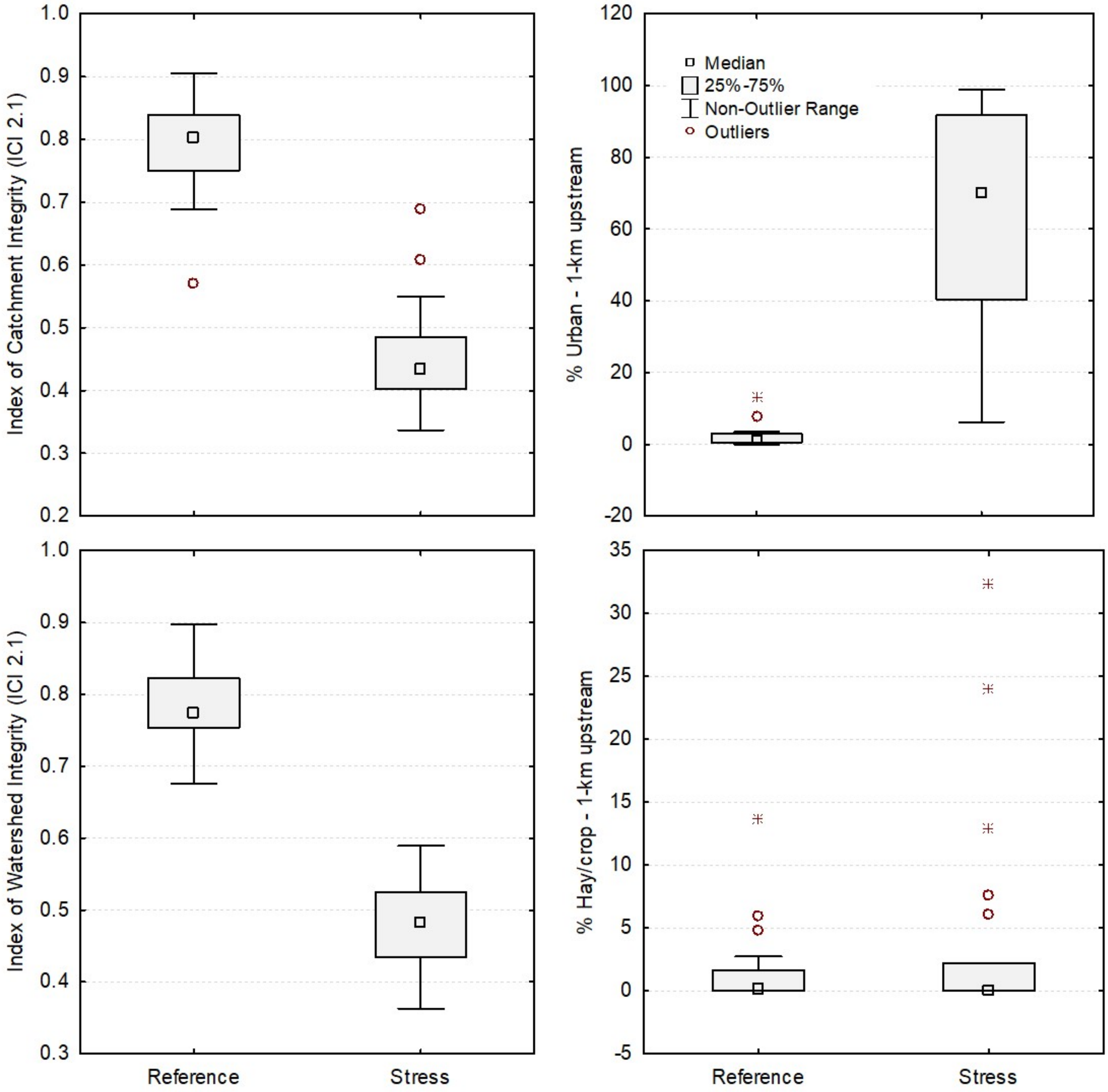


Figure 8. Box plots showing the range of disturbance represented in the reference (n=26) and stressed (n=23) sites, as measured by the ICI, IWI, percent urban, and percent agricultural land cover.

4 Classification

Site classification addresses the recognition that even with the least disturbance to streams, there might be different expectations of the sampled benthic assemblage due to natural effects and influences. Natural variation in stream slope, stream size, dominant substrates, temperature, and other factors are components of ecoregional characteristics that might cause a sample to contain more or less of certain taxa groups, sensitive taxa, or functionally specialized taxa. These types of taxa and some of the metrics derived from their traits are expected to exhibit variation not only with natural variation but also with human disturbance and unnatural stressors. When we use the benthic assemblage to indicate biological conditions relative to disturbance, we attempt to account for different expectations due to the background natural setting.

Accounting for different biological expectations was explored through an investigation of natural variation in samples from the least-disturbed reference sites. If the variation in taxa or metrics can be associated with natural categories or gradients, then those categories or gradients can be used to characterize different reference conditions. Comparisons of metrics between reference sites and those with high disturbance will be more sensitive to stressors if the natural variation is filtered out through site classification.

Site classification was expected to result in no classes or at most two classes. The low-gradient characteristics of the sites define the overall class in this data set. Only two discrete site classes could possibly be recognized before the separate classes became too small to robustly represent the reference condition in each class or to allow comparisons between reference and disturbed data within each class. The results of the classification exploration are summarized here because there was evidence of natural influences on the taxonomic composition. However, the details of the analysis are only included in an appendix because the ultimate decision was to address all low-gradient streams as a single category with no further site classification (Appendix E). General characteristics of the reference and highly stressed site groups and in all sites are shown in Table 7.

Table 7. Minimum and maximum values for selected characteristics of reference (Ref) and highly stressed (Strs) site groups and in all sites (All).

Variable	Ref Min	Ref Max	Strs Min	Strs Max	All Min	All Max
Drainage area (km^2)	3.1	91.1	1.8	175.2	1.7	188.8
Stream slope, 500m	0.00	2.94	0.03	1.76	0.00	2.94
% wetland/open water	8.5	34.3	0.2	34.2	0.2	44.4
Elevation (ft)	25	159	12	185	7	188
IWI	0.68	0.90	0.36	0.56	0.36	0.90
ICI	0.57	0.91	0.34	0.55	0.33	0.91
% urban	0.76	5.70	6.32	98.9	0.8	98.9
Road density	1.5	4.0	2.8	19.6	1.4	19.6

4.1 Exploratory Classification Analysis

The classification investigation proceeded through ordination of taxa and metrics in reference sites so that samples could be organized by similar biological characteristics. Non-metric multidimensional scaling (NMS) ordination was used to find sites with similar taxa. Principle components analysis (PCA) was used to organize sites by similar metric values, using 45 selected metrics. In each of these ordinations, the biological gradients were mapped in two dimensions, with each axis describing orthogonal composite aspects of the community. Any strong associations of environmental factors with the axes prompted further investigation of the factors as possible classification variables.

Level 4 ecoregions were fairly distinct for reference SNEP sites using presence/absence ordinations. On the first axis of the NMS ordination, sinuosity, longitude, land slope, and substrate characteristics, and percent water and wetland cover in the watershed are the major correlated natural variables that might be useful for site classification. Drainage area was also correlated but might not be appropriate for classification. In more disturbed non-reference sites, watersheds were up to 189 km^2. If drainage area was used in site classification, the reference condition derived mostly from small sites (<25 km^2) might represent a natural condition that would not be applicable to large non-reference sites. Sinuosity was on the same axis as land slope and drainage area. These three variables are often related, as large catchments are generally in flatter valleys with low slopes and meandering streams.

Longitude is related to ecoregion and could be used as a continuous variable for classification whereas ecoregions could define categorical classes. However, there was no distinctive break-point or threshold along the longitudinal gradient and the categorical ecoregions would be better classification variables than longitude.

To explore the effects of environmental variables on metric distributions, a PCA was performed with 45 metrics that represented a variety of metric formulations and taxa characteristics. The PCA identified the same variables on the first axis as were identified in the NMS of taxa presence absence, though in a slightly different order of importance. These included sinuosity, land slope, percent water and wetland cover in the watershed, longitude, and drainage area. Substrate characteristics were also correlated, though not as strongly.

4.2 Classification Summary

Classification schemes related to Level 4 ecoregions and drainage area were considered but ruled out based on results from the NMS and PCA analyses. Level 4 ecoregion did not cluster distinctly in the PCA ordination of metrics. Moreover, defining site classes based on Level 4 ecoregions might be untenable because it would result in small sample sizes for index calibration. All the reference sites in the NBL were <15 km2, which is smaller than the bulk of stressed sites, suggesting that a classification scheme based on drainage area or ecoregion would result in insufficient comparable samples for index calibration.

Continuous variables that showed potential for classification included: annual air temperature (PRISM 1981-2010), sinuosity, longitude, land slope, substrate types, and drainage area. Because there are no clear break-points to distinguish classes based on the continuous variables, scores for individual metrics that showed strong correlations with these natural variables were adjusted during index development (see Section 5.1).

5 Index Development

During the calibration of the SNEP low gradient IBI, a parallel project (statewide MassDEP low gradient IBI development) was also underway. Several members of the SNEP workgroup were also members of the MassDEP workgroup. There was also overlap across the two datasets (the SNEP samples were included in the statewide MassDEP IBI dataset). Thus, the two projects were not completely independent and often were informing one another, as described in the ensuing sections.

Index development consisted of the following steps:

- Metric scoring
- Metric selection
- Index compilations and performance evaluation
- Selection of final IBI
- Index verification

5.1 Metric scoring

Evaluation and selection of metrics typically involve testing of many more metrics than end up in the final index. We calculated and evaluated over 150 metrics (Appendix B). Formulae were applied to the metrics to standardize them to a 100-point scoring scale (as in Hughes et al. 1998, and Barbour et al. 1999). The scoring scale was based on the percentile statistics (and minimum values) of metric values across all sites (as opposed to only reference sites). 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 (zero) was the lower end:

$$Decreaser\ metric\ score = 100 * \frac{Metric\ value - minimum\ possible\ value}{95th\ percentile - minimum\ possible\ value}$$

For metrics that increased with increasing stress (referred to as 'increasers'; an example is the number of tolerant taxa metric), we used the following equation in which the 95th percentile was the upper end of the scoring scale and the 5th percentile was the lower end:

$$Increaser\ metric\ score = \ 100 * \frac{95th\ percentile - metric\ value}{95th\ percentile - 5th\ percentile}$$

A metric adjustment procedure was implemented for metrics that were strongly correlated with the classification variables (drainage area, mean annual air temperature (PRISM 1981-2010), longitude, percent wetland and open water in the watershed, mean land slope in the watershed). The procedure included the following steps:

1. Run a Spearman correlation analysis on all metrics and classification variables
    a. Include all reference samples
2. Identify metrics that were correlated at |r| > 0.50.

    a. At this level of correlation, the variable seems to be affecting the reference metric values
3. Identify variables that are correlated with more than one metric
    a. Variables that are consistently correlated are likely to have robust effects
4. Plot the 95th quantile regression line for all reference sites
    a. Included non-reference sites as points on the plots, though they do not drive the quantile regression
5. Identify plateaus in the relationships so the effective adjustment range is limited
    a. Extrapolation beyond the effective range might result in unreasonable metric expectations
    b. Define the plateau subjectively
6. Define the optimal end of the metric scoring range as the 95th quantile regression line and the plateaus intersecting that regression line
7. Score metrics on a 0-100 scale, interpolating between 0 and the optimal scoring range, based on the observed metric value and adjustment variable value

An example of an adjustment is shown in Figure 9. The number of taxa was higher in reference sites in larger drainage areas than smaller drainage areas (r = 0.61). The optimal number of taxa greater than 10 km^2 ($\log_{10}$ = 1.0) was about 65 taxa. For drainage areas smaller than 10 km^2, the optimal number of taxa is defined by the 95th quantile line and the actual drainage area of the site. A site with a drainage area of 6.0 km^2 would be expected to have about 52 taxa and the actual expectation would be calculated from the regression equation. Metric adjustments were made by converting metric values to metric scores on a 100-point scale, using the optimal metric value as the top of the scale (100), and interpolating down to 0. For example, a site with a drainage area of 6.0 km^2, expected to have 52 taxa, but truly having 48 taxa would have a score of 100 * 48/52 = 92.3.

The complexity of adjustment was also considered. If a metric showed a high correlation coefficient with a classification variable, then using the unadjusted metric might cause bias in evaluation and the unadjusted metric should not be used. If a similar responsive metric was available, but it did not require adjustment, then that similar metric might be a better choice for inclusion in the index. Those adjustments were applied and tested. However, if metrics based on relative richness (percent of taxa) did not require adjustment and performed as well as the adjusted metric, then the relative richness metric should be selected.

Seventeen of the biological metrics were adjusted to one or more classification variables. However, in the end, only the drainage area adjustment for the number of total taxa metric was considered in index development. All other adjusted metrics had similar performance to their non-adjusted equivalents (based on DE and z-score, as described in Section 5.2). Therefore, the non-adjusted metric versions were favored as they were conceptually easier to calculate and communicate.

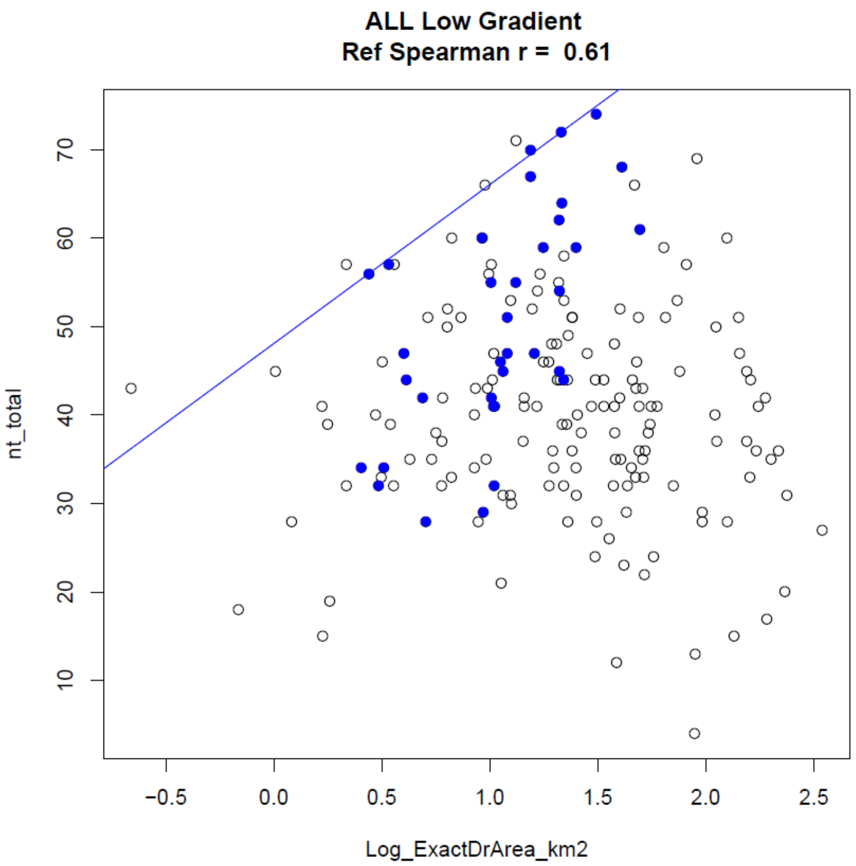


Figure 9. Bi-plot of total taxa (nt_total) and the log(10) transformation of drainage area, showing reference sites as solid blue markers, non-reference sites as open circles, and the reference 95th quantile regression line as a blue sloping line. Subjective limits to the regression adjustment were applied below 0.5 km^2 and above 1.5 km^2.

5.2 Metric selection

Metrics were 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 (richness, composition, evenness, tolerance, functional attribute, habit, thermal preference, and life cycle)
- Precision

The discrimination efficiency (DE) and *Z*-score were the primary performance statistics used to determine metric sensitivity. 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 10). Higher DE denotes a 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 the index. However, metric selection was usually dependent on relative DE values within a metric category.

The *Z*-score was calculated as the difference between mean 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 summarize the difference in distributions at critical potential threshold levels and incorporate the precision of the reference distribution. They were used in favor of a t-test or signal to 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 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.

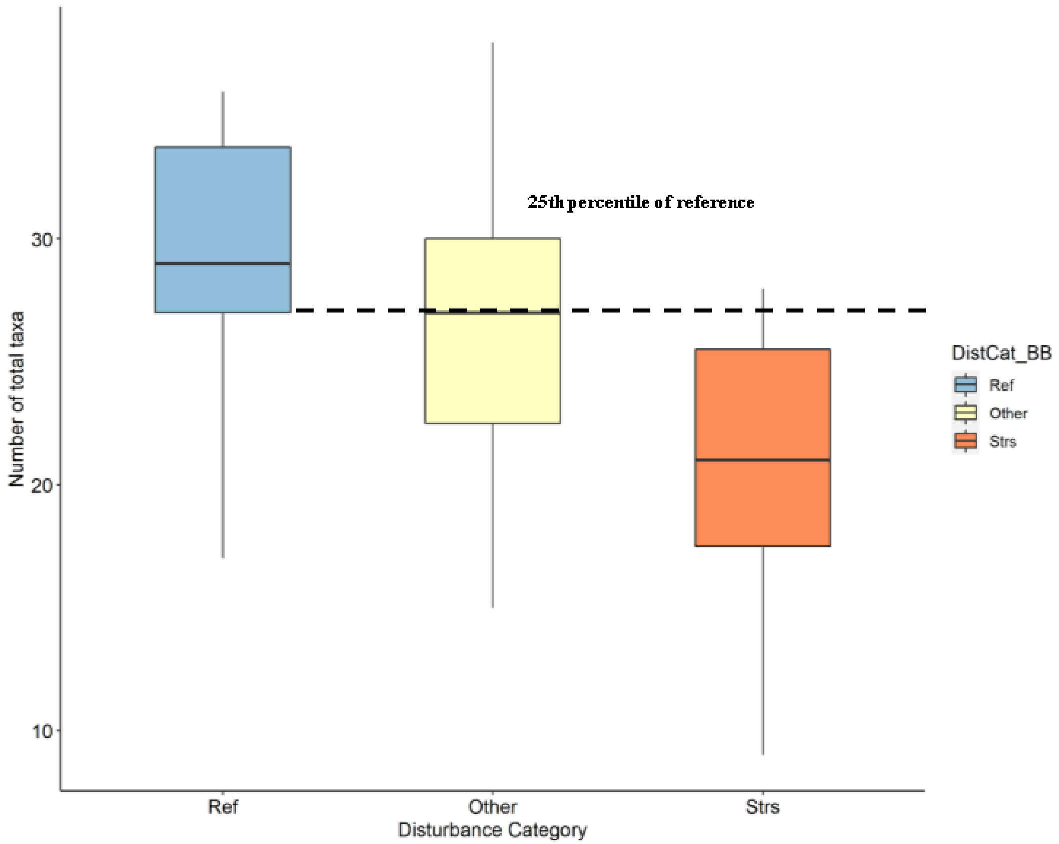


*Figure 10. Discrimination efficiency (DE). In this example, which uses the total number of taxa (a metric that decreases with stress), the 25th percentile of the reference distribution is used as the standard (and we calculate what percent of stressed sites were below that threshold; for example, if 15 out of 20 stressed sites have # total taxa metric values below the threshold (in this case, 27), the DE would equal 75%; if metric values for all 20 of the stressed sites were < 27, the DE would equal 100%). If it were a metric that increased with stress, we would have used the 75th percentile of the reference distribution as the standard (and calculated what percent 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 reference 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).*

Table 8 contains a list of the metrics that had the best performance (with high DE and *Z*-scores) within each metric category and were selected to be tested in the index compilations. The list of candidate metrics was further culled by identifying redundant metrics (metrics that represent similar taxa or traits) and removing the poorer performing metrics. Finally, the remaining metrics and those being considered in the SNEP IBI project were favored since having the same IBI for both projects would simplify application across the region. In the MA/SNEP dataset, the best performing metrics had DE of 100%. Each metric category was represented by at least one metric with DE > 50%. Spearman correlation analyses were performed on all pairwise combinations of candidate metrics (Table 9). Metric pairs with Spearman |r| ≥ 0.85 were considered redundant and were not both used in any index alternative. Metrics correlated at Spearman |r| ≥ 0.75 were evaluated for possible exclusion.

Table 8. Candidate metrics considered for inclusion in index development. 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 code	Metric description	In MA Project	Category	Trend	5th	95th	Scoring formula	*Z*-score	DE
pi_EPT	percent indivs - Orders Ephemeroptera, Plecoptera and Trichoptera	YES	COMP	Dec.	1.3	41.0	100*(metric)/41	1.46	82.6
pi_OET	percent indivs - Orders Odonata, Ephemeroptera, and Trichoptera	YES	COMP	Dec.	3.8	41.4	100*(metric)/41.4	1.28	78.3
pi_NonIns	percent indivs - Class not Insecta	YES	COMP	Inc.	10.7	79.8	100*(79.8-metric)/69.2	-1.41	78.3
nt_ffg_pred	number taxa - Functional Feeding Group - predator	YES	FFG	Dec.	4.0	13.8	100*(metric)/13.8	1.36	69.6
pt_ffg_col	percent taxa - Functional Feeding Group - collector-gatherer	NO	FFG	Inc.	32.6	48.3	100*(48.3-metric)/15.7	-1.17	69.6
pi_habit_swim	percent indivs - Habit - swimmers	YES	HABIT	Dec.	0.0	10.4	100*(metric)/10.4	0.74	87.0
pt_habit_climb	percent taxa - Habit - climbers	YES	HABIT	Inc.	6.3	22.7	100*(22.7-metric)/16.4	-1.59	69.6
nt_EPT	number taxa - Orders Ephemeroptera, Plecoptera, and Trichoptera	NO	RICH	Dec.	2.1	10.8	100*(metric)/10.8	1.56	91.3
nt_POET	number taxa - Orders Plecoptera, Odonata, Ephemeroptera, and Trichoptera	YES	RICH	Dec.	3.0	12.0	100*(metric)/12	1.43	82.6
pt_EPT	percent taxa - Orders Ephemeroptera, Plecoptera, and Trichoptera	YES	RICH	Dec.	5.5	26.3	100*(metric)/26.3	1.66	87.0
nt_CruMol	number taxa - Phylum Mollusca and SubPhylum Crustacea	NO	RICH	Inc.	5.0	9.9	100*(9.9-metric)/4.9	-1.90	95.7
pt_Amph	percent taxa - Order Amphipoda	NO	RICH	Inc.	2.0	9.1	100*(9.1-metric)/7.1	-2.24	91.3
pt_NonIns	percent taxa - not Class Insecta	YES	RICH	Inc.	24.3	46.3	100*(46.3-metric)/21.9	-2.21	95.7
pt_tv_intol	percent taxa - tolerance value - intolerant	YES	TOL	Dec.	0.0	0.0	100*(metric)/0	1.58	100.0
x_Becks	Becks Biotic Index	NO	TOL	Dec.	0.0	0.0	100*(metric)/0	1.34	100.0
pt_tv_toler	percent taxa - tolerance value - tolerant	YES	TOL	Inc.	11.6	28.3	100*(28.3-metric)/16.6	-2.72	100.0
x_HBI	Hilsenhoff Biotic Index	YES	TOL	Inc.	5.2	6.8	100*(6.8-metric)/1.6	-2.13	95.7
pt_volt_semi	percent taxa - semivoltine	YES	VOLT	Dec.	0.0	5.5	100*(metric)/5.5	1.21	87.0
pt_volt_multi	percent taxa - multivoltine	YES	VOLT	Inc.	11.7	30.3	100*(30.3-metric)/18.6	-1.39	69.6

In MA project: indicates that the same metric was under consideration in the MA multihabitat IBI development project; **Trend:** Decreasing (Dec.) or increasing (Inc.) trend with increasing stress; **5th**: 5th percentile of all sample metrics in the site class; **95th**: 95th percentile of all sample metrics in the site class; **Scoring Formula:** Replace "metric" with the sample metric value for calculation of an index; **DE:** Discrimination Efficiency.

Table 9. Spearman rho correlation among candidate metrics. Coefficients ≥ 0.80 are emphasized with bold type.

		Metric #																		
#	Metric	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	nt_CruMol	1																		
2	nt_EPT	-0.39	1																	
3	nt_ffg_pred	-0.07	0.46	1																
4	nt_POET	-0.33	**0.97**	0.57	1															
5	pi_EPT	-0.31	0.77	0.25	0.71	1														
6	pi_habit_swim	-0.14	0.52	0.48	0.53	0.44	1													
7	pi_NonIns	0.50	-0.60	-0.29	-0.57	-0.53	-0.24	1												
8	pi_OET	-0.20	0.72	0.30	0.69	**0.96**	0.45	-0.49	1											
9	pt_Amph	0.52	-0.58	-0.43	-0.54	-0.48	-0.25	0.50	-0.39	1										
10	pt_EPT	-0.47	**0.91**	0.22	**0.85**	0.78	0.41	-0.53	0.72	-0.49	1									
11	pt_ffg_col	0.27	-0.41	-0.35	-0.40	-0.39	-0.28	0.29	-0.39	0.36	-0.45	1								
12	pt_habit_climb	0.64	-0.26	0.02	-0.21	-0.26	-0.08	0.23	-0.16	0.20	-0.33	0.05	1							
13	pt_NonIns	0.67	-0.74	-0.34	-0.72	-0.51	-0.39	0.69	-0.44	0.51	-0.72	0.41	0.46	1						
14	pt_tv_intol	-0.49	0.59	0.47	0.59	0.37	0.34	-0.54	0.31	-0.57	0.49	-0.25	-0.34	-0.59	1					
15	pt_tv_toler	0.49	-0.78	-0.49	-0.77	-0.61	-0.47	0.59	-0.57	0.61	-0.74	0.45	0.34	**0.80**	-0.66	1				
16	pt_volt_multi	0.21	-0.50	-0.53	-0.53	-0.43	-0.35	0.27	-0.43	0.45	-0.41	0.21	0.15	0.28	-0.51	0.52	1			
17	pt_volt_semi	-0.46	0.61	0.33	0.62	0.45	0.18	-0.58	0.43	-0.48	0.61	-0.41	-0.34	-0.64	0.56	-0.66	-0.31	1		
18	x_Becks	-0.46	0.62	0.52	0.62	0.38	0.38	-0.54	0.32	-0.57	0.48	-0.24	-0.33	-0.59	0.99	-0.66	-0.51	0.54	1	
19	x_HBI	0.44	-0.66	-0.43	-0.65	-0.59	-0.30	**0.86**	-0.57	0.64	-0.58	0.35	0.21	0.65	-0.62	0.71	0.38	-0.66	-0.62	1

5.3 Index compilation and performance

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 index alternative was then evaluated for discrimination efficiency and other measures of representativeness and sensitivity. Index formulations were created and evaluated in two ways: automatic all-subsets modeling and manual metric substitutions.

The all-subsets analysis allowed consideration of a plethora of diverse index compositions that simply could not be computed by hand. Nineteen candidate metrics were selected for inclusion in index trials based on DE, *Z*-score, and professional opinion of the working group. An “all subsets” routine in R software (R Core Team 2020) 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 8 were included in the all-subsets analysis. The all-subsets model calculation and screening resulted in thousands of valid index combinations. Initially, the all-subsets analysis resulted in approximately 103,000 different index combinations. To identify the most sensitive, comprehensive, and practical index alternatives, the characteristics of the alternatives were screened for favorable characteristics such as high DEs and representation of multiple metric categories. Metrics with conceptual redundancy and unexplained response mechanisms were excluded. Habit metrics were not preferred because they did not have plainly understandable response mechanisms. To narrow down the long list of index alternatives, two reviewers (Ben Block (Tetra Tech) and James Meek (MassDEP)) were provided an Excel worksheet with results from the all-subsets analysis. The number of index alternatives was reduced to approximately twenty. The screening and exclusion criteria are summarized in Table 10. The resulting subset of index alternatives had similar performance statistics (Table 11), therefore, the final selection process involved subjective decisions on metric preference and performance.

The workgroup decided to pick indices with familiar metrics (composition, functional feeding group (FFG), richness, tolerance, and voltinism). Voltinism metrics were emphasized because they indicate ecosystem stability. Multivoltine taxa are short lived and have multiple generations per year. The presence/abundance of these taxa indicate a system that can experience more variability (e.g., flow) and potentially more disturbance overall. Semivoltine taxa require more than one year to complete their life cycle and thereby tend to require a more stable environment. The workgroup rationalized their choice based on empirical performance and ecological characteristics of the individual and combined metrics. They selected an index that was a top selection for both the MassDEP and SNEP projects. The final choice was Model 6_13784, which included six metrics (Tables 12 & 13).

Table 10. Reviewer screening and exclusion criteria for narrowing the list of index alternatives. Initially, the all-subsets model resulted in over 100,000 alternative index compositions.

Criteria #	Model Elimination Criteria (eliminated models with these criteria)	# Remaining models
1	Contains any Habit metrics	27388
2	Insect/Non-Insect Metrics > 1	22892
3	Contains both pt_EPT and pi_EPT	20995
4	Contains both nt_EPT and pi_EPT	19095
5	Contains both pt_tv_toler and pt_tv_intol	15940
6	Contains both pt_volt_semi and pt_volt_multi	11989
7	Contains no FFG metrics	8990
8	Contains no Tolerance metrics	7910
9	Number of Metrics < 5 OR > 7	5483
10	DE < 100	5032
11	*Z*-Score > -2.25	3740
12	Ref. q25 – Str. q75 < 18	3358
13	Ref. cv > 0.22	1366
14	Contains x_Beck and x_HBI	1224
15	Ref. q10 – Str. q90 < 3	861
16	Contains both nt_CruMol and pt_Amph	600
17	Contains both nt_ffg_pred and pt_ffg_col	425
18	Number of metric categories < 4	369
19	Number of Richness metrics > 2	249
20	Only Richness metrics are nt_CruMol, pt_Amph, or pt_NonIns	130
21	Contains no Composition metrics	95
22	Only Composition metric is pi_NonIns	84
23	Contains nt_CruMol or pt_Amph metrics	33

Evaluation of subsample size

After Model 6_13784 was selected, we performed an additional analysis on the full dataset to evaluate how much the IBI was affected by subsample size since some regional partners may lack sufficient resources to process 300-organisms (instead they may be limited to 200 or 100-count samples). Of particular interest was the effect on the two richness metrics (number of Plecoptera, Odonata, Ephemeroptera, Trichoptera (POET) taxa and number of predator taxa), since the number of taxa found in samples generally decreases with a decrease in the number of individuals collected (Gotelli and Graves 1996). With this consideration in mind, the working group wanted to explore: 1) the magnitude that subsample size affected the two richness metrics vs. the percent taxa versions of those metrics; and 2) if the percent taxa POET and predator metrics were substituted into IBI model 6_13784, did the alternative IBI perform equally well or better (as measured by DE, Z-score, and coefficient of variation (cv)] when using 300, 200, or 100-count samples). Ideally, the working group wanted to select an IBI that not only performed well in both the SNEP and MA/SNEP datasets but also performed well in 100, 200, and 300-count samples. For clarity's sake, we refer to Model 6_13784 as the ‘NumTaxaIBI’ and the alternative model, which contains the percent taxa metric equivalents, as the ‘PctTaxaIBI’ (Table 12).

The analyses showed the PctTaxaIBI to have similar performance as the NumTaxaIBI (DEs of 97.6 vs. 100, respectively, accounted for by one sample) (Table 13). There were, however, differences in metric scoring formulae. With the PctTaxaIBI, the same metric scoring formulae could be used in 100-, 200-, and 300-count samples in both the MA/SNEP and SNEP datasets, whereas the scoring formulae for the two richness metrics in the NumTaxaIBI would need to be adjusted based on subsample size (Block et al. 2020). Thus, although the NumTaxaIBI (Model 6_13784) was initially selected by the working group through the all-subsets model routine, the PctTaxaIBI alternative was decided upon as the final model in both projects to eliminate the need to adjust metric scoring formula and simplify the application of the IBI across the region. We do, however, recommend 300-count samples (or the highest subsample size resources permit) because those samples do perform better based on z-scores and cv statistics (Table 13) (Block et al. 2020).

Table 11. The nine best model alternatives (selected by the working group). Metrics used in each alternative are listed as "1". 0 = not included. The model initially chosen by the working group is highlighted in green (Model 6_13784). See Table 8 for metric descriptions.

Model ID	7_49898	6_18508	7_33461	7_31921	7_43415	6_15092	**6_13784**	7_38340	7_22450
nt_CruMol	0	0	0	0	0	0	**0**	0	0
nt_EPT	0	0	0	0	0	0	**0**	0	1
nt_POET	0	0	1	1	1	1	**1**	0	0
pt_Amph	0	0	0	0	0	0	**0**	0	0
pt_EPT	1	1	0	0	0	0	**0**	1	0
pt_NonIns	1	1	1	1	1	1	**1**	1	1
pi_habit_swim	0	0	0	0	0	0	**0**	0	0
pt_habit_climb	0	0	0	0	0	0	**0**	0	0
nt_ffg_pred	0	1	1	1	0	1	**1**	1	1
pt_ffg_col	1	0	0	0	1	0	**0**	0	0
pt_volt_multi	0	0	0	0	0	0	**0**	0	0
pt_volt_semi	1	1	1	1	1	1	**1**	1	1
pi_EPT	0	0	0	1	0	1	**0**	0	0
pi_NonIns	0	0	0	0	0	0	**0**	0	0
pi_OET	1	1	1	0	1	0	**1**	1	1
pt_tv_intol	0	0	0	0	0	0	**0**	0	0
pt_tv_toler	1	1	1	1	1	1	**1**	1	1
x_Becks	0	0	0	0	0	0	**0**	0	0
x_HBI	1	0	1	1	1	0	**0**	1	1
Str.DE	100	100	100	100	100	100	**100**	100	100
z	-2.51	-2.56	-2.50	-2.54	-2.45	-2.49	**-2.45**	-2.61	-2.52

*Table 12. Metric codes and names for the index selected by Mass DEP (6_13784). *Denotes the richness metrics that were affected by subsample size. The "alternative" index (PctTaxaIBI) replaces the two richness metrics with percent taxa versions of those metrics.*

Index	Metric Code	Metric Name
6_13784 (NumTaxaIBI)	*nt_POET	number taxa - Orders Plecoptera, Odonata, Ephemeroptera, and Trichoptera (POET)
	*nt_ffg_pred	number taxa - Functional Feeding Group (FFG) - predator (PR)
	pt_NonIns	percent (0-100) taxa - not Class Insecta
	pt_volt_semi	percent (0-100) taxa - semivoltine (SEMI)
	pi_OET	percent (0-100) individuals - Orders Odonata, Ephemeroptera, and Trichoptera
	pt_tv_toler	percent (0-100) tolerant taxa
Alternative (PctTaxaIBI)		
	pt_POET	percent (0-100) taxa - Orders Plecoptera, Odonata, Ephemeroptera, and Trichoptera (POET)
	pt_ffg_pred	percent (0-100) taxa - Functional Feeding Group (FFG) - predator (PR)

Table 13. Performance statistics for the two versions of the selected model (NumTaxaIBI vs. PctTaxaIBI). Coefficient of variation (CV) equals the ratio of the standard deviation to the mean, based on reference sites. Lower values are more desirable as they indicate less variability.

Dataset	NumTaxaIBI*			PctTaxaIBI		
	DE	*Z*-score	CV	DE	*Z*-score	CV
MA/SNEP 300-count	100.0	2.87	0.18	97.6	2.96	0.16
MA/SNEP 200-count	97.6	2.69	0.19	97.6	2.74	0.17
MA/SNEP 100-count	97.6	2.45	0.21	97.6	2.50	0.19
SNEP only, 300-count	100.0	2.45	0.21	95.65	2.72	0.18
SNEP only, 200-count	100.0	2.30	0.23	100.0	2.48	0.19
SNEP only, 100-count	100.0	2.22	0.23	100.0	2.40	0.20

* the scoring formulae for the two richness metrics in the NumTaxaIBI would need to be adjusted based on subsample size (Block et al. 2020) therefore PctTaxaIBI was ultimately selected

5.4 Final index selection and performance

The team of MassDEP and RI DEM biologists used the following empirical and logical criteria to select their final index:

- Relatively high index DE and *Z*-scores
- Index metrics representing as many metric categories as practical
- Not including redundant metrics
- Performs well at different subsample sizes (tested 100-, 200-, and 300-count versions)
- Inclusion of individual metrics having the following characteristics:
  - High overall DE
  - Response mechanisms that were plausible and ecologically important
  - Straightforward metric calculations

The component metrics in the SNEP low gradient, multihabitat IBI are listed in Table 14, along with performance statistics and scoring formulae. The metrics have comprehensible mechanisms of response to increasing environmental stress, as described in Appendix F. The percent tolerant taxa metric (pt_tv_toler) is strongly correlated with percent non-insect taxa (pt_NonIns) (rho=0.80), percent POET taxa (pt_POET) (rho=-0.75), and percent semi-voltine taxa (pt_volt_semi) (rho=-0.66) (Table 15); however, the workgroup did not think that these metric were fundamentally redundant with one another but instead evaluated unique components of the macroinvertebrate community. The IBI discriminates well between reference and stressed samples, as shown in Figure 11.

Index scores do not always match the disturbance categories. For example, a tributary of the Wading River east of Attleboro (TAU-W2910) is a reference sites with a low index score. This is a sub-reference site with a small watershed (5.0 km^2). There is no immediate explanation for the high percentages of non-insects and tolerant taxa in this sample, so it might take additional investigation to associate site conditions with the index score. On the Moshassuck River near Providence, there are two highly stressed sites with very different index scores. The upper site, LO-Worst-P1, has an unusually high IBI score of 67.9 and the lower site, LO-Worst-R1, has an index score of 32.6, as expected for a highly stressed site. Because of possible confusion of the contributing watershed (downstream of an impoundment of the Blackstone River Canal), it is possible that the watershed delineation was incorrect and that the upstream site with the better IBI score is actually only moderately stressed. In this case, the incongruent index score might indicate that the disturbance category was incorrect as the biology indicates.

Table 14. Metrics in the low gradient IBI, with scoring formulae, DE values, and trend. This index was chosen for both the SNEP and MassDEP low gradient projects.

Metric Name	Category	5th	95th	Scoring formula	DE	Trend
% OET individuals (pi_OET)	COMP	3	49	100*Metric/49	78.3	Dec.
% Predator taxa (pt_ffg_pred)	FFG	9	32	100*Metric/32	69.6	Dec.
% Non-insect taxa (pt_NonIns)	RICH	4	46	100*(46-Metric)/42	95.7	Inc.
% POET taxa (pt_POET)	RICH	9	40	100*Metric/40	78.3	Dec.
% Tolerant taxa (pt_tv_toler)	TOLER	3	36	100*(36-Metric)/33	100.0	Inc.
% Semivoltine taxa (pt_volt_semi)	VOLT	0	12	100*Metric/12	87.0	Dec.

5th: 5th percentile of all sample metrics; 95th: 95th percentile of all sample metrics
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 15. Correlation coefficients (Spearman rank rho) for the IBI input metrics, based on the SNEP dataset.

	pi_OET	pt_ffg_pred	pt_NonIns	pt_POET	pt_tv_toler	pt_volt_semi
pi_OET	1					
pt_ffg_pred	0.03	1				
pt_NonIns	-0.44	-0.09	1			
pt_POET	0.70	0.11	-0.73	1		
pt_tv_toler	-0.57	-0.24	0.80	-0.75	1	
pt_volt_semi	0.43	0.11	-0.64	0.64	-0.66	1

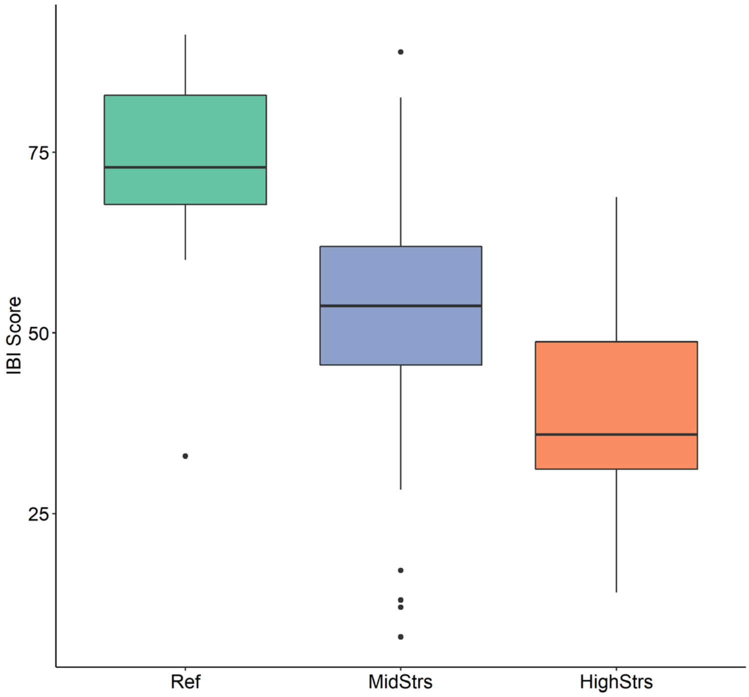


Figure 11. Distribution of SNEP IBI scores across disturbance categories, reference (Ref), intermediate (MidStrs), and stressed (HighStrs).

We also evaluated the relationship between IBI scores and four measures of disturbance (ICI, IWI, percent urban, and percent agriculture). IBI scores were positively correlated with the ICI (rho = 0.53) and IWI (rho = 0.61) and had a strong negative correlation with percent urban land cover (rho = -0.62) (Figure 12). IBI scores were weakly correlated with percent agriculture land cover (rho = 0.05) but most sites had low percent agriculture (<10%) (Figure 12).

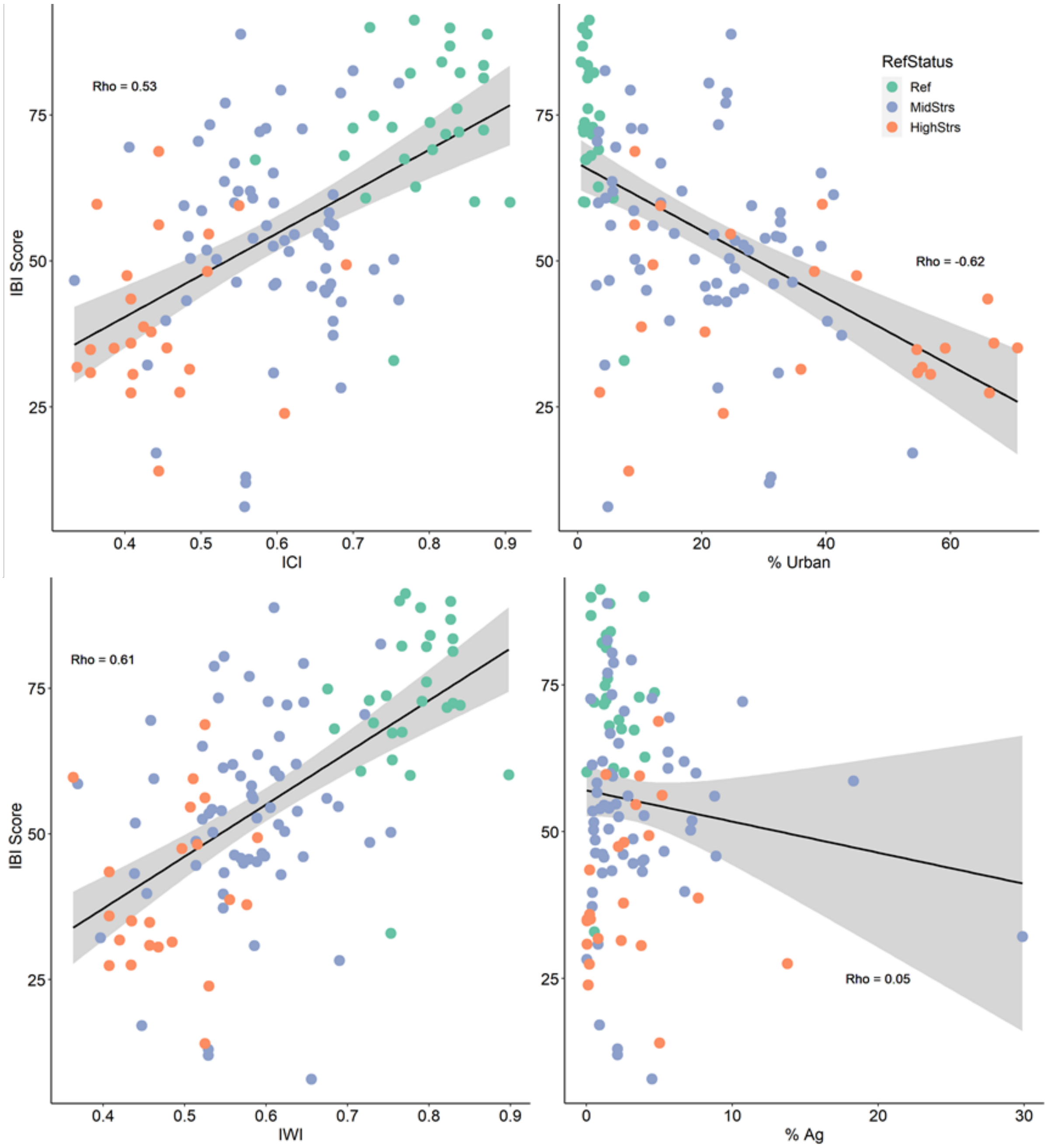


Figure 12. Relationship between the low gradient, multihabitat IBI vs. ICI (upper left), IWI (lower left), percent urban (upper right) and percent agriculture (lower right). The black line is the regression line and the rho value is the Spearman rank correlation coefficient.

5.5 Index verification

We had few sites to use in calibrating the index, so all were used in index calibration and none were reserved for independent application of the index and comparison to reference designations. Instead, index values were compared to stressors that were not used in defining the index calibration stressor gradient. Relationships with these independent indicators would show that the index was responsive along the stressor gradient, and it would be validated. The stressor variables that were compared included habitat scores, dissolved oxygen (DO), conductivity, and percent forest cover in the watershed. Other variables were compared, though they were not necessarily stressors in the low gradient streams. These included acidity (pH), substrate, and temperature.

When evaluated in relation to the RPB habitat score (maximum score = 189), maximum IBI scores declined as the habitat scores decreased from 120 (Figure 13). Not all IBI scores were high with better habitat scores. This suggests that other stressors might affect the macroinvertebrate community even when habitat conditions were fair or good. The individual habitat variables that went into the total habitat score show that some components of habitat were more influential on IBI scores than others. The most effective habitat components include available cover, sediment deposition, riparian vegetation, and bank stability (Figure 14). As with the total habitat score, these and other habitat variables only seem to affect the IBI scores when the values were low and IBI scores were variable with less habitat stress.

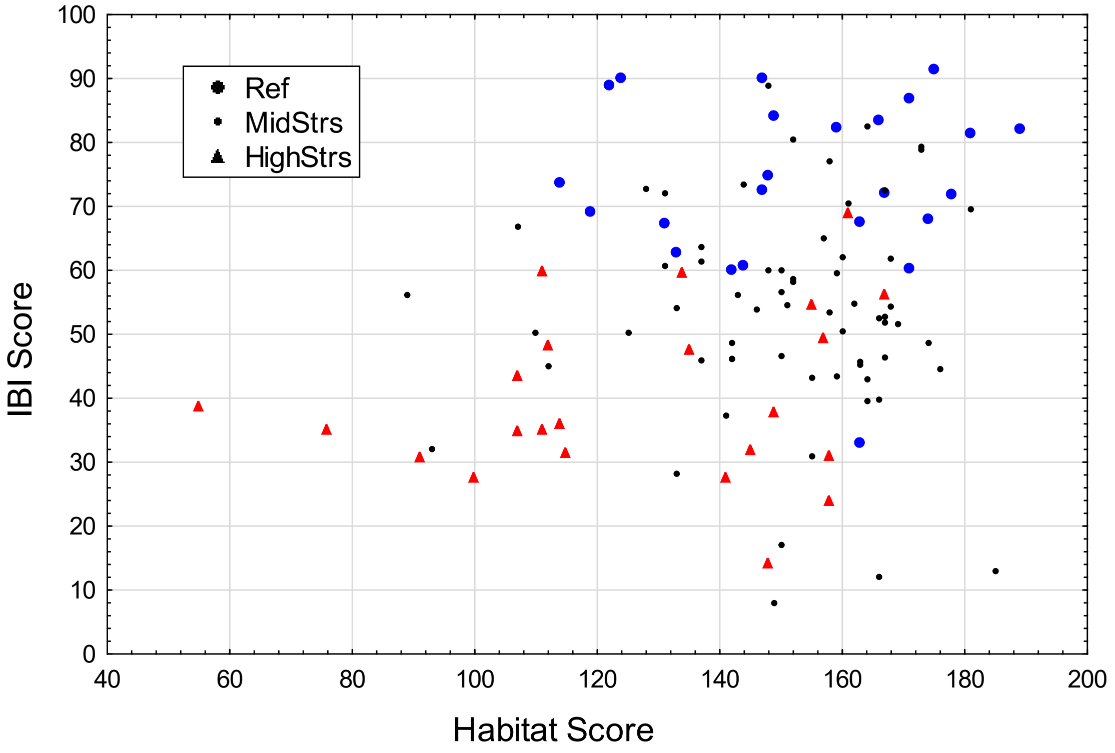


Figure 13. IBI scores in relation to RBP total habitat scores, marked by disturbance category.

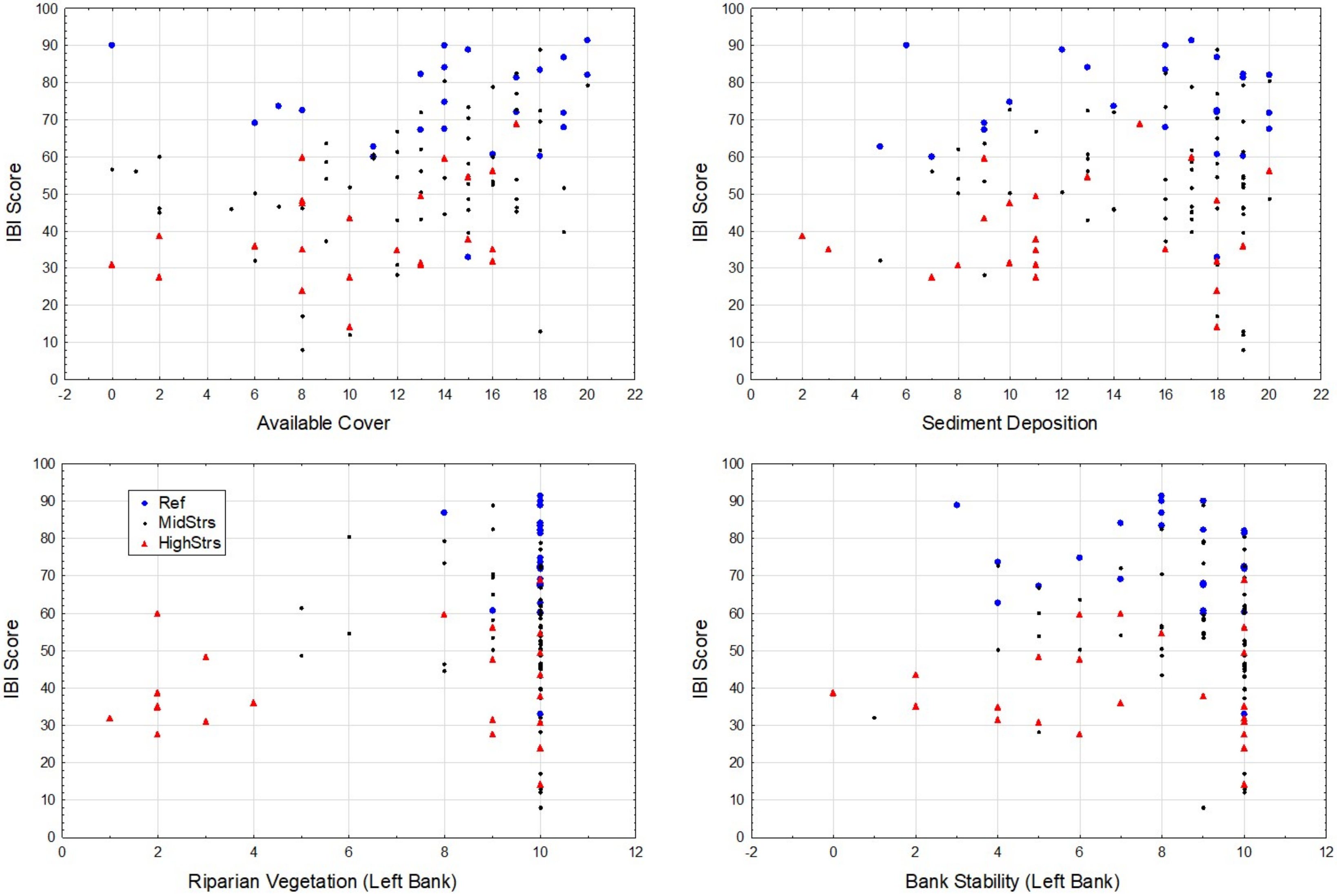


Figure 14. IBI scores in relation to effective habitat variables, including available cover, pool variability, riparian vegetation, and sediment deposition.

DO appears to affect IBI scores when concentrations are below 6 mg/L and above 14 mg/L (Figure 15). However, there were only eight sites that had DO at these extremes. The DO signal is also tenuous because the data are from grab samples taken at the time of the macroinvertebrate sampling and readings could fluctuate during the day depending on light intensity and temperature. However, the observed low DO might be associated with eutrophic conditions in which oxygen is stripped from the water due to excessive respiration by consumers and decomposers of the excessive algae. Very high DO might also be associated with algal productivity. Resulting high respiration can cause an extreme DO flux between night and day conditions. This flux was not confirmed for these examples.

The IBI shows a strong correlation with specific conductivity, especially as conductivity increases above 0.10 mS/cm (100 µS/cm) (Figure 16). Conductivity can be an indicator of general inputs of salts and other contaminants that could affect the macroinvertebrates. Greater inputs suggest more human activity in general and the relationship between the IBI and conductivity could be due to the multiple stressors associated with human activity (Burns et al. 2005, Hatt et al. 2004, Lussier et al. 2008).

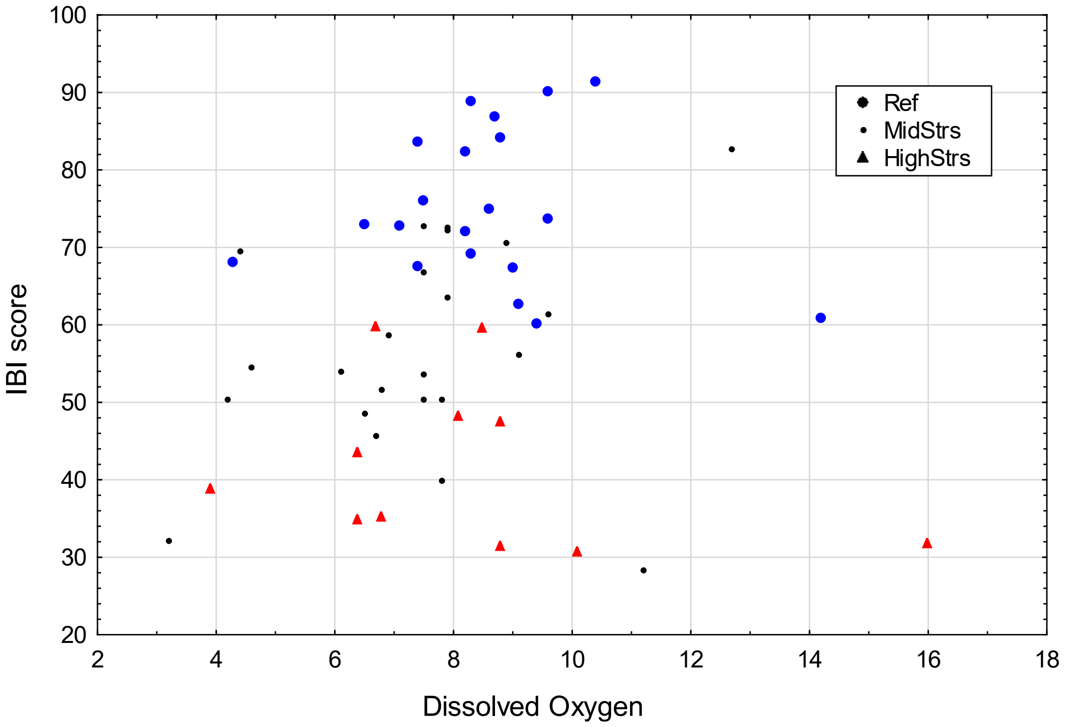


Figure 15. IBI scores in relation to dissolved oxygen (DO) in sites with DO data, marked by disturbance category.

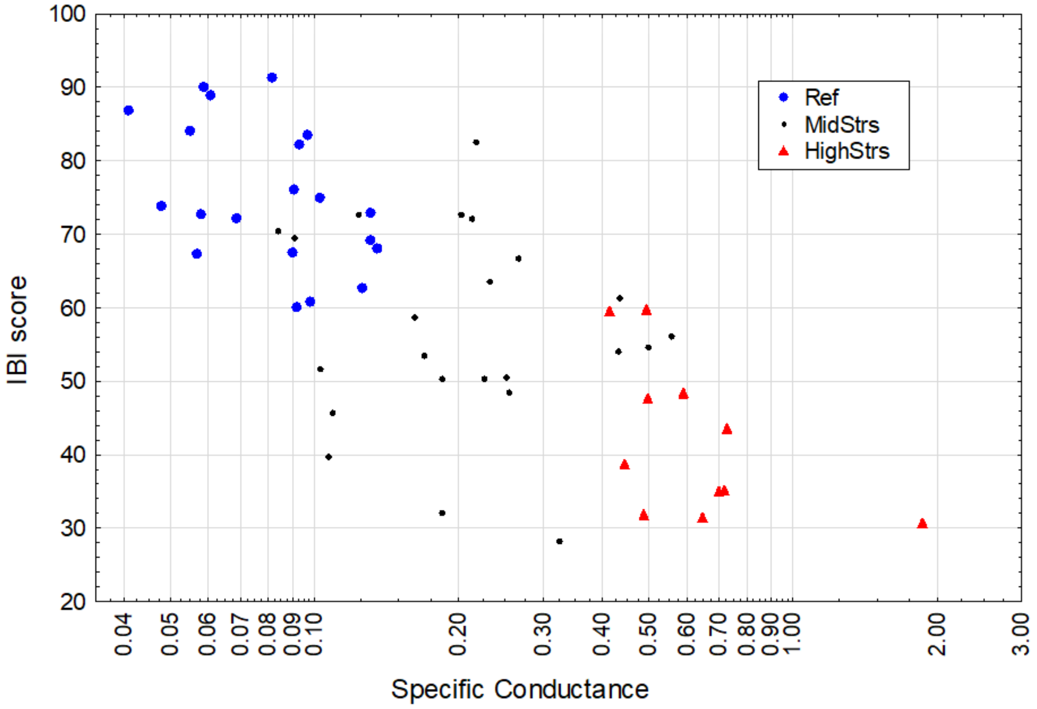


Figure 16. IBI scores in relation to conductivity (on a log-transformed axis) at sites with conductivity data, marked by disturbance category.

The IBI has higher values at sites with a greater percentage of forested land in the watershed (Figure 17). Forest cover is generally the complement of developed land cover, whether developed for urban or agricultural uses. Forest cover was not directly used as a criterion for the calibrated disturbance gradient, while urban and agricultural covers were.

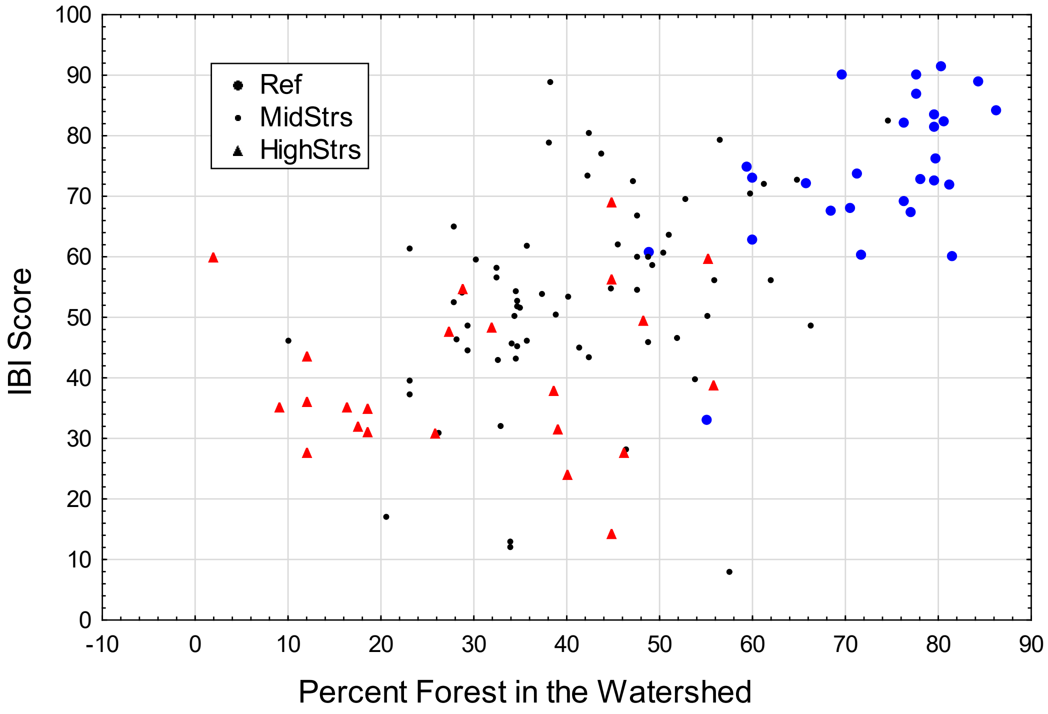


Figure 17. IBI scores in relation to percent forest cover (watershed-scale), marked by disturbance category.

Indications from habitat, DO, conductivity, and percent forest cover are that the IBI responds as expected to these stressor indicators and is validated. While the relationships between the IBI and habitat and DO are somewhat variable over the whole range of stressor intensity, the relationships show a limitation of biological potential with the most intensive stresses. The strongest IBI relationships are with conductivity and percent forest. Conductivity increases steeply with increasing urban land uses (Figure 18). The urban land uses were also considered in defining the disturbance categories for IBI calibration. This connection between land use, conductivity, and disturbance status might suggest an inevitable relationship between the IBI and conductivity. However, it also provides a mechanistic link between the source of stress (urban intensity) and the macroinvertebrate assemblage through inputs such as salts.

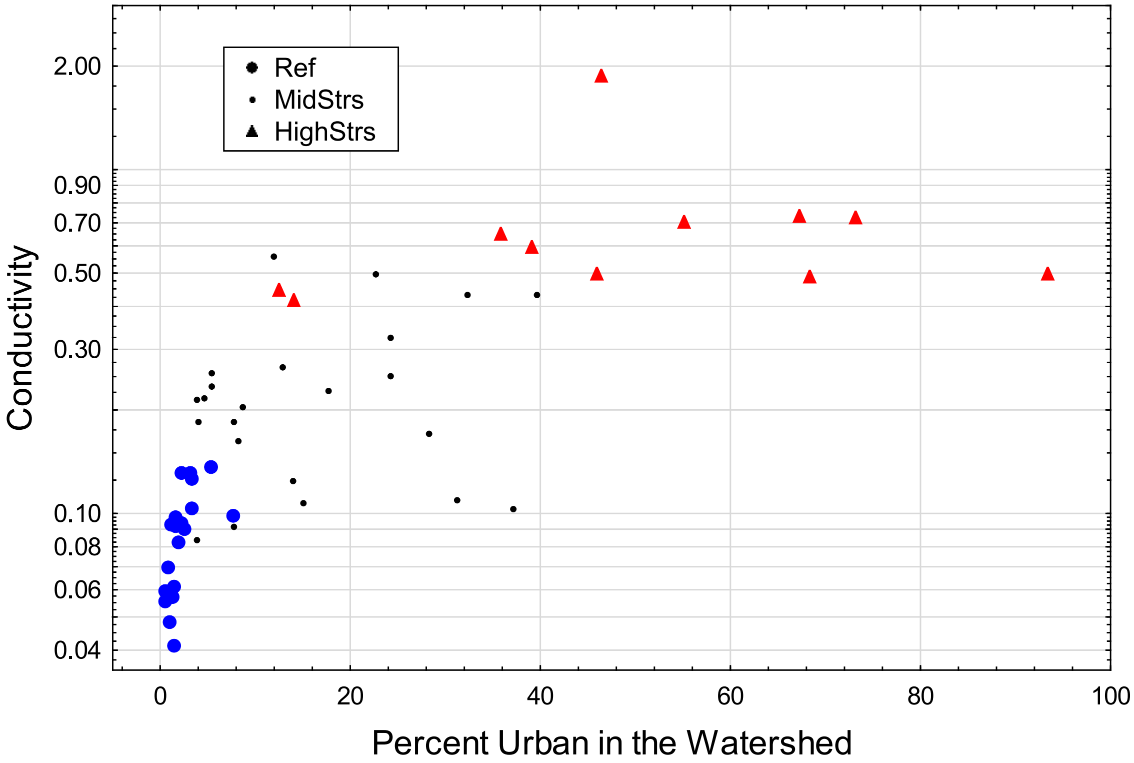


Figure 18. Conductivity (on a log-transformed axis) at sites with conductivity data in relation to percent urban land uses in the watershed, marked by disturbance category.

In this data set, there is a strong correlation between pH and conductivity, with low pH associated with low conductivity (Figure 19). The IBI is also associated with pH, showing better scores with low pH, even below 5.0 su. The reference streams used in calibrating the IBI all had pH < 6.5 su and conductivity < 0.30 mS/cm. These relationships suggest that the natural condition of the low gradient streams in the SNEP region are acidic. The natural setting includes greater canopy cover than in developed areas and therefore greater input of leaf litter as well as cooler temperatures (Figure 20). The soils apparently have low buffering capacity, as is seen in the neighboring pine barrens of Cape Cod. As conductivity increases with human activity, the salts provide buffering capacity and pH increases. Higher pH might not be a stressor, but it is certainly associated with higher conductivity and higher urban land use intensity.

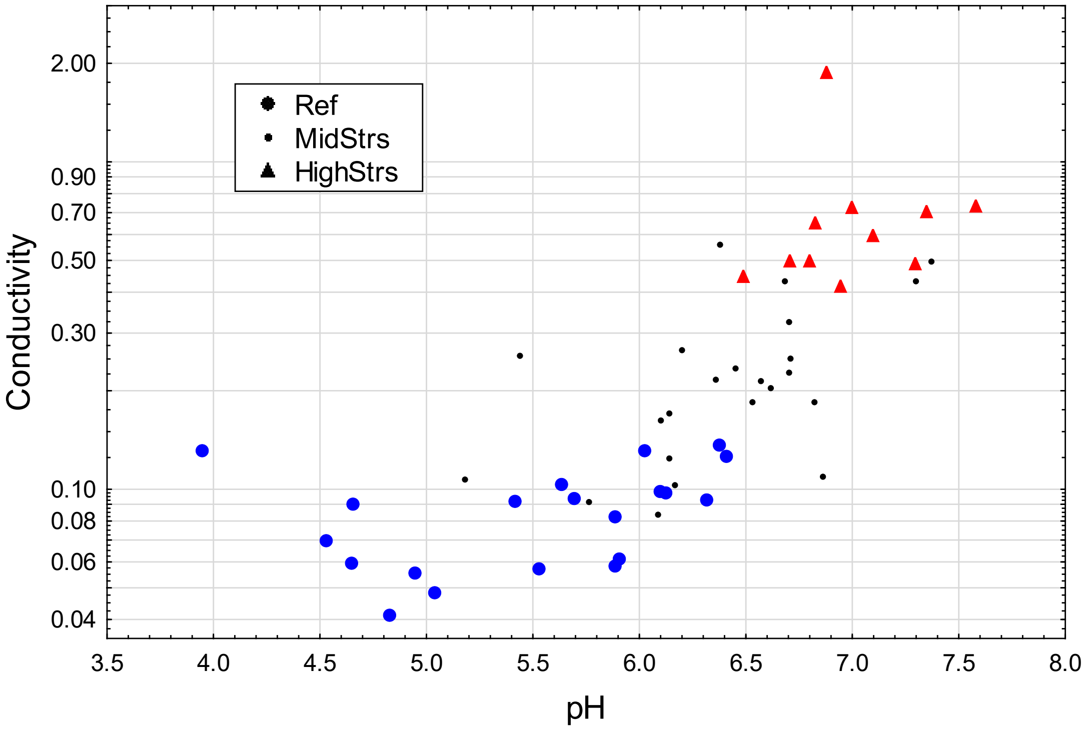


Figure 19. Conductivity (on a log-transformed axis) at sites with conductivity data in relation to pH, marked by disturbance category.

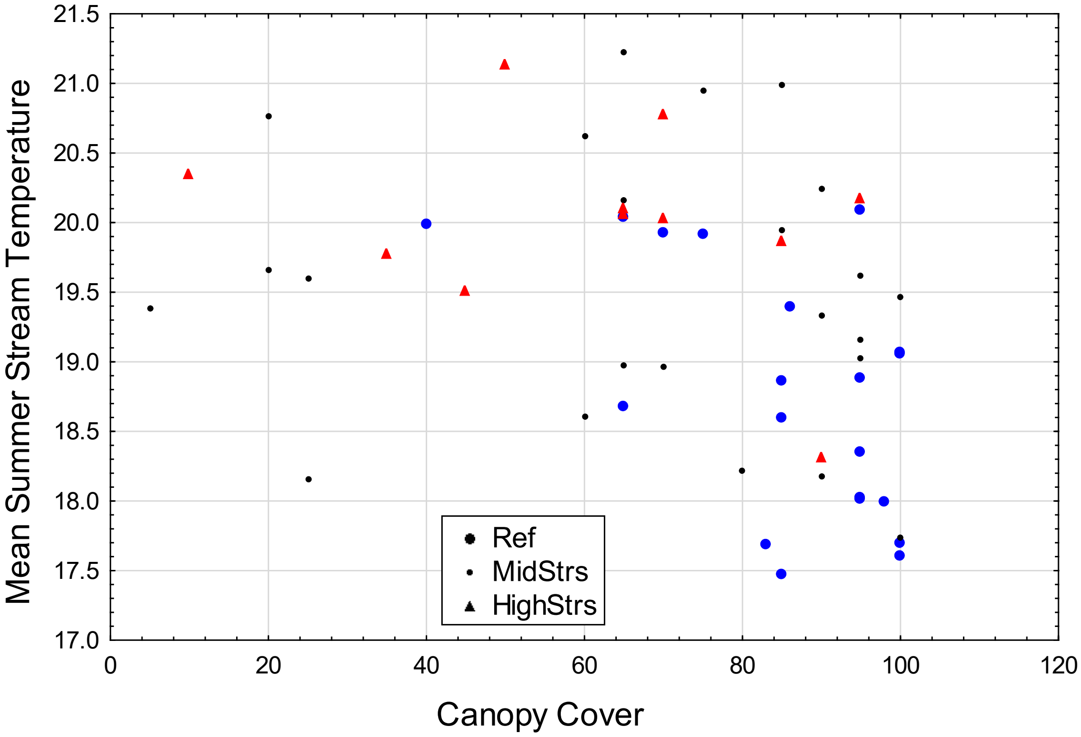


Figure 20. Modeled mean summer stream temperature (MSST; Hill et al. 2013) in relation to percent canopy cover, marked by disturbance category.

The IBI responds negatively to percent sand, silt, and clay in the stream substrate (Figure 21). Reference sites have the full range of fine sediments. IBI scores in the reference sites decline slightly as fines increase, as do values in non-reference sites. Fine sediments were not identified as a classification factor when calibrating the IBI. However, the response is slight and no accounting for substrate is needed for index assessments.

The IBI is relatively unresponsive to stream size (as measured by drainage area) (Figure 22) and water temperature, as measured by modeled summer stream temperature (Figure 23) and *in situ* water temperature from the SNEP sites (Figure 28). These variables were explored and discounted as classification variables in the site classification analysis. Stressed sites have warmer predicted summer temperatures and have lower IBI scores than the cooler reference sites (Figure 23). Within reference sites, the IBI was unresponsive to modeled summer and *in situ* water temperatures (Figures 23 and 24).

Though classification analysis indicated possible differences in reference sample composition across ecoregions and with varying percentage of water and wetland in the watershed, the index does not show a strong relationship with these variables within reference sites (Figures 25 and 26). The reference site with a low IBI score is in the Narragansett-Bristol Lowlands and has relatively high percent water and wetland, but does not indicate a strong pattern or bias of the index. Index values in sites with >20% water and wetland did not have the highest IBI scores, but the scores were aligned with the range of other reference scores, except for the one outlier.

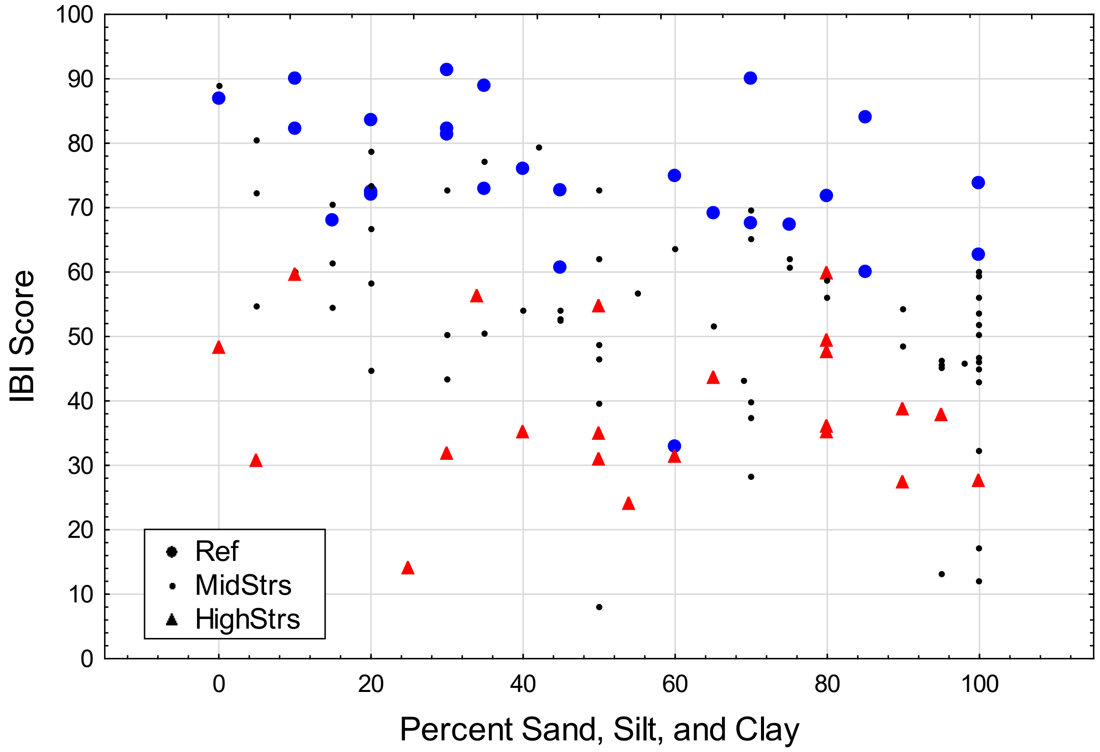


Figure 21. Percent sand, silt, and clay substrates in relation to IBI scores, marked by disturbance category.

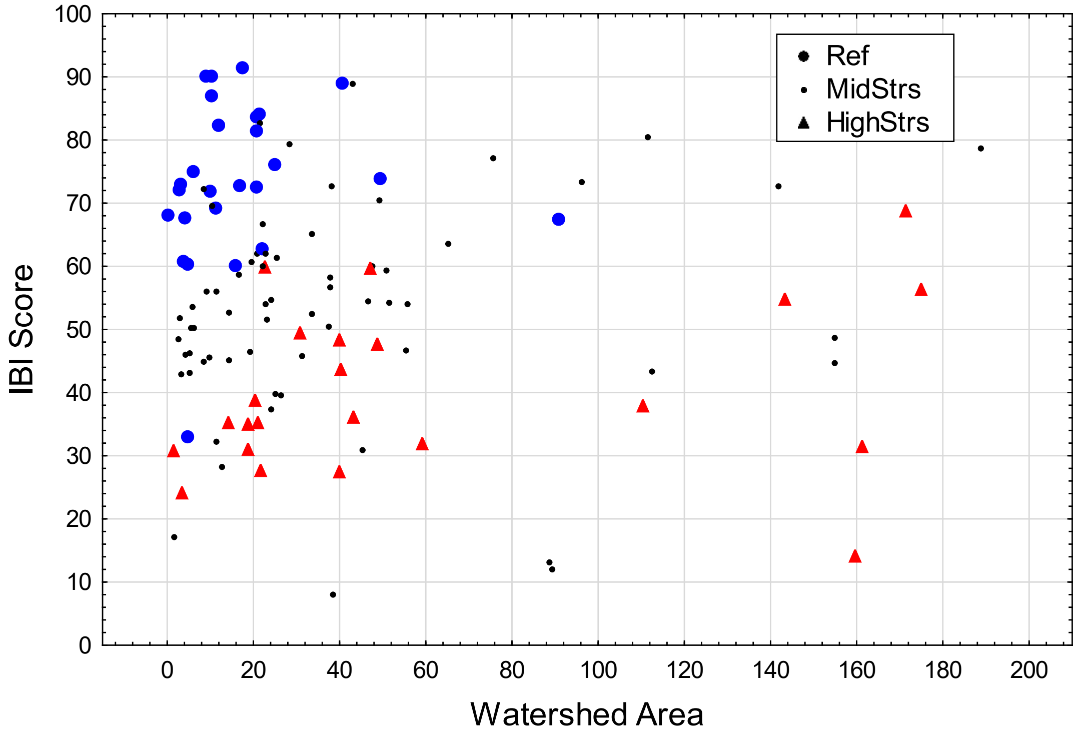


Figure 22. Site drainage area (on a log-transformed axis) in relation to IBI scores, marked by disturbance category.

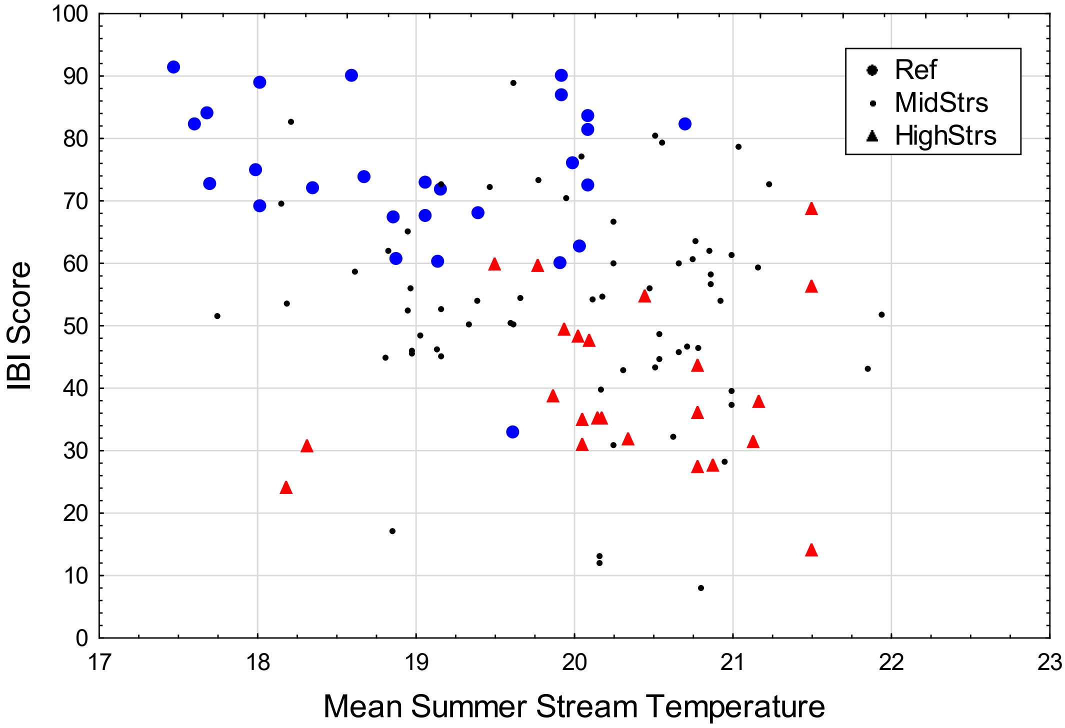


Figure 23. Mean Summer Stream Temperature (MSST) in relation to IBI scores, marked by disturbance category.

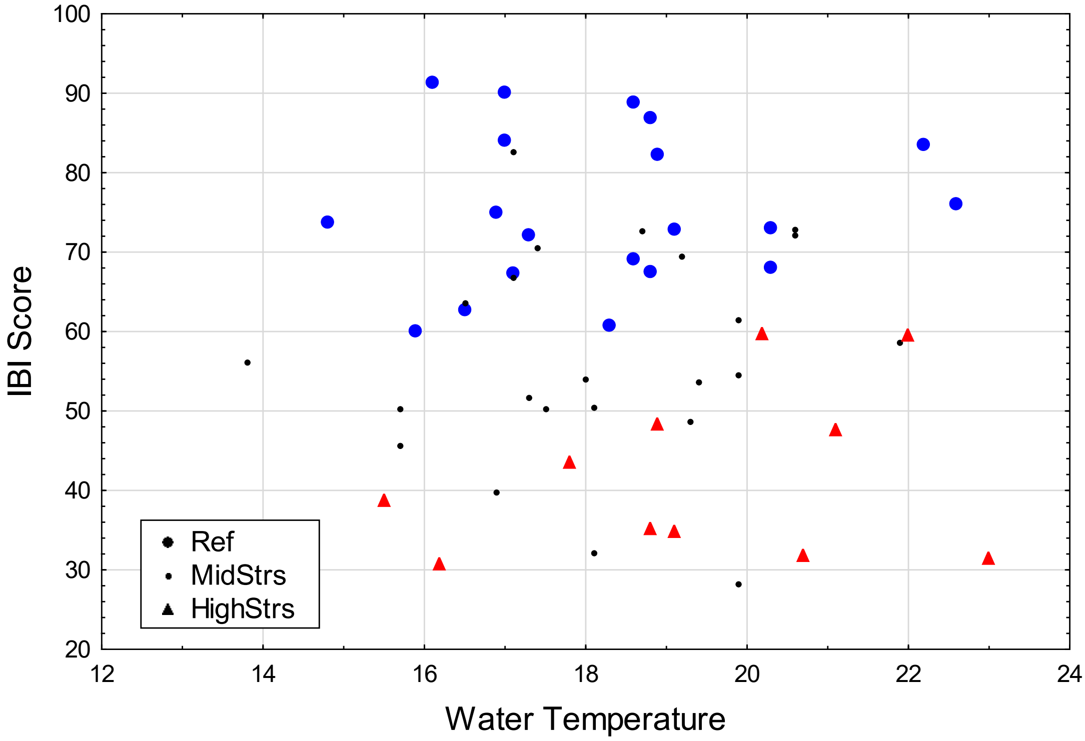


Figure 24. In situ (measured) stream temperature in relation to IBI scores, marked by disturbance category.

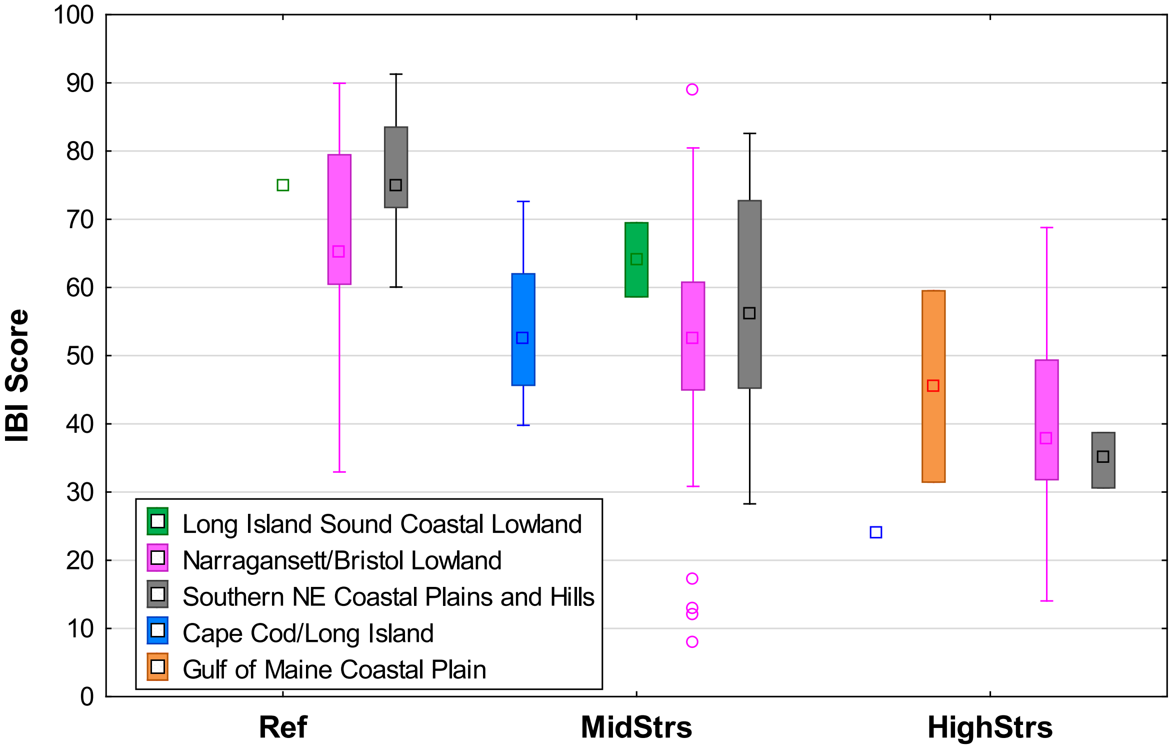


Figure 25. IBI score distributions (medians, interquartile ranges, non-outlier ranges, and outliers) in Level 4 ecoregions and disturbance categories; reference (Ref), intermediate (MidStrs), and stressed (HighStrs).

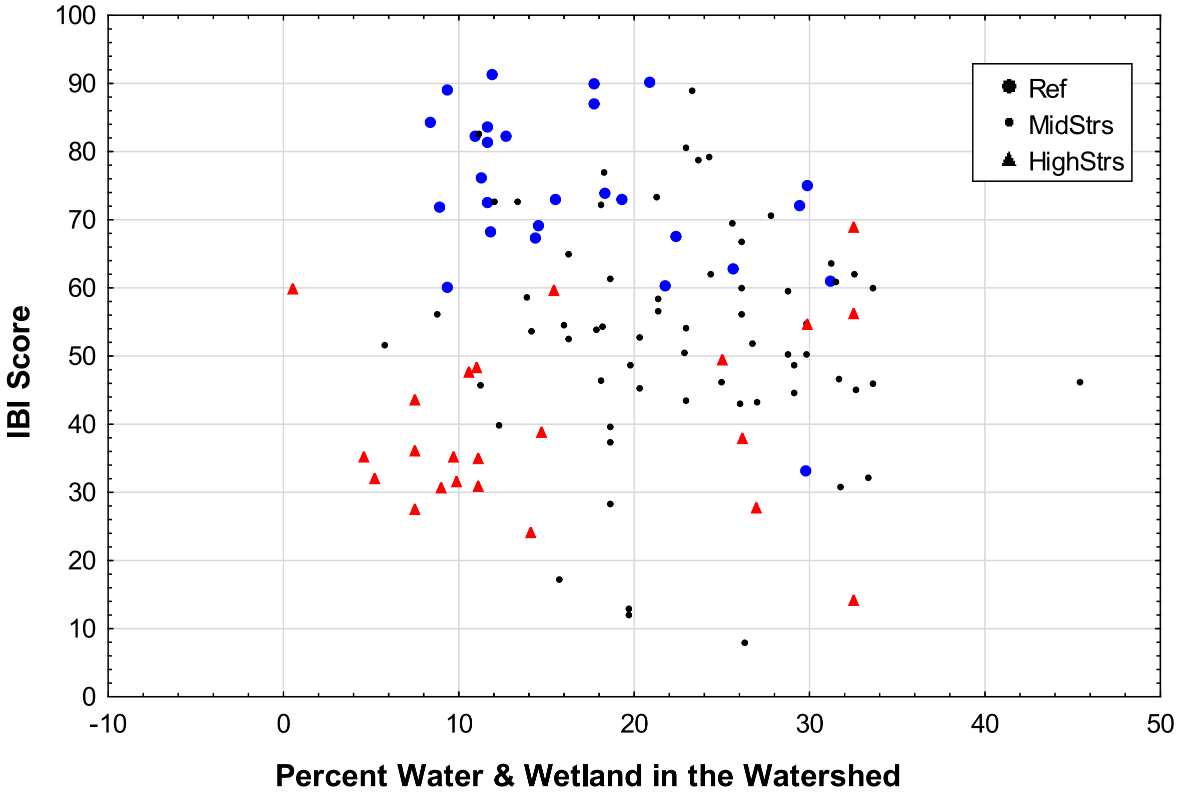


Figure 26. Percent water and wetland in the watershed in relation to IBI scores, marked by disturbance category.

6 Exploration of assessment thresholds

Once site classes are established and indices are calibrated, some entities establish thresholds for numeric biocriteria. We used multiple analyses to identify possible thresholds associating ranges of index values with biological condition categories. However, before identifying thresholds, we found that revisions to the taxa traits were needed, and this changed the index scores when compared to index scores calculated on the calibration traits. The shift in index scores was acknowledged and incorporated into this analysis of thresholds.

Explanation of Trait Changes and Index Adjustments

The data used for index calibration was based on taxa traits that were available at the time of the analysis and using metric scoring formulae based on the 300-count data. Taxa lists are not static and should be updated with new and better information as it becomes available, as it did over the project timeline. As the project progressed, the taxa traits were updated based on conferences with MassDEP biologists (Bob Nuzzo and Allyson Yarra) and the contract taxonomist (Mike Cole). The metric scoring formulae were based on distribution statistics first in the calibration data and then in the combination of calibration data and virtually subsampled data. Changes in traits and scoring formulae resulted in changes in metric and index scores between the original calibration data and the metric and index values in Attachment C.

The taxa traits that were changed over time included tolerance values and voltinism traits. In most cases, missing values were completed based on new information or association with similar taxa with existing traits. There were no changes in tolerance values, only additions of new values. For example, *Physa* (Mollusca) first had no tolerance value and then the value was added (8, tolerant). For voltinism traits, an important change was applied to Elmid beetles. Per feedback from Mike Cole, we assigned all Elmids to the 'semi-voltine' category (vs. previously, Elmid taxa were assigned to a mix of categories (blank, uni-voltine, semi-voltine). Revised taxa traits are tabulated in Attachment B. The trait revisions resulted in higher percentages of semi-voltine taxa in the revised metric calculations compared to the calibrated metrics. When the original scoring formula was applied to the pt_volt_semi metric, there were many high scores and many scores of 100 because of the increased number of recognized semi-voltine taxa.

The 5th and 95th percentiles of metrics based on 300-count data were used in calibration. As the project evolved to consider application with 100-count and 200-count data, the scoring formulae were changed to include the percentiles of those data also (as an average value for the three data sets). The changes to the scoring formulae were minor and were not expected to substantially affect metric and index scores.

The overall effect of the changes in metric traits and scoring were an upward shift in index values (Figure 27). The regression line for the calibration and revised index scores has a slope of almost 1 (0.99), indicating that the adjustment is applicable along the whole index gradient. The revised index is 4.9 points higher than the calibration index, in general. This shift should be accounted for when applying the index. Threshold development proceeded using index scores calculated from the revised taxa traits and the scoring formulae in Table 14.

Reference Distribution Statistics

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.

The low gradient, multihabitat SNEP IBI was developed using reference condition concepts to identify sites with relative degrees of disturbance due to human activities. The reference and highly stressed conditions for low gradient sites were defined using quantitative criteria of measures of stressors and stressor sources. The absolute degree of disturbance is undefined, though there are relatively fewer stressors in the reference condition compared to intermediate and high-stress conditions.

Distribution statistics in reference sites and all sites can inform possible thresholds, allowing assessment of sites that are similar to reference. These reference sites have few stressors and a biological condition representing a somewhat natural standard. Any index value above the minimum of reference index values might be a reference site. However, given that the reference sites were defined with relative, not absolute, stressor criteria and that there is variability in biological conditions, it is likely that the minimum value is not representative of acceptable reference conditions. Rather, the minimum reference index value probably should not be recognized as an acceptable natural standard. In contrast, a threshold set at the median of index values would discount half of the reference sites, which would suggest that the reference sites were poorly defined and the reference condition has substantial errors.

Thresholds based on a lower percentile of reference index scores describe points on the index scale above which conditions represent predominantly natural community types and below which biological conditions are departing from the core natural standard and might be impacted, erroneously designated reference sites, or simple errors due to biological and site variability. The 10th - 25th percentiles of reference index values are common thresholds used in bioassessments. One of these percentiles could be selected as a threshold for assessing low gradient biological conditions using the index. In our data set, using the revised traits and scoring formulae in Table 14, these percentiles correspond to index values of 63 – 70 index points, respectively (Table 16). Because of the uneven distribution of reference and highly stressed index values, the percentage of highly stressed sites that are below 63 – 70 index points ranges from 91 – 96%. Using the index derived from revised traits and applying these thresholds, 91 – 96% of highly stressed sites in the current data set would be identified as biologically impacted .

One strategy for selecting a threshold is to balance errors in assessing reference and highly stressed sites: there should be as many reference sites identified as impacted as there are highly stressed sites identified as unimpacted. This is based on the premise that each data set and condition was identified with equal degrees of certainty and therefore error should be the same. Type I and Type II errors are associated with reference sites erroneously identified as impacted and highly stressed sites identified as unimpacted, respectively. In our data set, Type I and Type II errors are equal at index values at the 10th percentile, at approximately 63 index points (Table 16).

The standard deviation of the reference index distribution was 12.8 index points. A threshold of 63 index points is a little more than 1 standard deviation from the reference mean. The mean reference index score (76.4) minus 1 standard deviation is 63.7 index points.

Table 16. Low gradient IBI distribution statistics for the index calculated after trait revisions.

	All sites distribution statistics	Reference distribution statistics	Type I error	DE	Type II error
Valid N	114	27			
Minimum	7.9	34.1	0%	26.1	73.9
5th Percentile	26.2	59.5	5%	82.6	17.4
10th Percentile	33.9	63.1	10%	91.3	8.7
15th Percentile	40.5	67.1	15%	91.3	8.7
20th Percentile	43.4	69.2	20%	95.7	4.3
Lower Quartile	47.5	70.1	25%	95.7	4.3
Mean	59.9	76.4			
Median	62.4	79.1			
Upper Quartile	73.5	86.6			
Maximum	94.0	94.0			

Regression on the Calibrated Index

Similar analyses of potential thresholds were conducted using the index values derived from the calibration data; unadjusted for trait revisions. In those analyses, an index value of 60 points was the 10th percentile and balanced the Type I and Type II errors. A regression of the calibration index and the revised index showed that revised index values were generally 5 index points greater than calibration index values (Figure 27). The regression equation was $y = 0.99\ x + 4.89$ ($r^2 = 0.95$). If the regression equation is applied to the suggested calibration index threshold, the interpolated revised index threshold would be 64.3 index points.

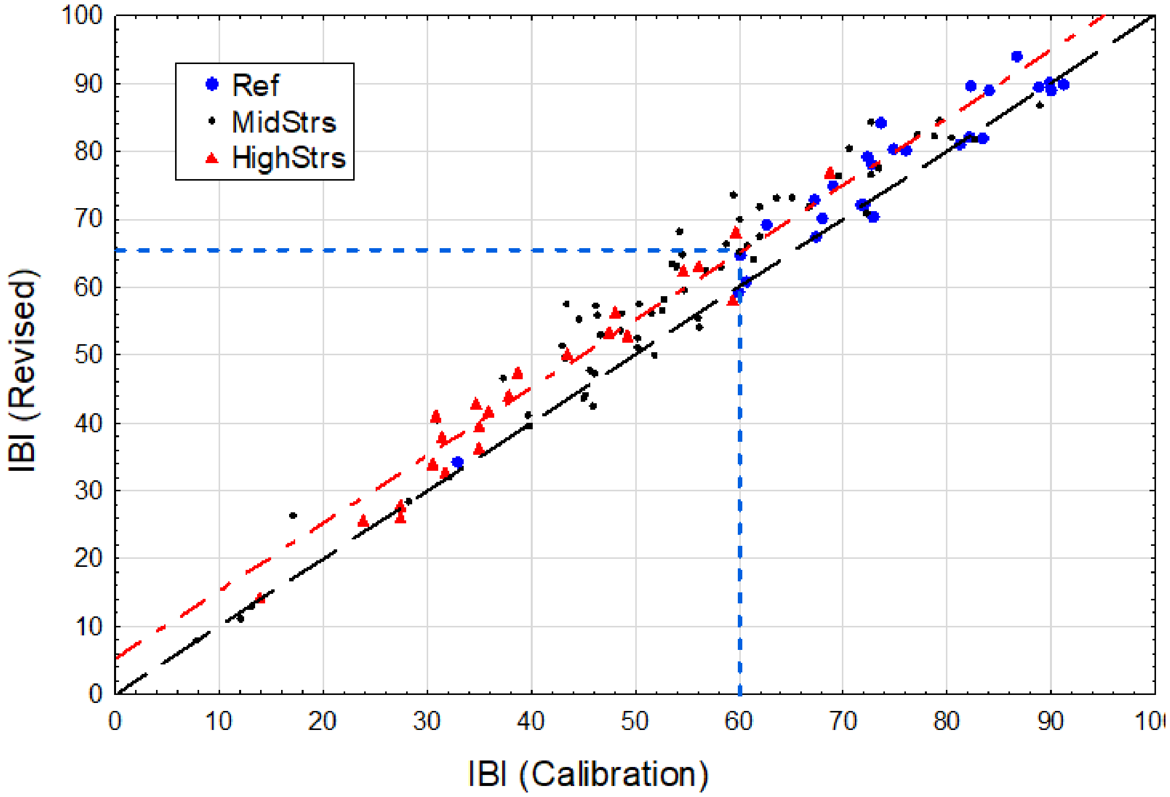


Figure 27. IBI values comparing calibration data and revised data, showing the unity line (black dashed), regression line (red dashed), and the central threshold at 60 in calibration data and 64 in revised data. The regression equation is y = 0.99 x + 4.89 (r^2 = 0.95).

These indications from reference distributions, balanced errors, standard deviations, and comparison to the preliminary calibration threshold suggest that a general condition threshold dividing satisfactory conditions from moderately degraded conditions should be in the range of 63 – 70 index points. If the balance of errors and the 10th percentile are given greater weight because they recognize potential error in both reference and highly stressed data sets and they are based on common precedent, then the threshold value would be closer to 63 index points. A general threshold of 63 index points is recommended.

Secondary Thresholds

As demonstrated in the MassDEP 100-count riffle habitat IBI threshold analyses (Stamp and Jessup 2020), secondary thresholds could be identified within the generally unimpacted and generally impacted index ranges. This would allow for refined emphasis in biological condition when prioritizing or justifying management decisions. Within the generally unimpacted index range, refined conditions could be described as Exceptional or Satisfactory based on a secondary threshold somewhat above 63 index points. A simple bisection of the unimpacted index range would suggest a threshold of 81.5 index points, half-way between the general threshold and the maximum of the index scale. In a similar fashion, the impacted range of the index scale could be bisected to describe a threshold between Moderately Degraded and Severely Degraded conditions at an index value of 31.5.

A more complex determination of secondary thresholds can be explored using proportional odds logistic regression. This technique estimates the probabilities of membership in the reference, moderately stressed, and highly stressed groups based on index values within those categories. The points at which there is equal probability between groups can describe a potential threshold that would evenly divide Exceptional and Satisfactory index values and also Moderately Degraded and Severely Degraded index values. Based on proportional odds logistic regression, a threshold between Exceptional and Satisfactory conditions was identified at 82 index points. The threshold between Moderately Degraded and Severely Degraded conditions was identified at 36 index points (Figure 28). These thresholds recognize the observed range of index values within disturbance groups, as opposed to the simple bisection, which uses the entire range of index values, regardless of the observed range. Recognition of the observed range of values is a more empirical method that is recommended. The crossover for highly stressed and reference membership probabilities is at 59 index points. We have less confidence in this potential general threshold because of the influence of the mid-stress distribution.

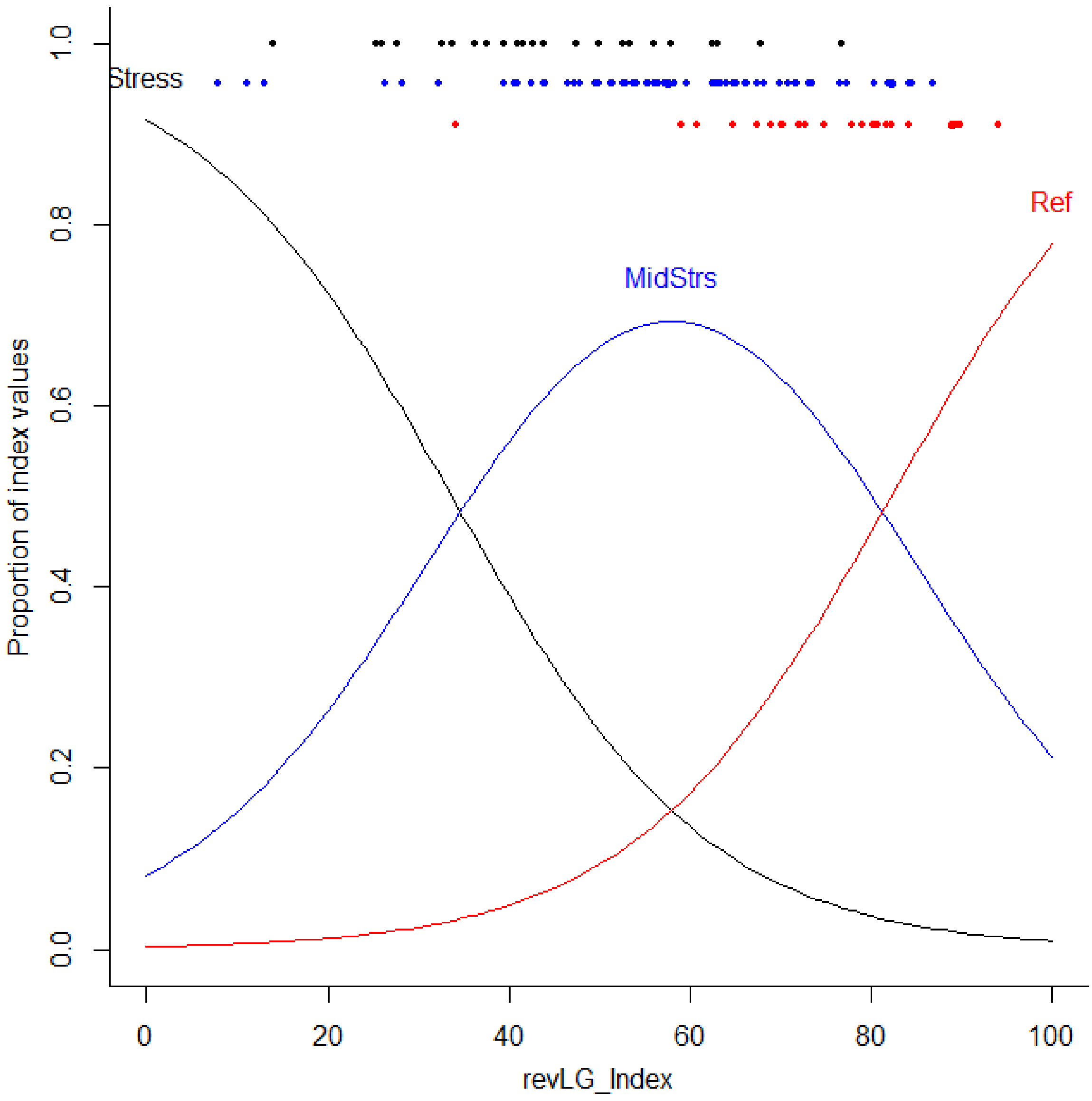


Figure 28. Proportional odds logistic regression graph, showing probability of membership in the highly stressed (Stress), moderately stressed (ModStrs), and reference (Ref) disturbance categories. Actual data points for the revised index are plotted at the top of the graph.

Based on the analyses described above, thresholds for the low gradient, multihabitat IBI with revised traits are as in Table 17 and Figure 29. The map in Figure 30 shows the spatial distribution of sites in the four biological condition categories based on the recommended thresholds. These thresholds are preliminary and are subject to further review, refinement, and approval by MassDEP and RI DEM before they are applicable in biological assessment programs.

Table 17. Threshold ranges and recommended SNEP IBI values for indication of biological conditions in low gradient streams.

<table>
<tr><td></td><td colspan="4">General unimpacted conditions</td><td colspan="4">General impacted conditions</td></tr>
<tr><td></td><td colspan="2">Exceptional Conditions</td><td colspan="2">Satisfactory Condition</td><td colspan="2">Moderately Degraded Condition</td><td colspan="2">Severely Degraded Condition</td></tr>
<tr><td>Index threshold range</td><td colspan="3">81.5 - 82</td><td colspan="2">63-70</td><td colspan="3">31.5 - 36</td></tr>
<tr><td>Recommended index threshold</td><td colspan="3">82</td><td colspan="2">63</td><td colspan="3">36</td></tr>
</table>

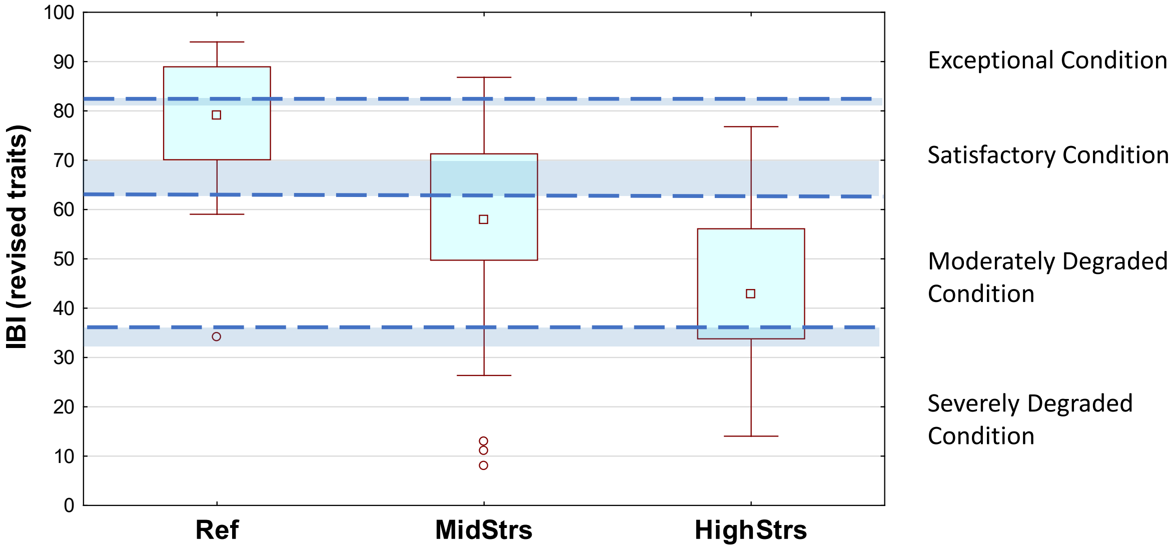


Figure 29. Low gradient SNEP index distributions plotted by disturbance category and showing recommended thresholds (dashed lines) and threshold ranges (shaded bars) to describe index values associated with narrative condition categories.

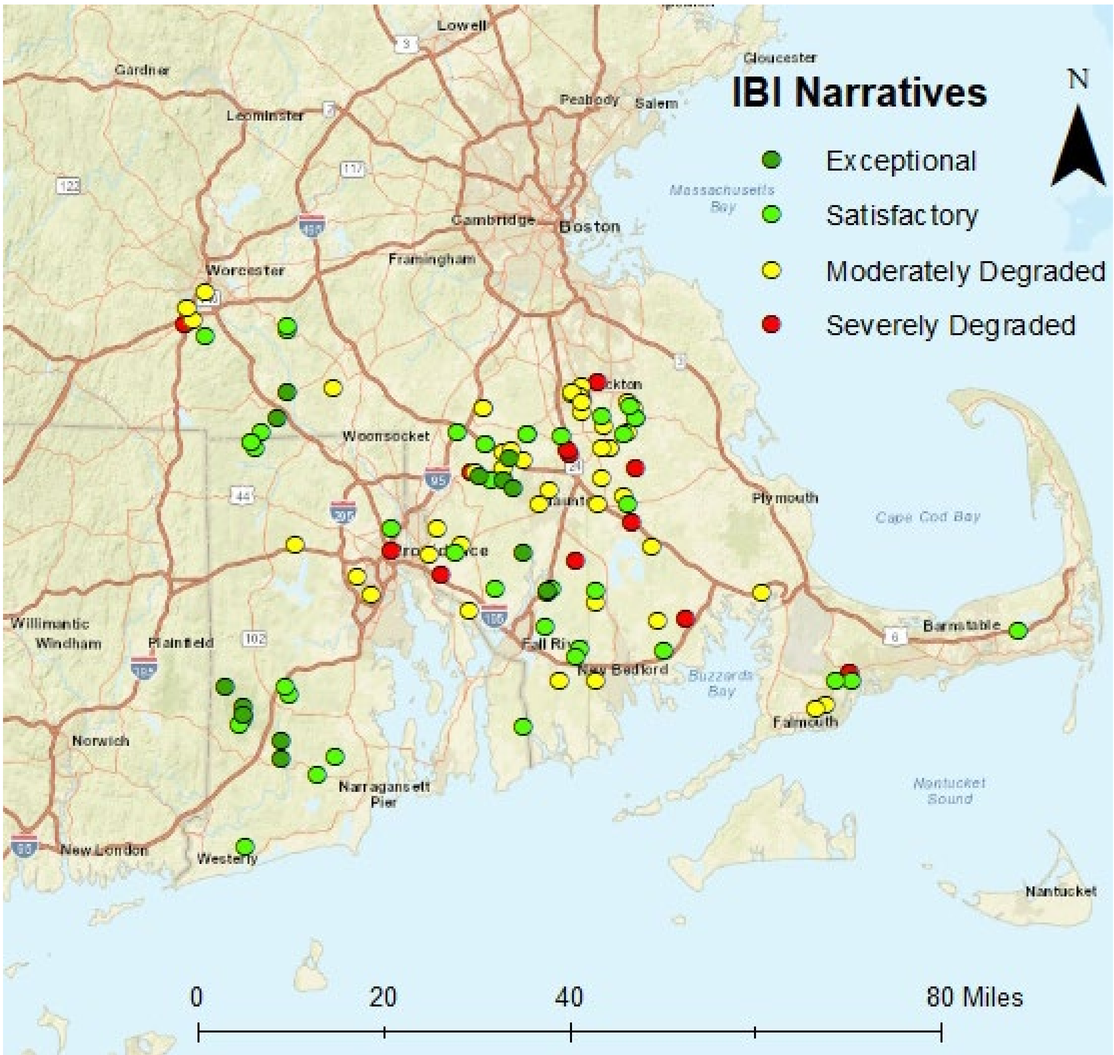


Figure 30. Low gradient, multihabitat sites in the SNEP region color-coded by biological condition category based on the recommended IBI thresholds in Table 17.

7 Applications

The SNEP IBI improves the ability of resource managers in the Southeast New England region to identify degradation in biological integrity and water quality in low gradient, non-tidal, wadeable streams. The IBI is comprised of biological metrics that were found to be responsive to a general stressor gradient, are ecologically meaningful, diverse in response mechanisms and represent multiple metric categories (composition, functional feeding group, tolerance, and voltinism). During calibration, the IBI had minimal error when discriminating between reference and stressed sites. When validated with independent data, the IBI also performed well, showing the expected direction of response in relation to various measures of anthropogenic disturbance. The IBI was calibrated using the Reference Condition approach, which bases biological expectations on least-disturbed reference sites. 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.

The IBI can be calculated using information presented in this report to assemble valid sample data, calculate metrics from revised traits, score metrics, and calculate the index. However, an option for calculating the IBI is also available through a free R-based tool (referred to as a Shiny app). The IBI calculator can be accessed via this weblink:

https://tetratech-wtr-wne.shinyapps.io/SNEPtools

Shiny apps are interactive web applications that are linked to R software, which is an open source programming language and software environment for statistical computing. The IBI calculator is easy to operate and only requires an input dataset (formatted in a specific way) to function. Users should keep in mind that they can run any data through the IBI calculator and get a result. However, if samples do not meet the criteria listed below, results should be interpreted with caution.

Criteria:

- Geographic area: the Southeast New England region of Massachusetts and Rhode Island (Figure 1)
- Stream type: low gradient, non-tidal, wadeable, perennial, slow moving streams with soft or hard substrate, with at least one of the following habitats: snags, root wads, leaf packs, aquatic macrophytes, undercut banks, overhanging vegetation, or hard bottom.
- Subsample size: 300-count samples are recommended for best performance, but the IBI can also be applied to 200 or 100-count samples
- Taxonomic resolution: lowest practical level
- Collection gear: Aquatic Kick Net with 500-µm mesh
- Collection method: 10 kicks, sweeps, and/or jabs from multiple habitats (listed above) taken over a 100-m reach and then composited into a single sample. Habitats are sampled in proportion to their occurrence
- Collection period: July 1–September 30

The macroinvertebrate IBI can be used to assess stream degradation relative to least-disturbed multihabitat streams. Some state biomonitoring programs take the additional step of establishing numeric IBI thresholds in their Surface Water Quality Standards (SWQS) to designate different categories of biological condition and to assess attainment of aquatic life use standards. MassDEP and RI DEM explored potential thresholds for four biological condition categories (Exceptional Condition, Satisfactory Condition, Moderately Degraded, and Severely Degraded). The thresholds proposed in this report are preliminary and subject to further review, refinement, and approval by MassDEP and RI DEM before they are applicable in biological assessment programs. Moving ahead, in addition to further exploring potential IBI thresholds, MassDEP, RI DEM and other biomonitoring programs in the SNEP region will continue to evaluate the performance of the low gradient IBI as new data are collected.

8 References

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.

Block, B. D., J. Stamp, and B. K. Jessup. 2020. MassDEP/SNEP index selection and evaluation with consideration of sub-sample sizes. Prepared for Massachusetts Department of Environmental Protection and Southeast New England Coastal Watershed Restoration Program.

Burns, D., T. Vitar, J. McDonnell, J. Hassett, J. Duncan, and C. Kedall. 2005. Effects of suburban development on runoff generation in the Croton River basin, New York, USA. *Journal of Hydrobiology* 311:266 – 281.

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

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.

Gibbs, D. A., B. Bierwagen. 2017. Procedures for delineating and characterizing watersheds for stream and river monitoring programs. (EPA/600/R-17/448F). Washington, DC: U.S. Environmental Protection Agency, Office of Research and Development.

Gotelli, N. J. and G. R. Graves. 1996. Null models in ecology. Washington, DC: Smithsonian Institution Press.

Hatt, B.E, T.D. Fletcher, C.J. Walsh, and S.L. Taylor. 2004. The influence of urban density and drainage infrastructure on the concentrations and loads of pollutants in small streams. *Environmental Management* 34:112 – 124.

Hill, R.A., C.P. Hawkins, and D.M. Carlisle. 2013. Predicting thermal reference conditions for USA streams and rivers. *Freshwater Science* 32(1):39-55. doi:10.1899/12-009.1.

Hill, R. A., M. H. Weber, S. G. Leibowitz, A. R. Olsen, and D. J. Thornbrugh. 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.

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. 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. 2020. Development of indices of biotic integrity for assessing macroinvertebrate assemblages in Massachusetts freshwater wadeable streams. Prepared for the Massachusetts Department of Environmental Protection.

Jessup, B., Block, B. and J. Stamp. 2021. Development of an Index of Biotic Integrity for macroinvertebrates in freshwater low gradient wadeable streams in Massachusetts. Prepared for the Massachusetts Department of Environmental Protection.

Johnson, Z., S. Leibowitz and R. Hill. 2018. Revising the index of watershed integrity national maps. *Science of the Total Environment*. 10.1016/j.scitotenv.2018.10.112.

Karr, J.R. 1981. Assessment of biotic integrity using fish communities. *Fisheries*, *6*(6):21-27.

Leppo, E.W., J. Stamp, and J. van Sickle. 2021. BioMonTools: Tools for Biomonitoring and Bioassessment. R package version 0.5.0.9039. https://github.com/leppott/BioMonTools.

Lussier, S.M., S.N. da Silva, M. Charpentier, J.F. Heltshe, S.M. Cormier, D.J. Klemm, M. Chintala, and S. Jayaraman. 2008. The influence of suburban land use on habitat and biotic integrity of coastal Rhode Island streams. *Environmental Monitoring and Assessment* 139:119 – 136.

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.

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

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.

Poff, N. L.; J. D. Olden, N. K. M. Vieira, D. S. Finn, M. P. Simmons, and B. C. Kondratieff. 2006. Functional trait niches of North American lotic insects: traits-based ecological applications in light of phylogenetic relationships. *Journal of the North American Benthological Society* 25(4):730−755.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Roth, N. E., M. T. Southerland, J. C. Chaillou, J. H. Vølstad, S. B. Weisberg, H. T. Wilson, D. G. Heimbuch, and 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-MANTAEA-97-2.

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. *Journal of the North American Benthological Society*, 2008, 27(4):878–891

Tetra Tech. 2019. Sampling and Analysis Plan - Data Collection for Development of an Index of Biotic Integrity for Freshwater Low-Gradient Wadeable Streams in Southern New England. Prepared for the NEIWPCC, Lowell, MA. Prepared by Tetra Tech, Montpelier, VT.

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

U.S. Environmental Protection Agency (U.S. EPA). 2011. Level III and IV ecoregions of the continental United States. U.S. EPA, National Health and Environmental Effects Research Laboratory, Corvallis, Oregon, Map scale 1:3,000,000. Available online at: https://www.epa.gov/eco-research/level-iii-and-iv-ecoregions-continental-united-states.

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 online: https://cfpub.epa.gov/ncea/global/recordisplay.cfm?deid=239585

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. Authorship: Jen Stamp, Anna Hamilton, Britta G. Bierwagen, Debbie Arnwine, Margaret Passmore and Jonathan Witt. Available online: https://cfpub.epa.gov/ncea/global/recordisplay.cfm?deid=307973

Yoder, C. O., and E. T. Rankin. 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.

Yuan, L. 2006. Estimation and Application of Macroinvertebrate Tolerance Values. Report No. EPA/600/P-04/116F. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, D.C.