Development of an Index of Biotic Integrity for Macroinvertebrates in Freshwater Low Gradient Wadeable Streams in Massachusetts

FINAL REPORT

***Prepared for*:**

Massachusetts Department of Environmental Protection
James Meek, Work Assignment Manager

***Prepared by*:**

Ben Jessup, Ben Block, and Jen Stamp
Tetra Tech
73 Main Street, Room 38, Montpelier, VT 05602

April 7, 2021

Executive Summary

The Massachusetts Department of Environmental Protection (MassDEP) is responsible for sampling and assessing Massachusetts's surface water quality pursuant to the Clean Water Act (CWA) Sections 305(b), 303(d), and 314. The Massachusetts Surface Water Quality Standards (SWQS) (314 CMR 4.00; MassDEP 2013) has narrative biological criteria that define biological integrity as "the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having species composition, diversity, and functional organization comparable to that of the natural habitat of the region." Waters supporting Aquatic Life Use should be suitable for sustaining "a native, naturally diverse, community of aquatic flora and fauna" that also support their "reproduction, migration, growth and other critical functions" (MassDEP 2013).

To assess the biological integrity of its freshwater, perennial, wadeable streams, MassDEP has been collecting macroinvertebrates since the 1980s, focusing primarily on streams with fast currents and rocky substrate. Recently, Indices of Biotic Integrity (IBIs) for macroinvertebrate assemblages in riffle habitats were calibrated for all but the southeastern portion of Massachusetts (Narragansett/Bristol Lowlands, Cape Cod, and the Islands) (Jessup and Stamp 2020). IBIs are numeric representations of biological conditions based on the combined signals of several different assemblage measurements (Karr 1981). The IBIs were comprised of biological metrics that were found to be responsive to a general stressor gradient. By scoring the metrics for each sample and averaging the scores in a multimetric index, the resulting IBI indicates the biological condition of the stream on a relative scale. The IBI scores in the reference sites are reasonable expectations for any stream in the region, and IBI scores that do not resemble the reference scores indicate that there might be stressors influencing the biological condition.

In 2010, MassDEP began to sample low gradient, slow-moving streams that either lack or have infrequent (< ~10%) riffle habitat. Low gradient streams can be found statewide and are prevalent in southeastern Massachusetts. Habitats include snags, root wads, leaf packs, aquatic macrophytes, undercut banks, overhanging vegetation, and hard bottom substrates. Structures and functions of macroinvertebrate assemblages in these low gradient streams also differ from those in fast-moving, rocky-bottom streams. To effectively sample these slow-moving streams, MassDEP developed a new collection method (referred to as the multihabitat method) in which organisms are collected from multiple habitats and then composited into a single sample.

In this report, we describe the development of a statewide low gradient multihabitat IBI for Massachusetts. The IBI calibration dataset included data from 178 sites, some of which were located in Rhode Island. Data from the Rhode Island sites and some of the Massachusetts sites were collected as part of a separate but concurrent IBI project in the Southeast New England Program (SNEP) region, which includes Cape Cod, Narragansett Bay, and Buzzards Bay. 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) in both the MassDEP and SNEP datasets and metrics that were familiar to the workgroup members, ecologically meaningful, and diverse in response mechanisms. The IBI

also 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 substantial (Spearman correlation coefficients ≥ |0.49|) and kept 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, Satisfactory, Moderately Degraded, and Severely Degraded) that can be used in the Consolidated Assessment and Listing Methodology (CALM) to interpret the narrative biological criteria in the SWQS. The thresholds proposed in this report are preliminary and subject to further review, refinement, and approval by MassDEP before they are applicable in biological assessment programs. The new low gradient IBI and preliminary thresholds improve MassDEP's diagnostic ability to identify degradation in biological integrity and water quality and will be re-evaluated in coming years as MassDEP obtains and analyzes more low gradient samples.

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

Metric (abbrev)	Response to 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

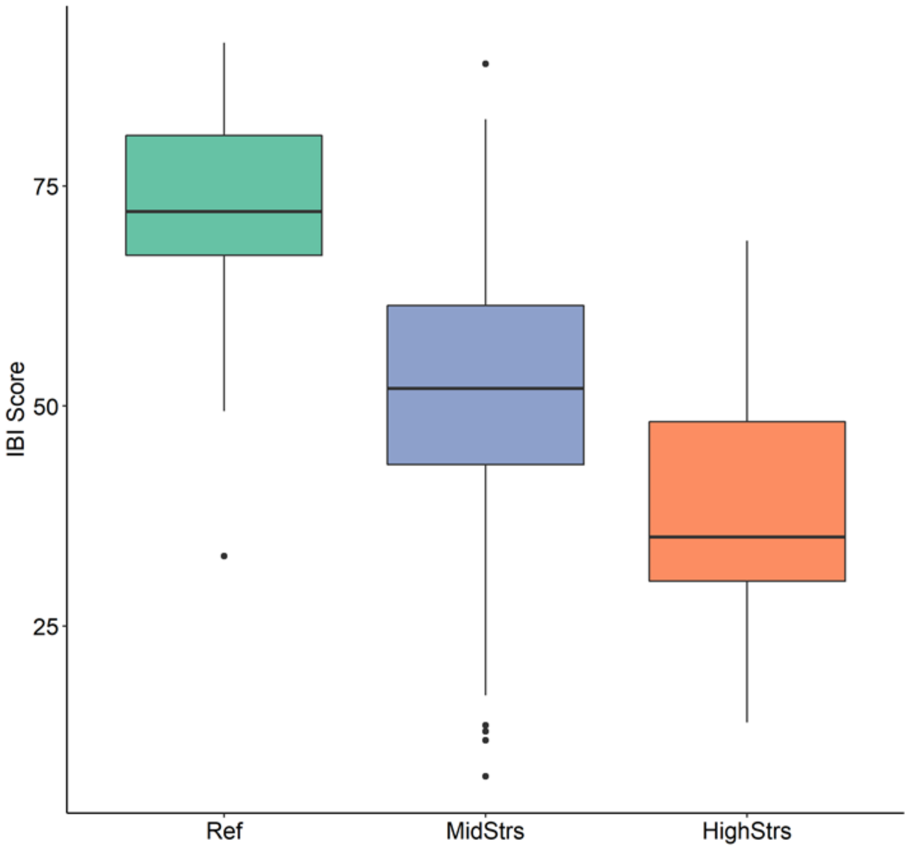


Figure ES-1. Distributions of low gradient IBI values in reference (Ref), intermediate (MidStrs), and stressed (HighStrs) sites.

Acknowledgments

The index development process was supported by the Massachusetts Department of Environmental Protection (MassDEP) through a contract with Tetra Tech, Inc. The contract was facilitated by James Meek of MassDEP. The index development team, which consisted of James Meek, Allyson Yarra, Robert Nuzzo, Arthur Johnson, Joan Beskenis, Anna Mayor, and Robert Maietta, participated in bi-weekly calls and provided feedback throughout the process. James Meek performed a thorough review of site disturbance assignments and index alternatives. Allyson Yarra and Robert Nuzzo provided data and reviewed taxa attribute assignments, site disturbance assignments, and index alternatives. Michael Cole of Cole Ecological, Inc. provided guidance on resolving nomenclature changes, feedback on taxa attribute assignments, and worked with Kelly Nolan and Natalie Ohl (Watershed Assessment Associates) on the taxonomic identification quality control process.

In addition, the U.S. Environmental Protection Agency (EPA) Southeast New England Program (SNEP) provided support for the collection and processing of data from 54 sites that were used for index calibration. Maryann Dugan and Richard Friesner from NEIWPCC provided contract support for the SNEP project. State partners provided low gradient data that were used in the regional taxa tolerance analyses. The regional partners included Katie DeGoosh and Jane Sawyers from Rhode Island Department of Environmental Management, 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 Massachusetts. Prepared for the Massachusetts Department of Environmental Protection, Worcester, MA. Prepared by Tetra Tech, Montpelier, VT.

Acronyms

Acronym	Description
BFI	Base Flow Index
BPJ	Best Professional Judgement
CALM	Consolidated Assessment and Listing Methodology
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

The Massachusetts Department of Environmental Protection (MassDEP) is responsible for sampling and assessing Massachusetts's surface water quality pursuant to the Clean Water Act (CWA) Section 305(b), as well as identifying waterbodies that are not meeting water quality criteria and require development of a Total Maximum Daily Load (TMDL) according to Section 303(d) of the CWA. To help meet these requirements, MassDEP monitors biological conditions and assesses the integrity of macroinvertebrate assemblages in freshwater streams and rivers (Massachusetts Division of Watershed Management Watershed Planning Program 2016). MassDEP's biomonitoring program has been collecting macroinvertebrates from riffle/run areas in freshwater wadeable streams with fast currents and rocky substrate since the early 1980s using the Rapid Bioassessment Protocol (RBP) kick net method. Indices of Biotic Integrity (IBIs) for macroinvertebrate assemblages in riffle habitats have been calibrated for all but the southeastern portion of Massachusetts (Narragansett/Bristol Lowlands, Cape Cod, and the Islands) (Jessup and Stamp 2020). IBIs are numeric representations of biological conditions 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.

In 2010, MassDEP began to sample low gradient, slow-moving streams that either lack or have infrequent (<~10%) riffle habitat. Low gradient streams can be found statewide but are most prevalent in southeastern Massachusetts. Habitats for macroinvertebrates include snags, root wads, leaf packs, aquatic macrophytes, undercut banks, overhanging vegetation, fine sediments, and hard bottom. To effectively sample these habitats, MassDEP developed a new collection method in which field crews sample multiple habitats and composite them into a single sample. Because there are natural differences in the structure and function of macroinvertebrate assemblages in low gradient streams versus those in fast-moving, rocky-bottom streams, a new IBI was needed. In this report, we describe the development of a statewide low gradient multihabitat IBI for Massachusetts. Steps for index development 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 low gradient IBI will improve MassDEP's diagnostic ability to identify degradation in biological integrity and associated stressors in freshwater, wadeable, perennial streams.

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 184 samples from 178 unique sites in Rhode Island (RI) and Massachusetts (MA). Twenty-two sites were located in RI and 156 in MA (Figure 1, Table 1). The data were collected by MassDEP and Tetra Tech field crews. Though MassDEP started collection in low-gradient streams in 2010, they were not confident that their methods were consistent until 2013 and later. Tetra Tech collected the samples from streams in RI and southeastern MA in 2019 as part of an IBI development project funded by the Southeast New England Coastal Watershed Restoration Program (SNEP). Of the 178 sites, 109 were located in the SNEP region, which includes watersheds of southern Cape Cod, Narragansett Bay, and Buzzards Bay (Figure 1). Most of the sites were located in the Northeastern Coastal Zone Level 3 ecoregion (n=169), two were located in the Northeastern Highlands, and seven were located in the Atlantic Coastal Pine Barrens (U.S. EPA 2011). The distribution of sites across Level 4 ecoregions is summarized in Table 1. For the RI sites, all but three were located in the Narragansett/Bristol Lowland and Southern New England Coastal Plains and Hills Level 4 ecoregions, which occur in both RI and MA. The other three RI sites were in the Long Island Sound Coastal Lowland ecoregion (Table 1, Figure 1). Sites from other surrounding states were not used because they were either distant from the focus area of low-gradient landscapes in Massachusetts, did not collect comparable samples, or both.

Table 1. Distribution of sites across states and Omernik Level 3 and 4 ecoregions.

Level 3 ecoregion	Level 4 ecoregion name	Level 4 code	Number of sites	
			MA	**RI**
Northeastern Coastal Zone	Boston Basin	59d	5	0
	Connecticut Valley	59a	20	0
	Gulf of Maine Coastal Lowland	59f	10	0
	Gulf of Maine Coastal Plain	59h	24	0
	Long Island Sound Coastal Lowland	59g	0	3
	Lower Worcester Plateau/Eastern Connecticut Upland	59b	7	0
	Narragansett/Bristol Lowland	59e	66	5
	Southern New England Coastal Plains and Hills	59c	15	14
Northeastern Highlands	Worcester/Monadnock Plateau	58g	2	0
Atlantic Coastal Pine Barrens	Cape Cod/Long Island	84a	7	0
		Total	***156***	***22***

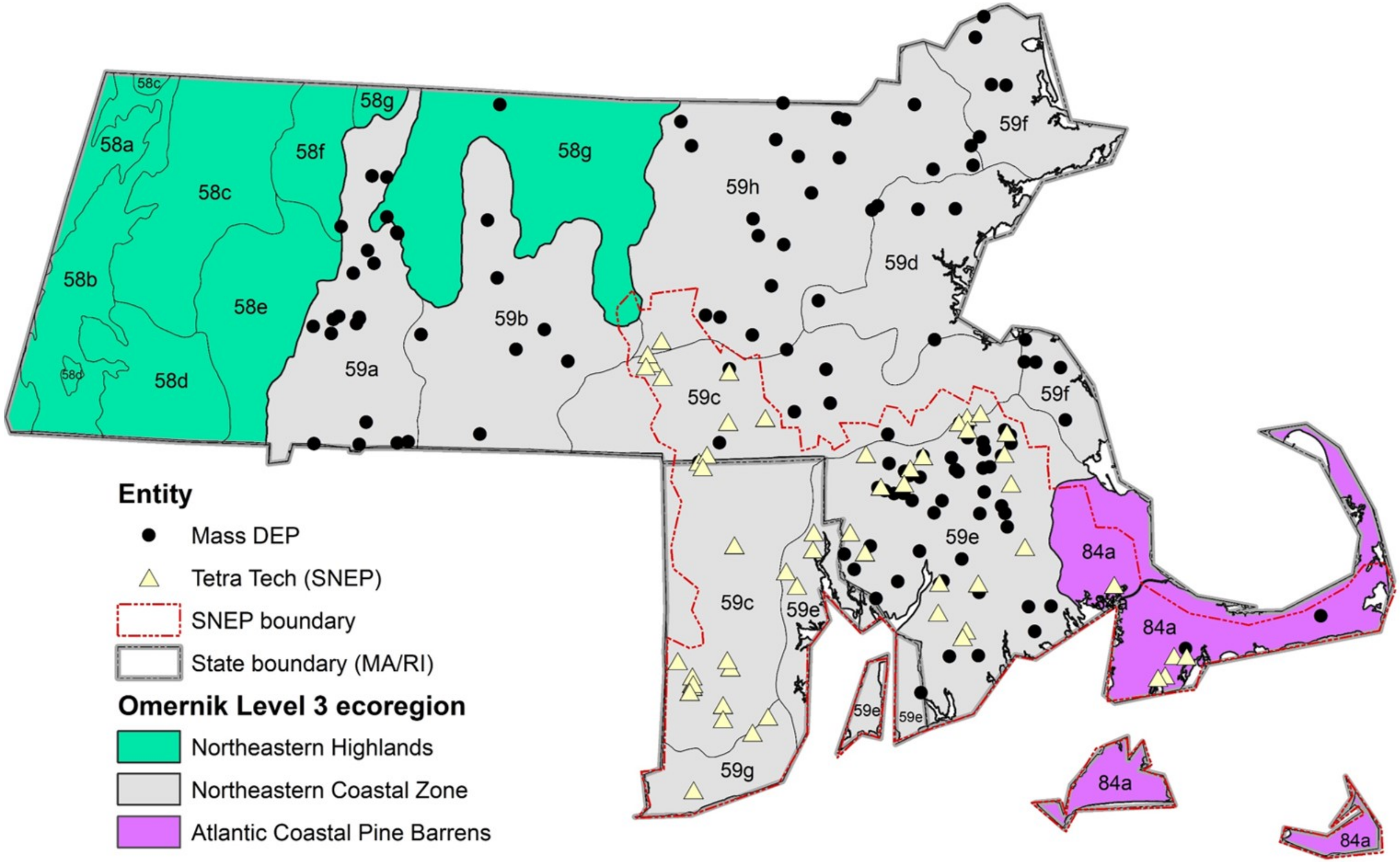


Figure 1. Locations of sites in the low gradient IBI development dataset (n=178 unique sites), coded by sampling entity (MassDEP or Tetra Tech), with Level 3 ecoregions as the backdrop. See Table 1 for Level 4 ecoregion code names.

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 at its peak. Major habitat types included submerged wood, submerged vegetation, undercut banks, overhanging vegetation, and hard bottom. 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 600µm 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 100-count 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 low gradient MassDEP/SNEP 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 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 565 taxa in the 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 MassDEP/SNEP dataset.

Attribute	Description	Categories	Sources*	Number of taxa with attribute assignments (out of 565)	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*	548	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	414	73.3%
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	260	46.0%
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	499	88.3%
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). However, those measures were not available for enough sites to include in the IBI calibration analyses.

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), and the lowest from vegetated margins/undercut banks (median = 2) (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 A 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.

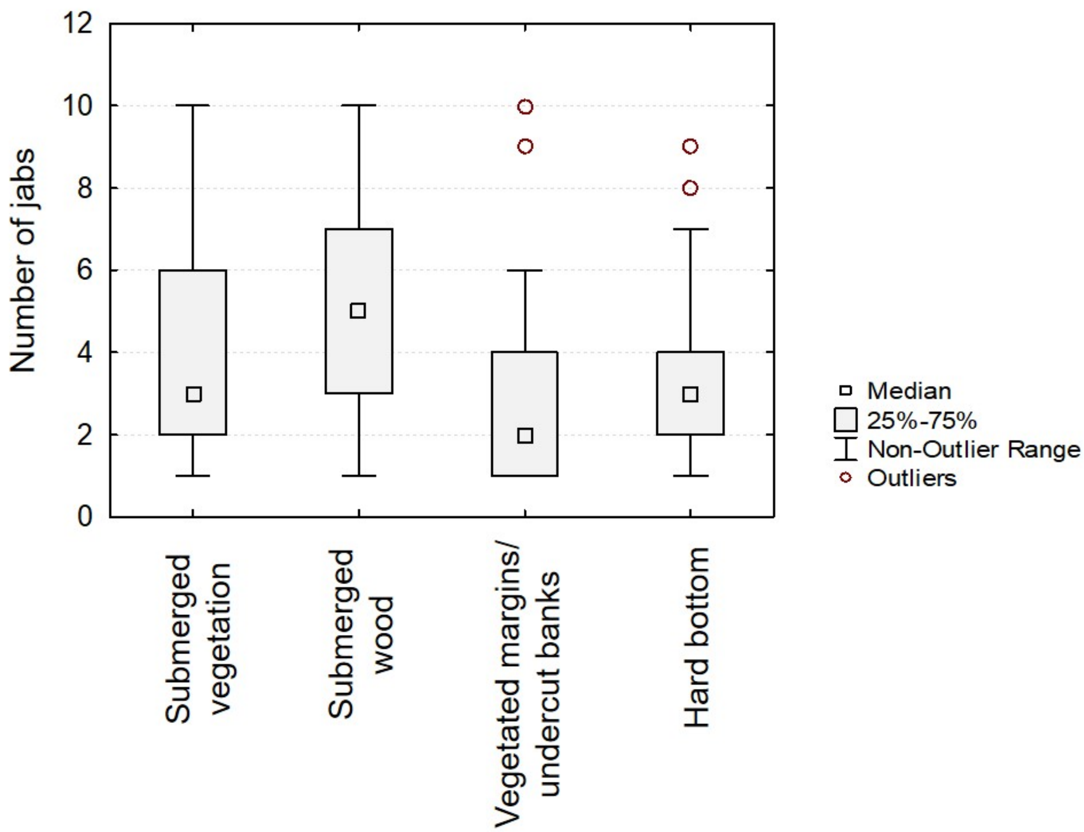


Figure 3. Distribution of jabs per site across the four major habitat types. A total of 10 jabs were taken per site. For more information on the habitat types, see Appendix A.

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.

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

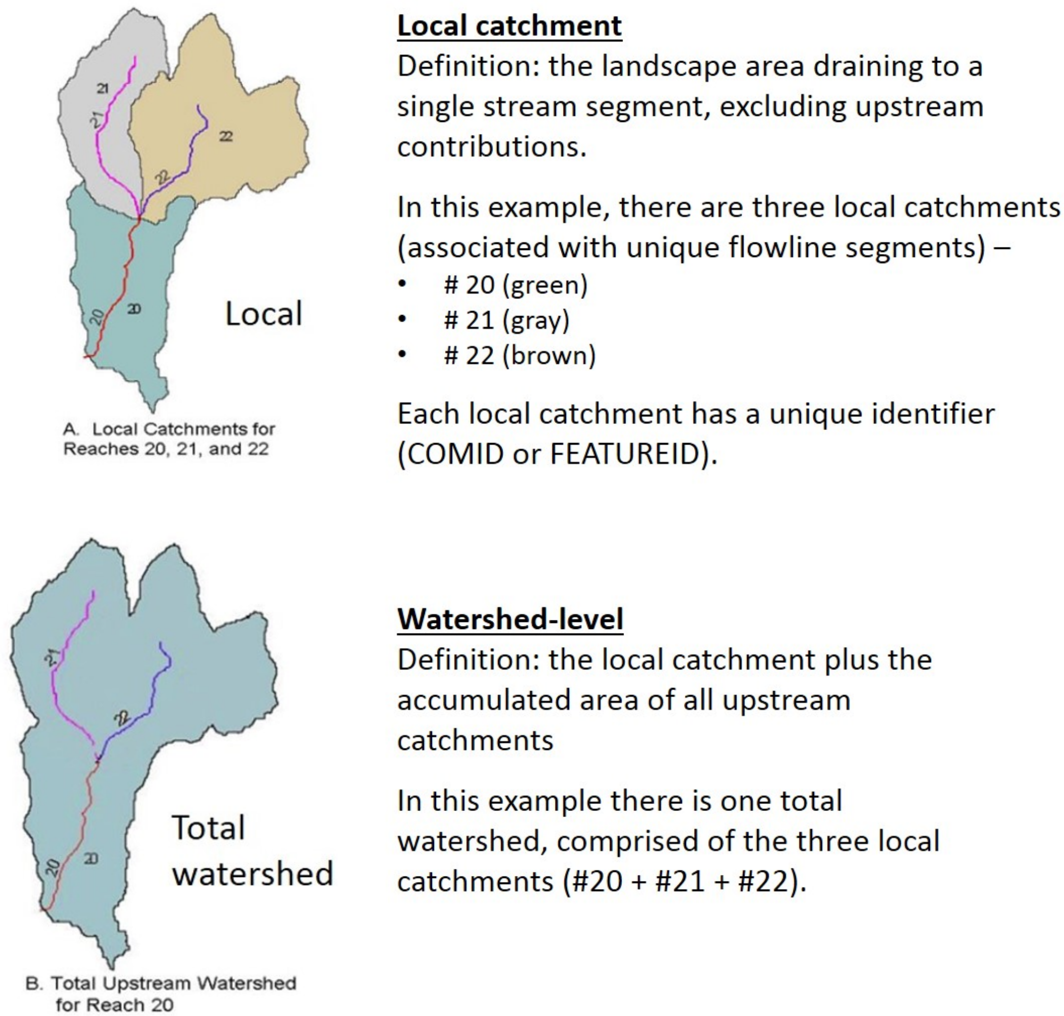


Figure 4. USEPA's StreamCat metrics (Hill et al. 2016) cover two spatial scales: local catchment and total watershed.

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 100-count riffle habitat IBIs (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, and manure application (Table 5). The low gradient disturbance index differed from the one used for the 100-count 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 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)

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).

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 178 sites, 41 were designated as reference sites, 41 as stressed sites, and 96 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, and elevation), and Attachment C contains the site list with preliminary and final disturbance category assignments.

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 100-count 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							

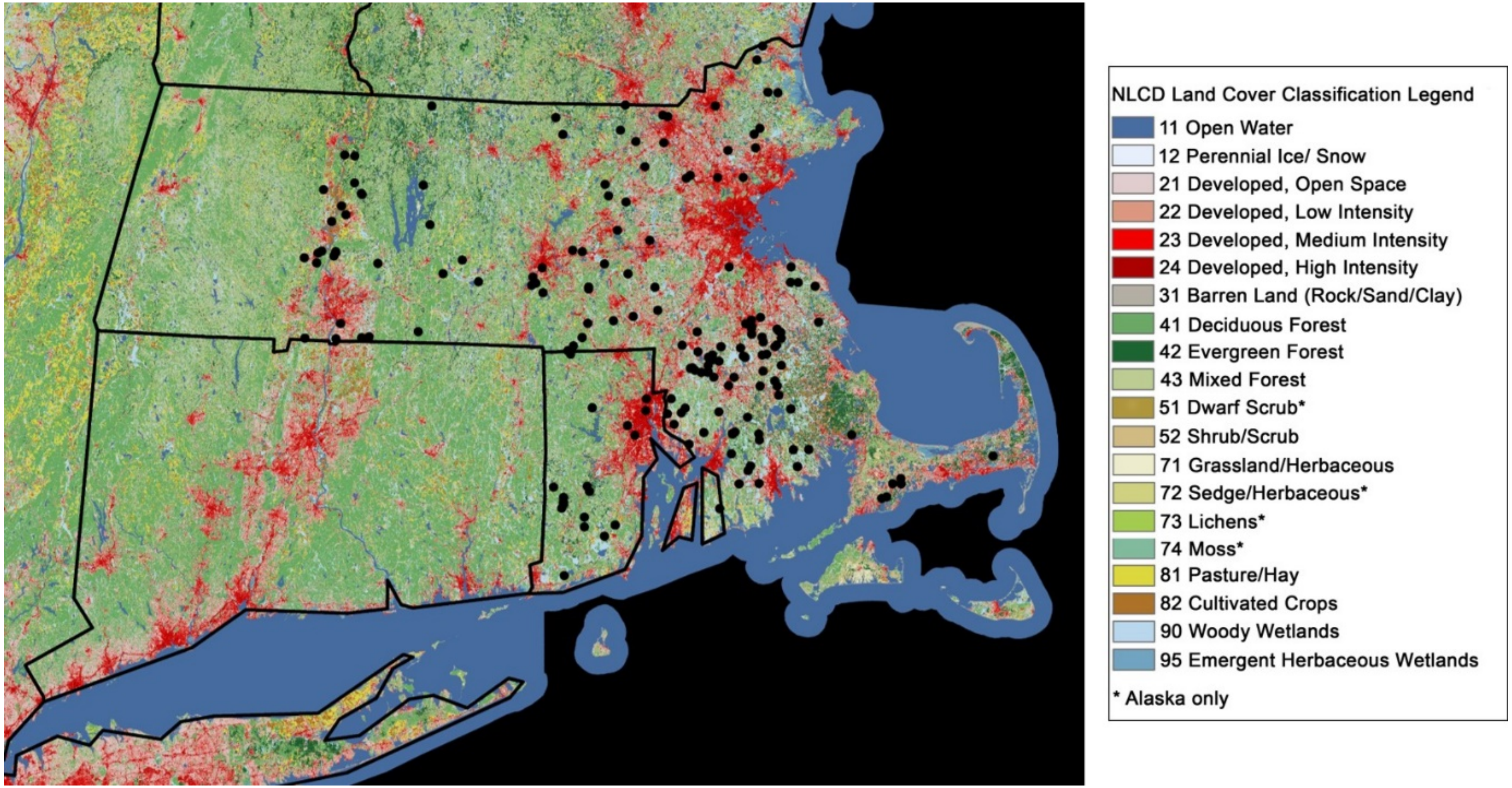


Figure 6. Low gradient streams are more prevalent in eastern Massachusetts, which has higher levels of urban land cover and human disturbance than western Massachusetts (source: NLCD 2016). Sample sites are shown as black dots.

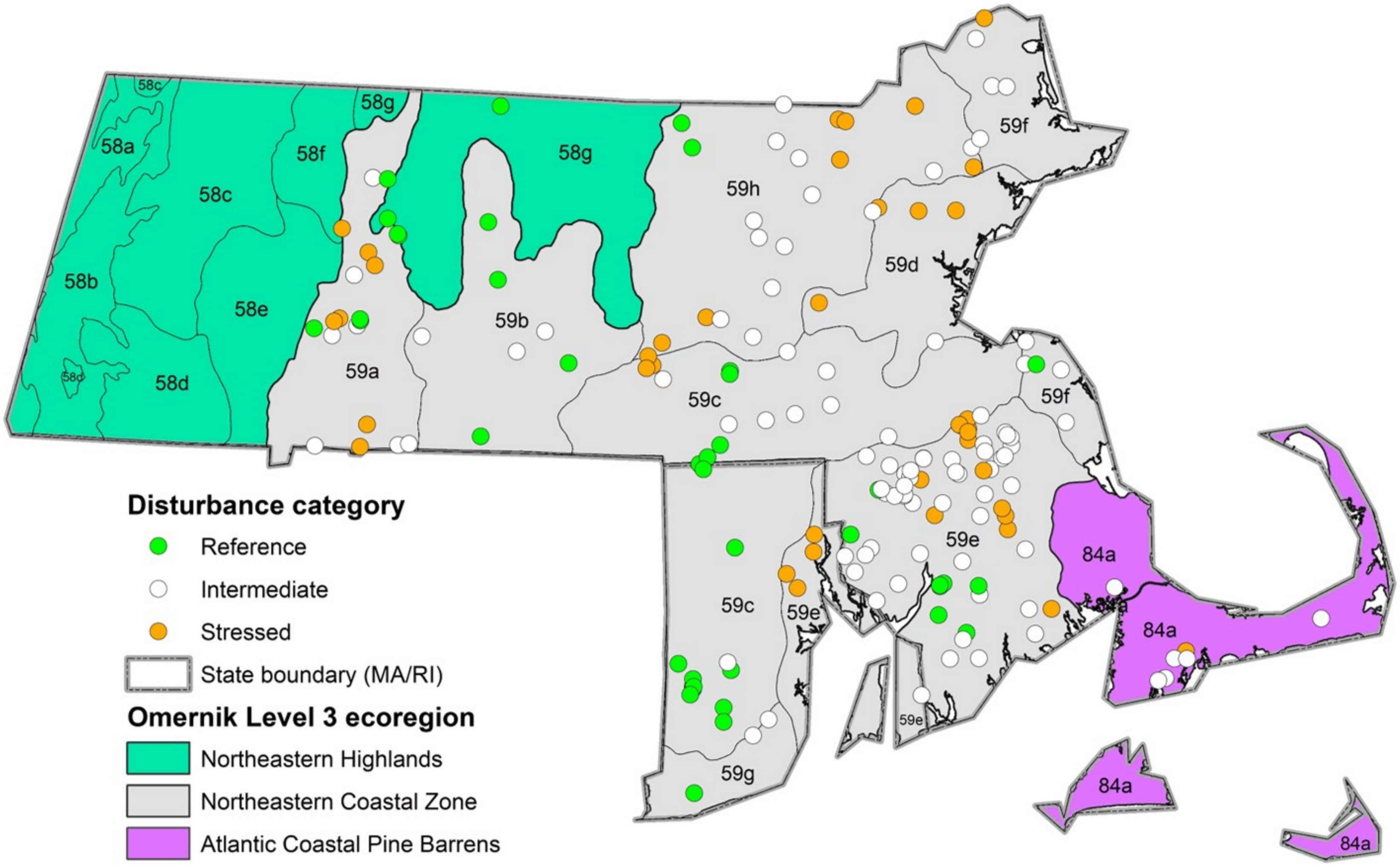


Figure 7. Spatial distribution of sites color-coded by disturbance category and overlaid on Level 3 and 4 ecoregions. See Table 1 for Level 4 ecoregion code names.

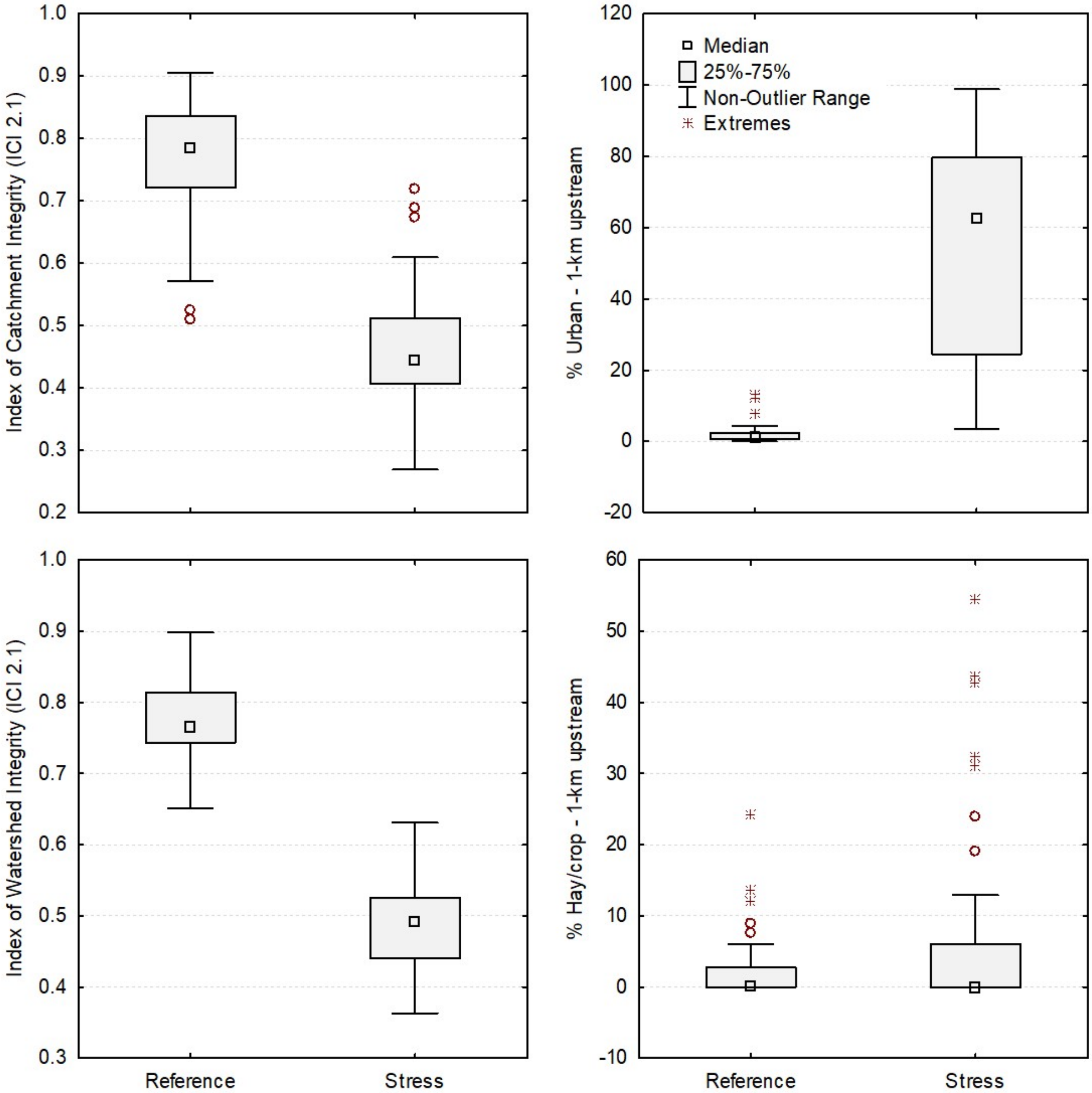


Figure 8. Box plots showing the range of disturbance represented in the reference (n=41) and stressed (n=41) sites, as measured by the ICI, IWI, percent urban, and percent agricultural land cover.

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 by 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)	2.8	91.1	0.7	238.0	0.7	346.4
Stream slope, 500m	0.00	2.94	0.00	1.76	0.00	2.94
Elevation (ft)	25.3	269.3	11.7	185.4	7.2	269.3
% wetland/open water	0.9	34.3	0.0	34.2	0.0	44.4
IWI	0.65	0.90	0.36	0.63	0.36	0.90
ICI	0.51	0.91	0.27	0.72	0.27	0.91
% urban	0.2	7.8	3.1	93.5	0.2	93.5
Road density	0.6	3.2	1.8	17.7	0.6	17.7

4.1 Exploratory Classification Analysis

The classification investigation proceeded through the 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. 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.

On the first axis of the NMS ordination, longitude and substrate characteristics were the natural variables that had potential for site classification. Percent forest, percent water and wetland, and percent urban cover were also correlated, but are inappropriate for classification because they could be directly influenced by human activities. Temporal variables (collection year and date) might help to explain differences in the current data set but are not reliable for extrapolation to future times. Substrate characteristics include the number of hard-bottom jabs (the habitat type sampled for macroinvertebrates) and percent muck-mud in the reach (estimated substrate areal percent).

Longitude is related to ecoregion and can be used as a continuous variable for classification whereas ecoregions could define categorical classes. It was evident that the level 4 ecoregions have some distinctive taxonomic characteristics. Longitude and hard-bottom jabs show a relationship with the first axis, but on the right side of the diagram, some sites span the gradients, making identification of class thresholds untenable.

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 first PCA axis was related to forest cover in steeper and larger watersheds versus eastern watersheds with more wetland cover. The metrics associated with the steeper, forested watersheds were related to taxa richness overall and in richness in specific trait groups (clingers, EPT, and intolerant taxa). Tolerant and non-insect individuals were associated with the eastern wetland sites.

On the second axis, warmer eastern streams were opposite northern, higher elevation streams. The northern streams also had more organic material (detritus). The warmer eastern streams had more sensitive insect individuals in contrast to the northern streams with more Diptera and short-lived, multivoltine individuals. The high percentage of midges are apparently in sites with more detritus. The relationships between the macroinvertebrate metrics and environmental variables were similar to those observed in the NMS presence/absence analysis. However, stream substrate was less important in the metric PCA than it was in the NMS analysis. Watershed land slope and annual air temperature were more important in the metric PCA than they were in the presence/absence ordination.

4.2 Classification Summary

Classification schemes related to Level 4 ecoregion, baseflow, and drainage area were considered but ruled out based on results from the NMS and PCA analyses. Level 4 ecoregions did not cluster distinctly in the ordinations. Moreover, defining site classes based on Level 4 ecoregions might be untenable because it would result in small sample sizes for index calibration. Patterns related to baseflow (BFI > 60) and watershed size (< 5 km^2) were evident in the NMS but were not strongly correlated with the biological metrics and did not show the same pattern in the PCA, so groupings based on these two metrics were not considered appropriate for site classification.

Of the continuous variables, the strongest associations were with percent forest, percent wetlands, and percent detritus. However, these are marginally-natural variables that are inappropriate for classification. Continuous variables that showed potential for classification included: mean annual air temperature (PRISM 1981-2010), latitude, longitude, elevation, watershed slope, 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

Index development consisted of the following steps:

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

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

5.1 Metric scoring

Evaluation and selection of metrics typically involve testing more metrics than end up in the final index. We calculated and evaluated over 150 metrics (Appendix C). 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, and mean land slope in the watershed; Section 4.3). 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 metric was available, but it did not require adjustment, then that similar metric might be a better choice. 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 used 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 understand and communicate.

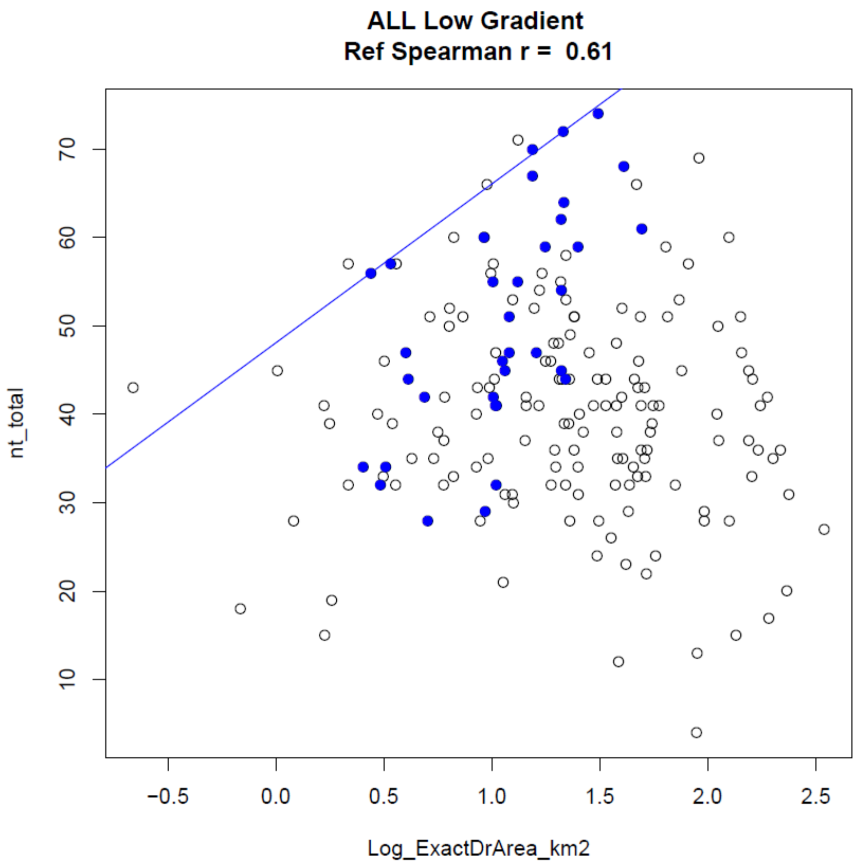


Figure 9. Bi-plot of total taxa (nt_total) and the log(10) transformation of drainage area, showing reference sites as solid blue markers, non-reference sites as open circles, and the reference 95th quantile regression line as a blue sloping line. Subjective limits to the regression adjustment were applied 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 25^{th} 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 75^{th} 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.

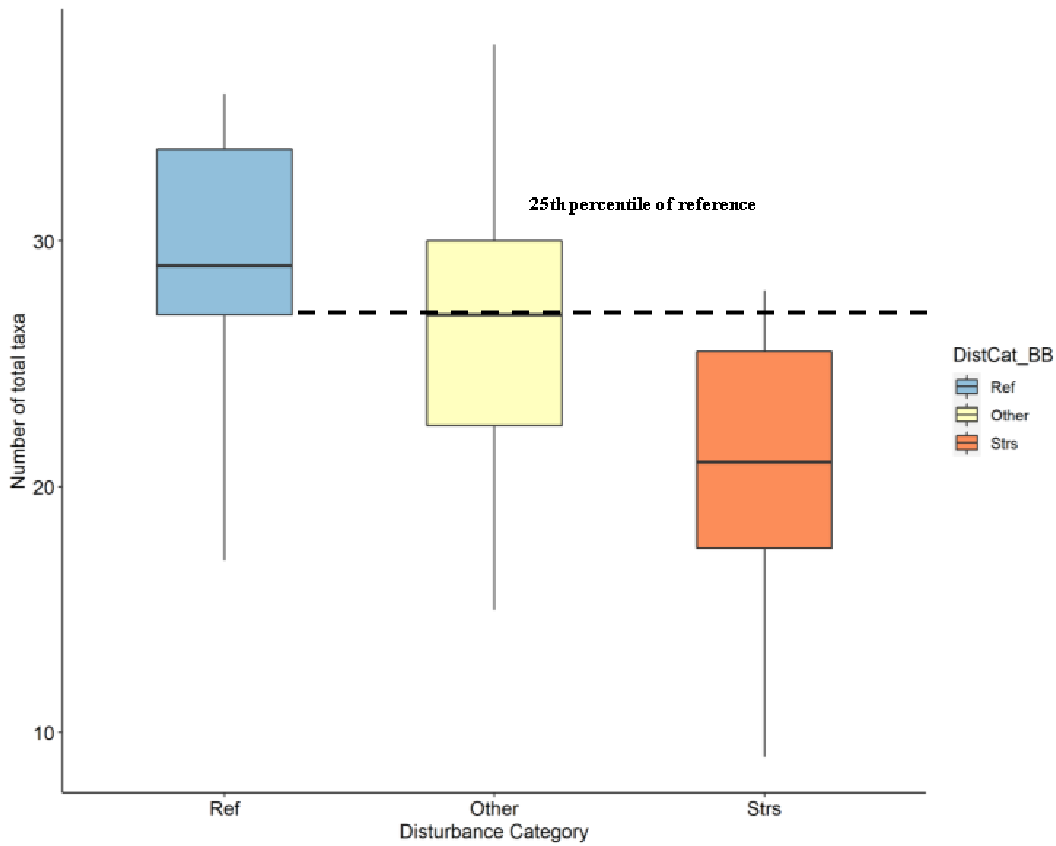


*Figure 10. Discrimination efficiency (DE). In this example, which uses the total number of taxa (a metric that decreases with stress), the 25th percentile of the reference distribution is used as the standard (and we calculate what percent of stressed sites were below that threshold; for example, if 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 Name	Metric Description	Category	Trend	5th	95th	Scoring Formula	*Z*-score	DE
SLog10_DrArea_km2.nt_total	number taxa - total adjusted for drainage area	RICH	Dec.	0.0	99.8	100*Metric/99.8	1.44	80.6
pt_Insect	percent taxa - Class Insecta	RICH	Dec.	0.0	94.9	100*Metric/94.9	2.66	87.1
pt_EPT	percent taxa - Orders Ephemeroptera, Plecoptera and Trichoptera	RICH	Dec.	0.0	34.1	100*Metric/34.1	2.04	93.5
nt_POET	number taxa - Orders Plecoptera, Odonanta, Ephemeroptera, and Trichoptera	RICH	Dec.	0.0	20.9	100*Metric/20.9	2.05	100.0
nt_Trich	number taxa - Order Trichoptera	RICH	Dec.	0.0	10.0	100*Metric/10	1.85	87.1
pt_NonIns	percent taxa - not Class Insecta	RICH	Inc.	5.1	44.9	100*(44.9-Metric)/39.8	-2.66	87.1
pi_EPT	percent individuals - Orders Ephemeroptera, Plecoptera and Trichoptera	COMP	Dec.	0.0	48.0	100*Metric/48	1.47	87.1
pi_OET	percent individuals - Orders Odonata, Ephemeroptera, and Trichoptera	COMP	Dec.	0.0	49.0	100*Metric/49	1.27	74.2
pi_TricNoHydro	percent individuals - Order Trichoptera and not Family Hydropsychidae	COMP	Dec.	0.0	16.3	100*Metric/16.3	0.87	80.6
pi_NonIns	percent individuals - Class not Insecta	COMP	Inc.	1.9	80.4	100*(80.4-Metric)/78.5	-1.82	80.6
pi_dom03	percent individuals - three most dominant taxa	EVENN	Inc.	24.1	77.7	100*(77.7-Metric)/53.7	-1.35	80.6
nt_ffg_pred	number taxa - Functional Feeding Group (FFG) - predator (PR)	FFG	Dec.	0.0	16.0	100*Metric/16	1.74	80.6
pi_ffg_pred	percent individuals - Functional Feeding Group (FFG) - predator (PR)	FFG	Dec.	0.0	22.8	100*Metric/22.8	0.95	80.6
pt_habit_climb	percent taxa - Habit - climbers (CB)	HABIT	Inc.	3.4	20.4	100*(20.4-Metric)/17	-1.72	67.7
pi_habit_swim	percent individuals - Habit - swimmers (SW)	HABIT	Dec.	0.0	13.8	100*Metric/13.8	0.79	90.3
x_HBI	Hilsenhoff Biotic Index	TOLER	Inc.	4.7	7.2	100*(7.2-Metric)/2.5	-2.16	100.0
pt_tv_intol	percent taxa - intolerant	TOLER	Dec.	0.0	11.1	100*Metric/11.1	1.78	100.0
pt_tv_toler	percent taxa - tolerant	TOLER	Inc.	3.6	34.3	100*(34.3-Metric)/30.7	-6.19	100.0
pt_volt_multi	percent taxa - multivoltine	VOLT	Inc.	10.6	31.2	100*(31.2-Metric)/20.6	-1.86	80.6
pt_volt_semi	percent taxa - semivoltine	VOLT	Dec.	0.0	11.6	100*Metric/11.6	1.34	90.3

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. See Table 8 for metric descriptions.

Metric #	Metric Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	SLog10_ DrArea_km2.nt_total	1																			
2	pt_Insect	0.48	1																		
3	pt_EPT	0.28	0.61	1																	
4	nt_POET	0.70	0.64	0.80	1																
5	nt_Trich	0.61	0.57	0.78	0.85	1															
6	pt_NonIns	0.48	1.00	0.61	0.64	0.57	1														
7	pi_EPT	0.32	0.47	0.71	0.65	0.61	0.47	1													
8	pi_OET	0.30	0.38	0.65	0.63	0.57	0.38	0.95	1												
9	pi_TricNoHydro	0.35	0.34	0.65	0.62	0.72	0.34	0.61	0.60	1											
10	pi_NonIns	0.48	0.73	0.46	0.55	0.50	0.73	0.53	0.50	0.31	1										
11	pi_dom03	0.67	0.47	0.36	0.59	0.58	0.47	0.39	0.42	0.46	0.51	1									
12	nt_ffg_pred	0.70	0.22	0.16	0.56	0.39	0.22	0.19	0.24	0.23	0.20	0.36	1								
13	pi_ffg_pred	0.52	0.17	0.02	0.32	0.21	0.16	0.00	0.14	0.18	0.24	0.49	0.66	1							
14	pt_habit_climb	0.18	0.51	0.29	0.17	0.26	0.51	0.24	0.15	0.20	0.34	0.19	-0.09	-0.08	1						
15	pi_habit_swim	0.35	0.29	0.40	0.50	0.34	0.29	0.40	0.42	0.26	0.20	0.19	0.44	0.20	-0.02	1					
16	x_HBI	0.51	0.65	0.50	0.61	0.55	0.65	0.56	0.53	0.35	0.86	0.45	0.34	0.31	0.26	0.24	1				
17	pt_tv_intol	0.51	0.57	0.49	0.59	0.50	0.57	0.36	0.29	0.31	0.53	0.34	0.41	0.28	0.30	0.32	0.60	1			
18	pt_tv_toler	0.53	0.75	0.64	0.73	0.69	0.75	0.56	0.53	0.48	0.60	0.44	0.40	0.25	0.31	0.40	0.67	0.64	1		
19	pt_volt_multi	0.31	0.07	0.36	0.44	0.47	0.07	0.31	0.32	0.43	0.03	0.17	0.47	0.23	-0.01	0.36	0.16	0.36	0.40	1	
20	pt_volt_semi	0.36	0.51	0.44	0.50	0.51	0.51	0.36	0.34	0.41	0.51	0.36	0.25	0.22	0.25	0.15	0.58	0.47	0.62	0.28	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. Twenty 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 index alternative 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 were unfamiliar to MassDEP biologists and they did not have plainly understandable response mechanisms. The MassDEP workgroup was tasked to reduce the number of index alternatives to approximately twenty and present these as options to the working group. Their screening and exclusion criteria are summarized in Table 10.

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 #	Models with these criteria were eliminated	Remaining Models
1	DE < 100	13988
2	*Z*-score > -2.5	4153
3	Number of Metrics < 5	4147
4	Metric Categories < 5	3739
5	Insect/Non-Insect Metrics > 1	2698
6	Contains both pt_EPT and pi_EPT	2479
7	Contains both nt_ffg_pred and pi_ffg_pred	1982
8	Contains both pi_habit_swim and pt_habit_climb	1594
9	Contains both pt_tv_toler and pt_tv_intoler	747
10	Contains both pt_volt_semi and pt_volt_multi	586
11	Contains SLog10_DrArea km2.nt_total	354
12	Contains no FFG metrics	319
13	Contains no Tolerance metrics	311
14	Reference 25th percentile - Stressed 75th percentile < 20	33

The index alternatives chosen by MassDEP were then compared to the short-list of index alternatives from the SNEP project's all-subset analysis. Many of the models considered by MassDEP were also considered in the SNEP project. Therefore, a short-list of index alternatives that could be applicable in both projects was selected (Table 11). The resulting subset of index alternatives had similar performance statistics; therefore, the final selection process involved subjective decisions on metric preference and performance. In the end, the workgroup decided to pick indices with metrics that were most familiar (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 index selection based on empirical performance and ecological characteristics of the individual and combined metrics. They also selected an index that was a top choice for both the MassDEP and SNEP projects. The final choice was Model 6_13784, which included six metrics (Tables 11 & 12).

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. 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 MA/SNEP and 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 formulae 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 macroinvertebrate 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

5.4 Final index selection and performance

The team of MassDEP 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 MassDEP 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=0.77), percent POET taxa (pt_POET) (rho=-0.66), and percent semi-voltine taxa (pt_volt_semi) (rho=-0.65) (Table 15); however, the workgroup did not think that these metrics were fundamentally redundant with one another but instead evaluated unique components of the macroinvertebrate community. All other metrics have pairwise correlations of less than 0.65. 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.

	pi_OET	pt_ffg_pred	pt_NonIns	pt_POET	pt_tv_toler	pt_volt_semi
pi_OET	1					
pt_ffg_pred	0.06	1				
pt_NonIns	-0.41	0.00	1			
pt_POET	0.63	0.17	-0.60	1		
pt_tv_toler	-0.56	-0.19	0.77	-0.66	1	
pt_volt_semi	0.41	0.12	-0.54	0.53	-0.65	1

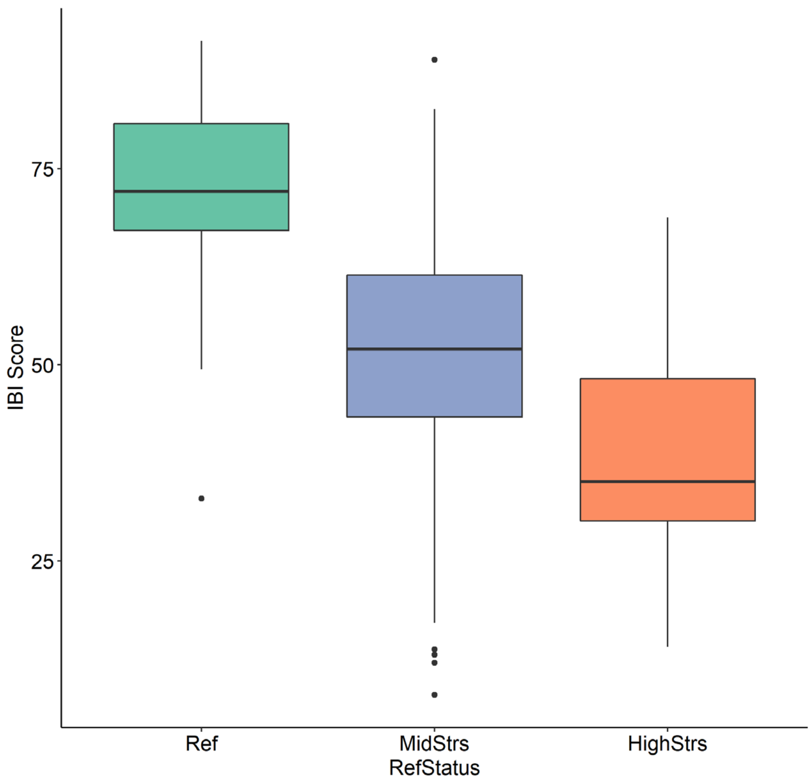


Figure 11. Distribution of IBI scores across disturbance categories, reference (Ref), intermediate (MidStrs), and stressed (HighStrs).

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 = 0.49) and IWI (rho = 0.59) and had a strong negative correlation with percent urban land cover (rho = -0.63) (Figure 12). IBI scores were weakly correlated with percent agriculture land cover (rho = 0.06) but most sites had low percent agriculture (<10%) (Figure 12).

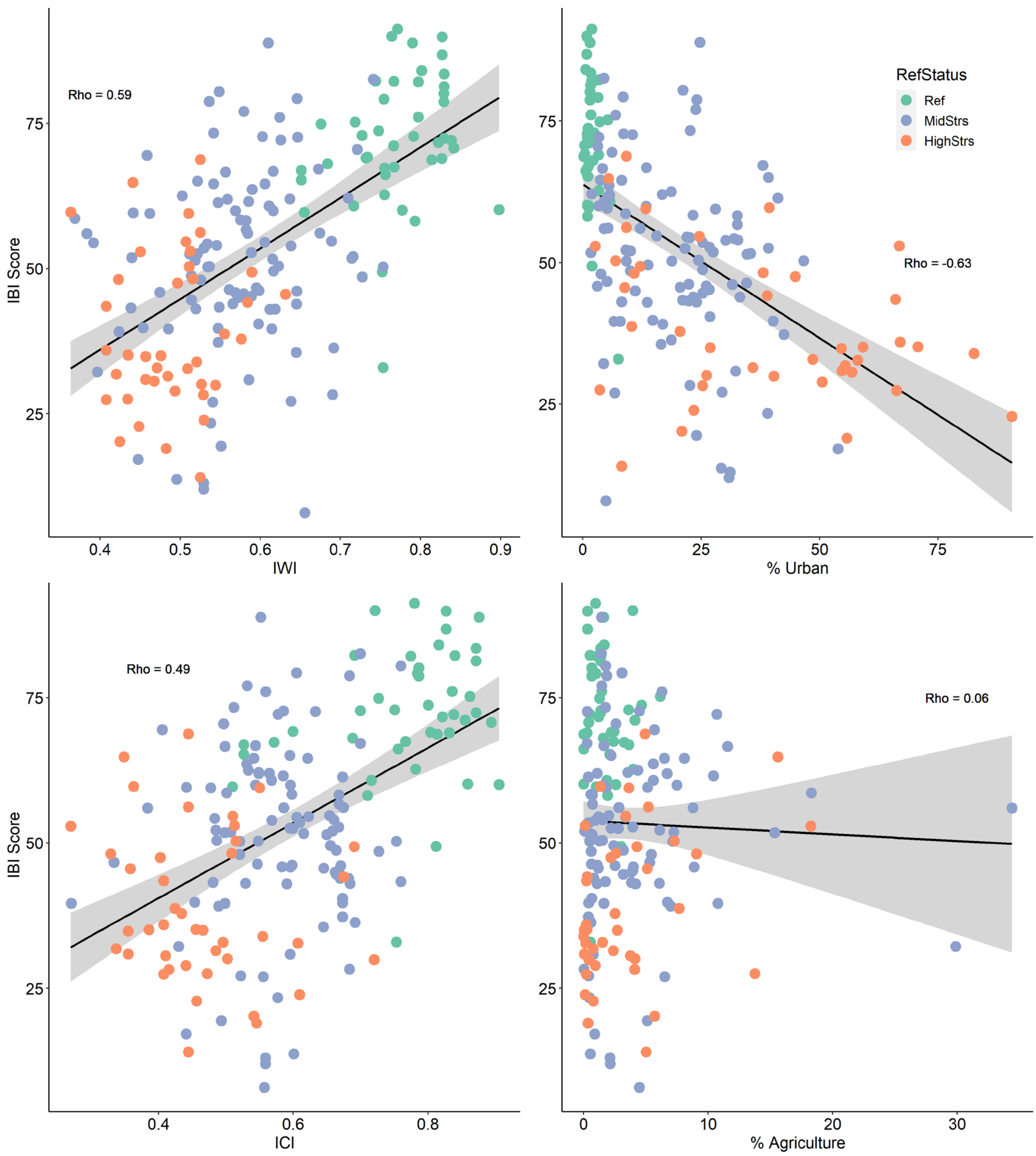


Figure 12. Relationship between the low gradient, multihabitat IBI vs. IWI (upper left), ICI (lower left), percent urban (upper right) and percent agriculture (lower right). The black line is the regression line and the rho value is the Spearman rank correlation coefficient.

5.5 Index verification

The low gradient, multihabitat IBI was validated through comparison of calibrated index values with stressor indicators that were not used in defining the index calibration stressor gradient. Relationship 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. The water quality variables were only available for the samples collected by Tetra Tech field crews using SNEP protocols in 2019.

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, pool variability, riparian vegetation, and sediment deposition (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.

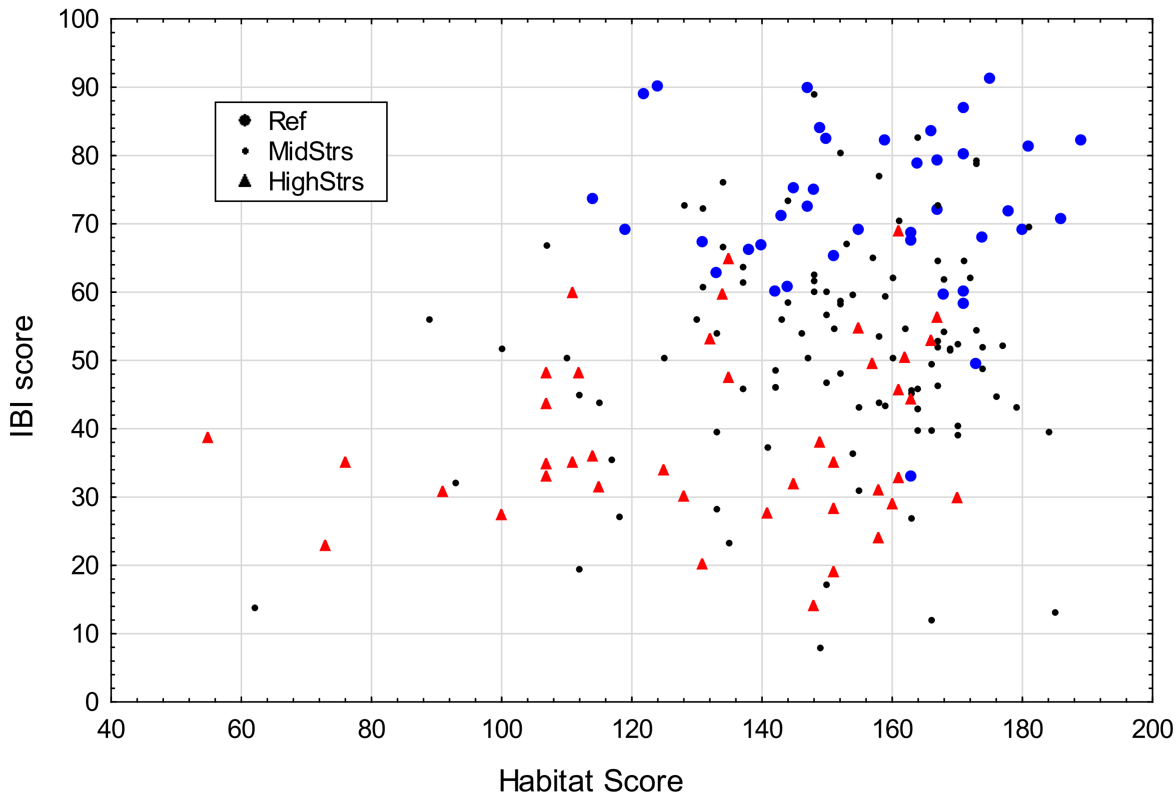


Figure 13. IBI scores in relation to RBP total habitat scores, marked by disturbance category.

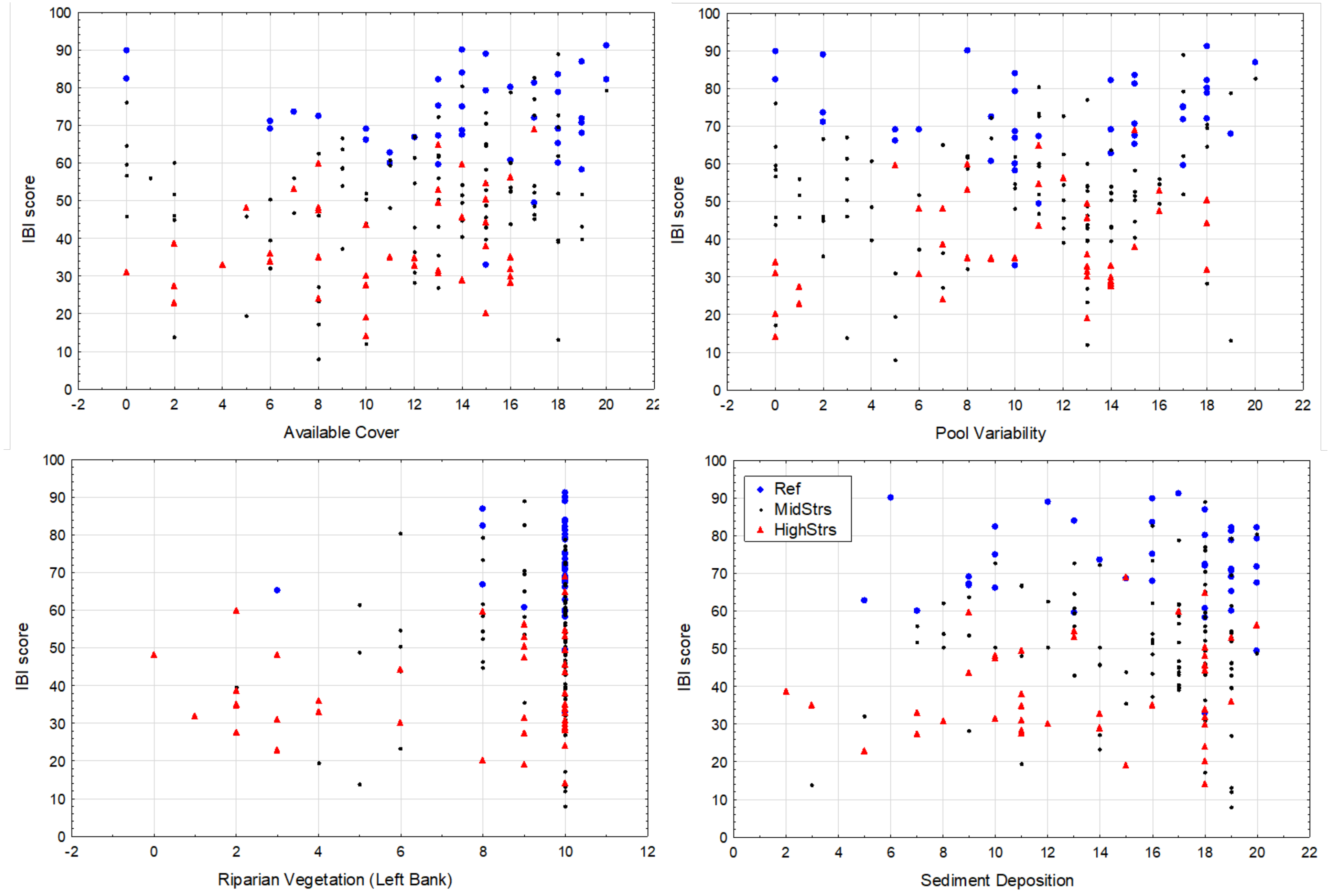


Figure 14. IBI scores in relation to effective habitat variables, including available cover, pool variability, riparian vegetation, and sediment deposition.

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 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).

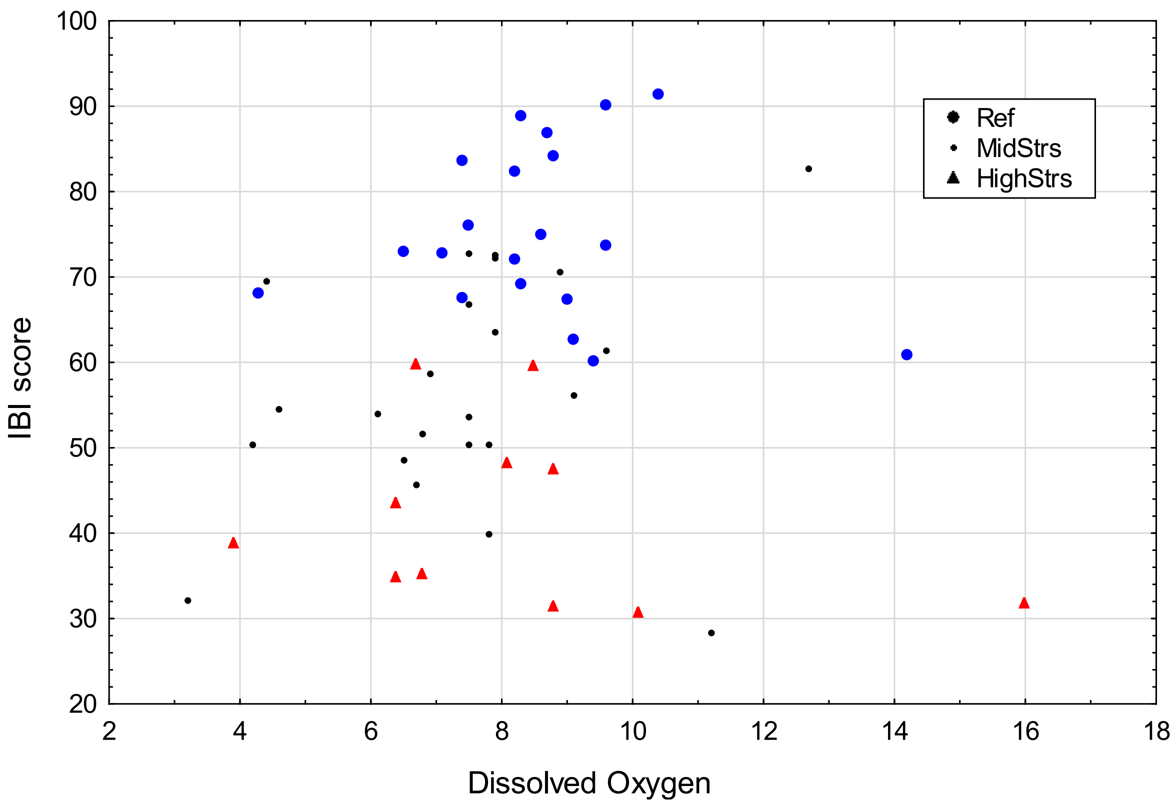


Figure 15. IBI scores in relation to dissolved oxygen (DO) in sites with DO data, marked by disturbance category.

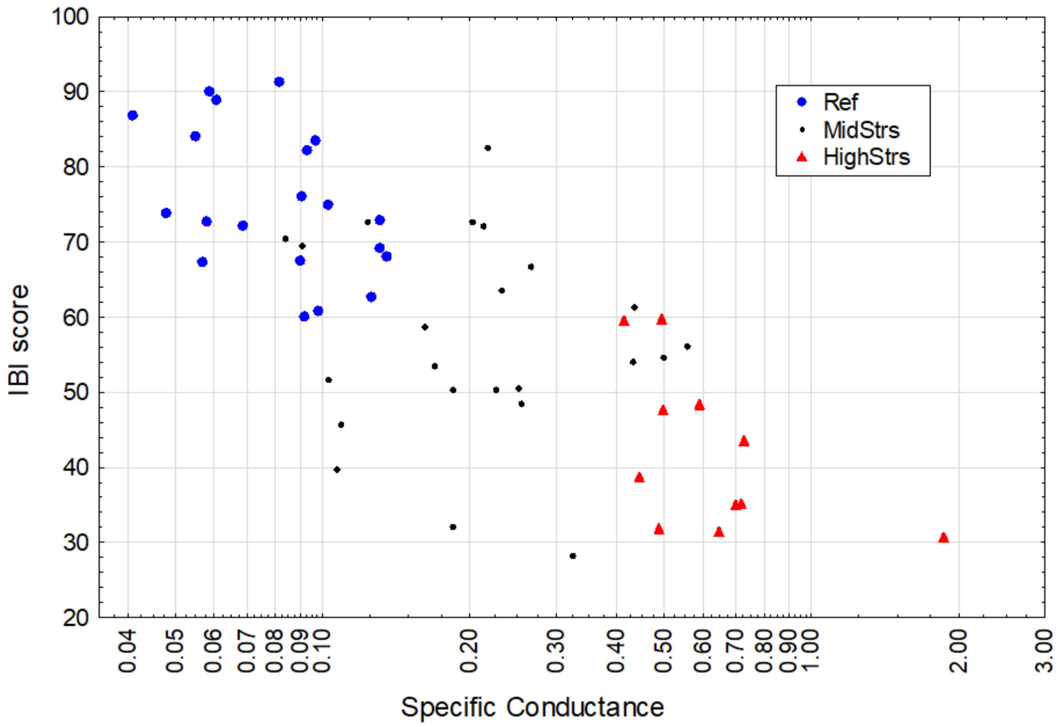


Figure 16. IBI scores in relation to conductivity (on a log-transformed axis) at sites with conductivity data, marked by disturbance category.

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.

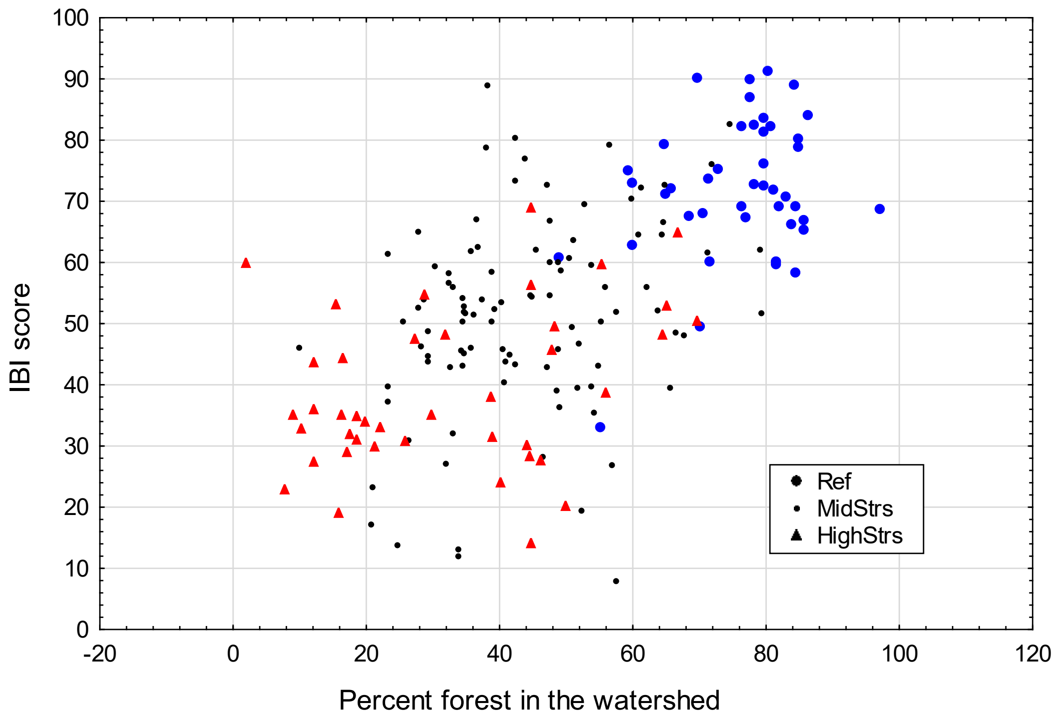


Figure 17. IBI scores in relation to percent forest cover (watershed-scale), marked by disturbance category.

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.

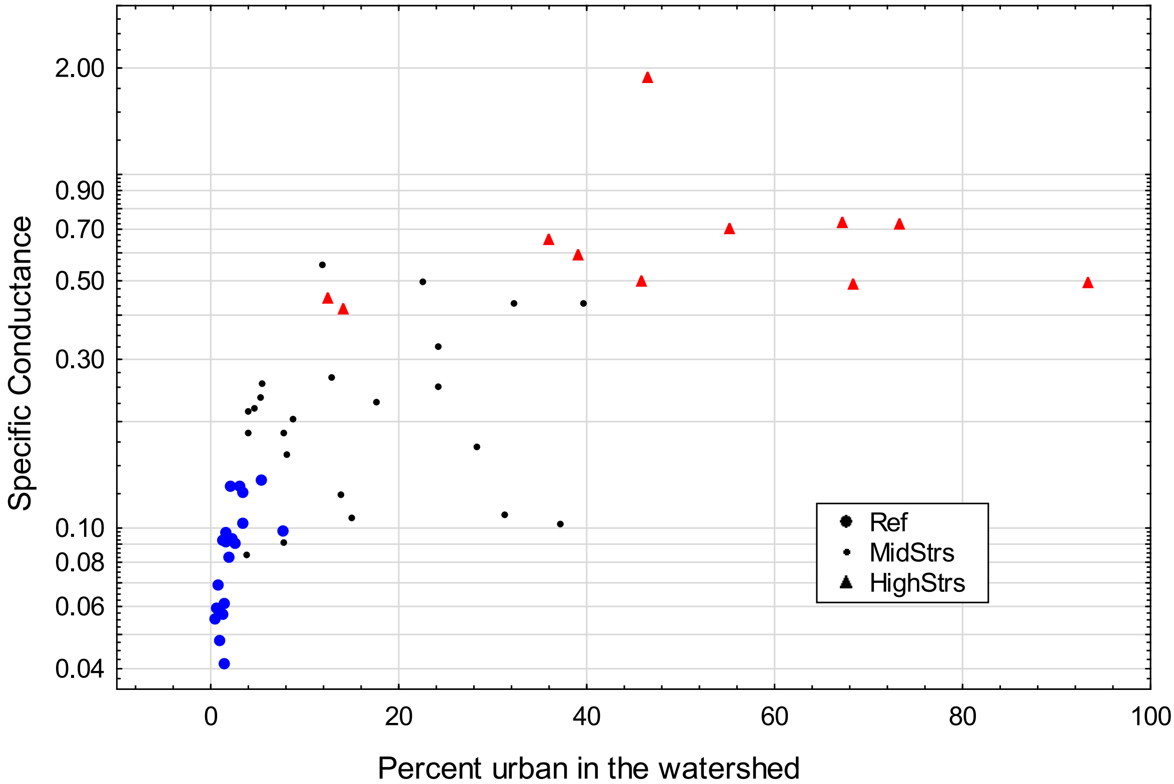


Figure 18. Conductivity (on a log-transformed axis) at sites with conductivity data in relation to percent urban land uses in the watershed, marked by disturbance category.

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 region are acidic, especially in the SNEP region. 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.

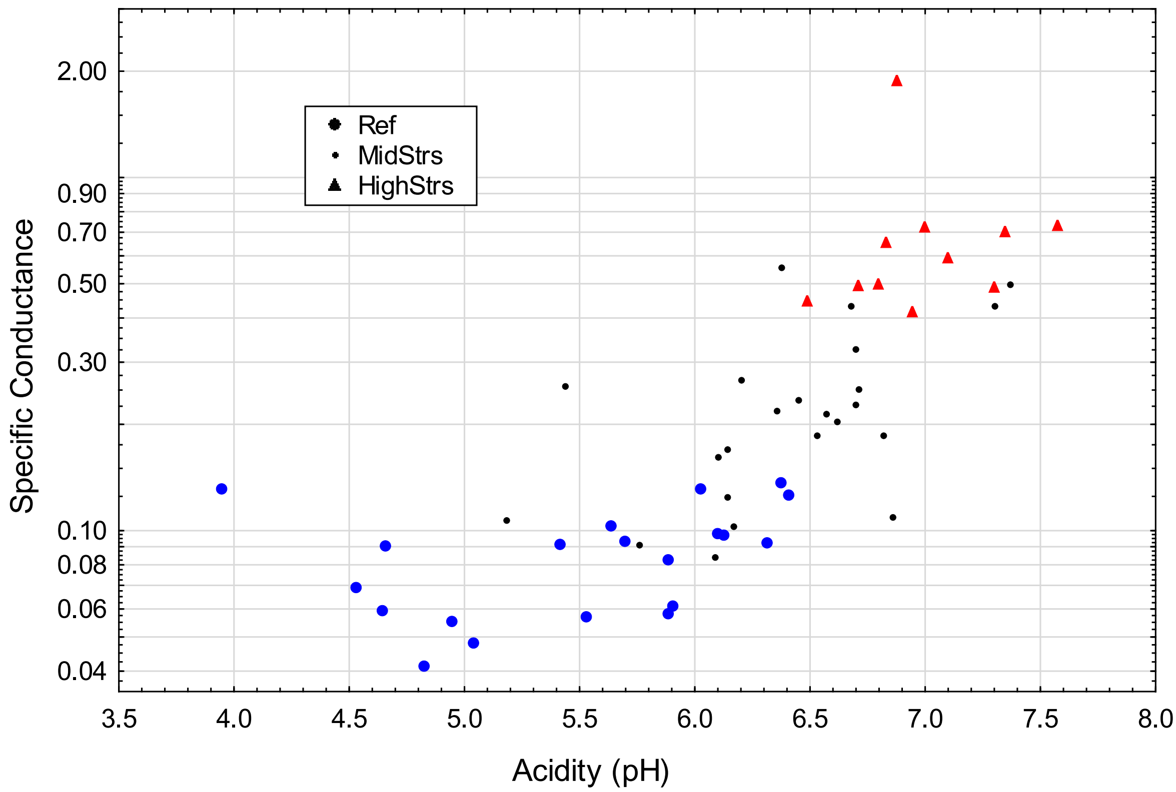


Figure 19. Conductivity (on a log-transformed axis) at sites with conductivity data in relation to pH, marked by disturbance category.

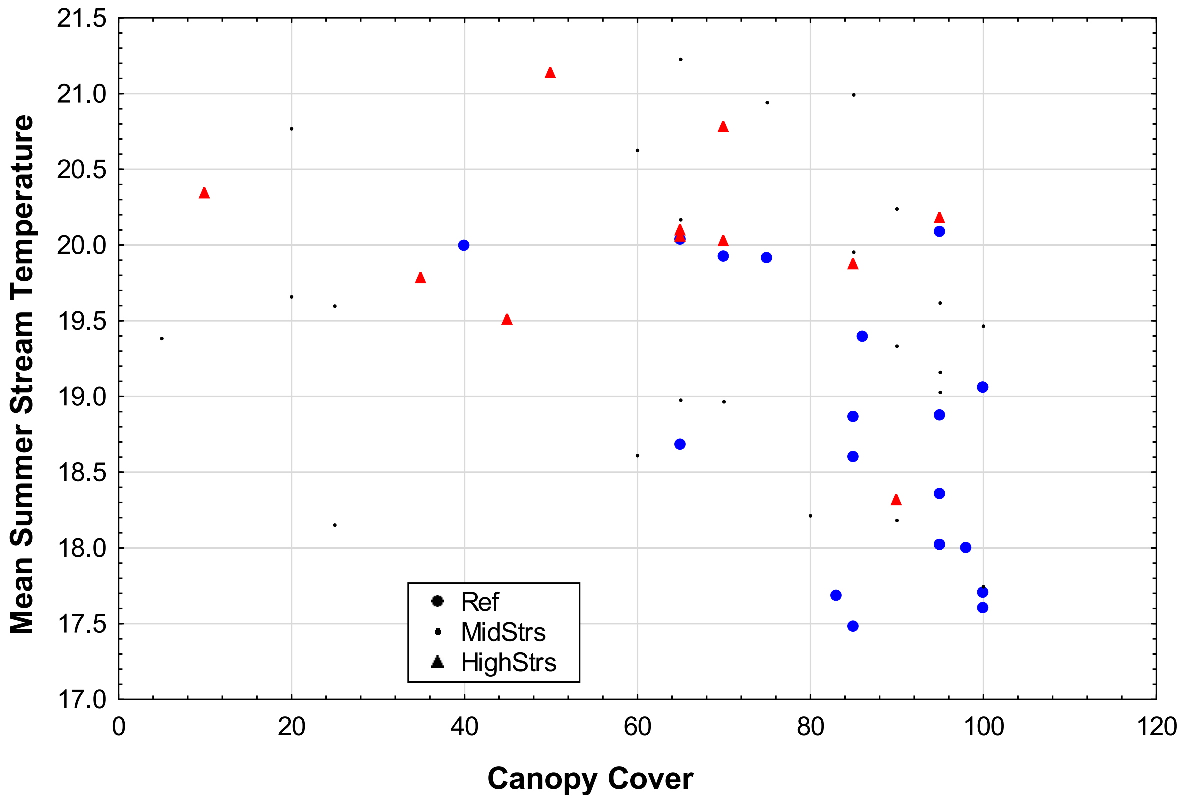


Figure 20. Mean summer stream temperature (MSST) in relation to percent canopy cover, marked by disturbance category.

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 24). 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 & 24).

Though classification analysis indicated possible differences in reference sample composition between the SNECPAH and NBL Level 4 ecoregions, reference IBI score distributions in these two ecoregions overlapped (Figure 25). IBI scores in reference sites that were further north and west in Massachusetts had somewhat lower values, though mostly >60 points. In addition, within each Level 4 ecoregion, IBI scores decreased with increasing stress and in comparison to reference scores. IBI scores in the Lower Worcester Plateau/Eastern Connecticut Upland did not decrease as expected, though only moderately stressed sites were compared to reference, not highly stressed sites.

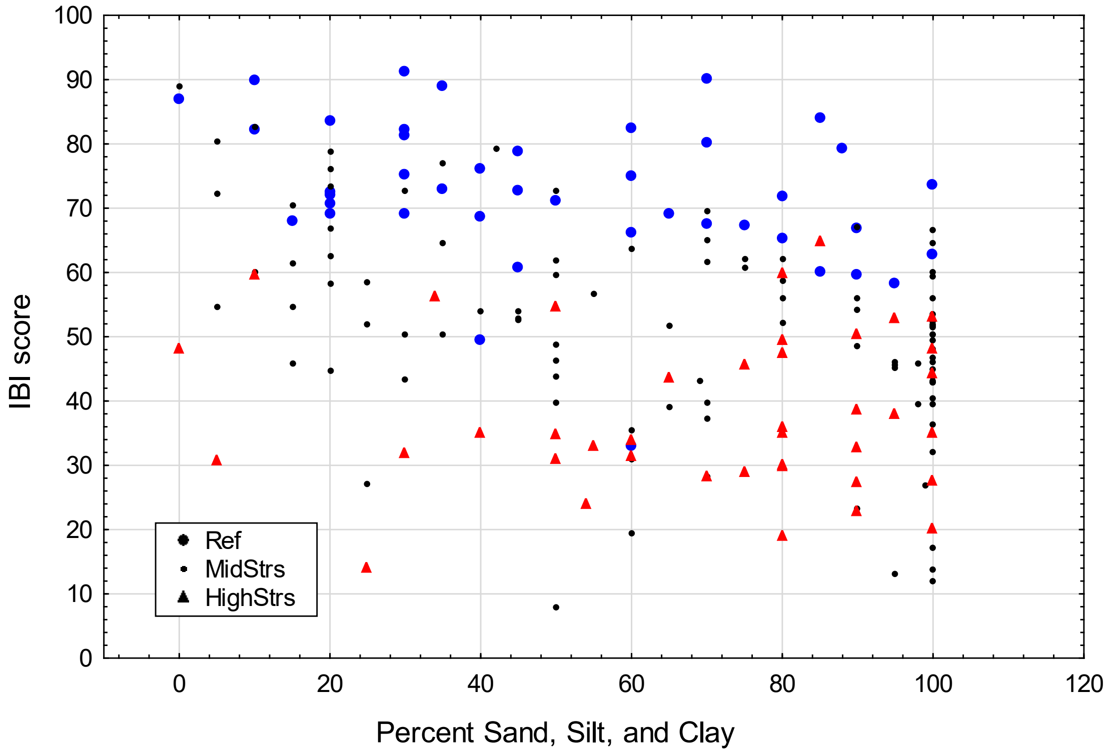


Figure 21. Percent sand, silt, and clay substrates in relation to IBI scores, marked by disturbance category.

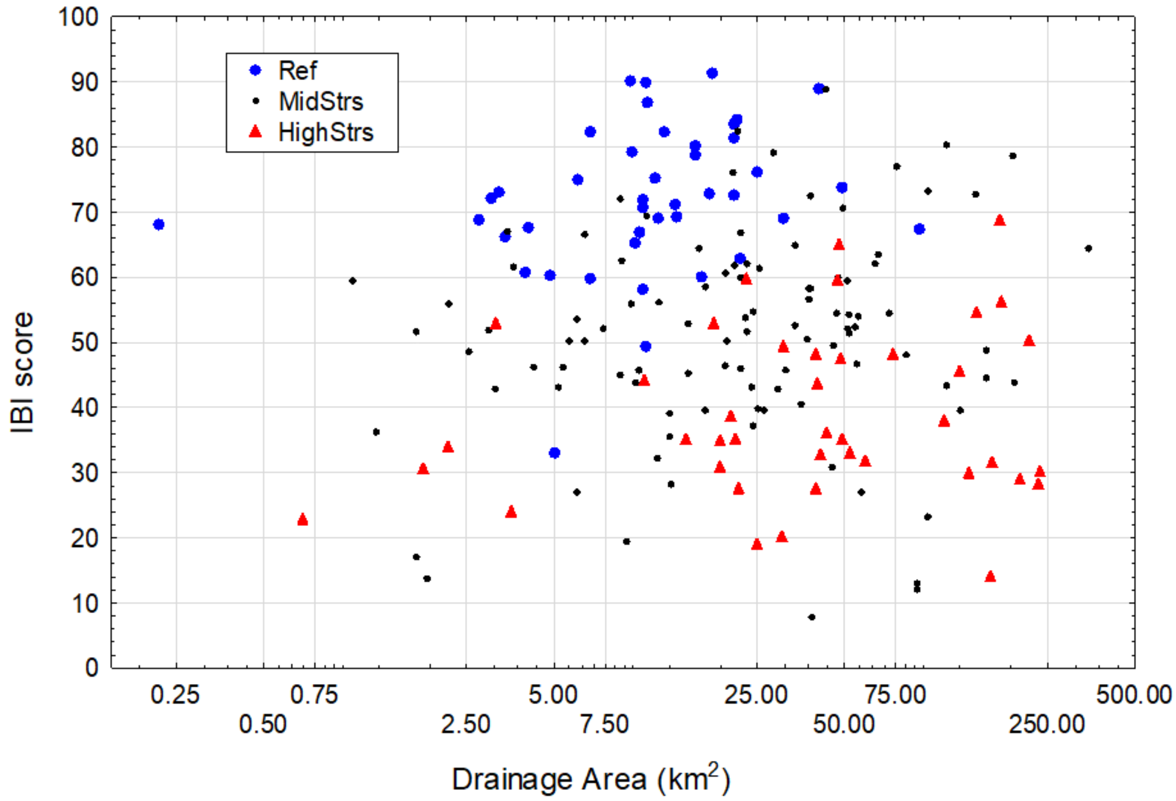


Figure 22. Site drainage area (on a log-transformed axis) in relation to IBI scores, marked by disturbance category.

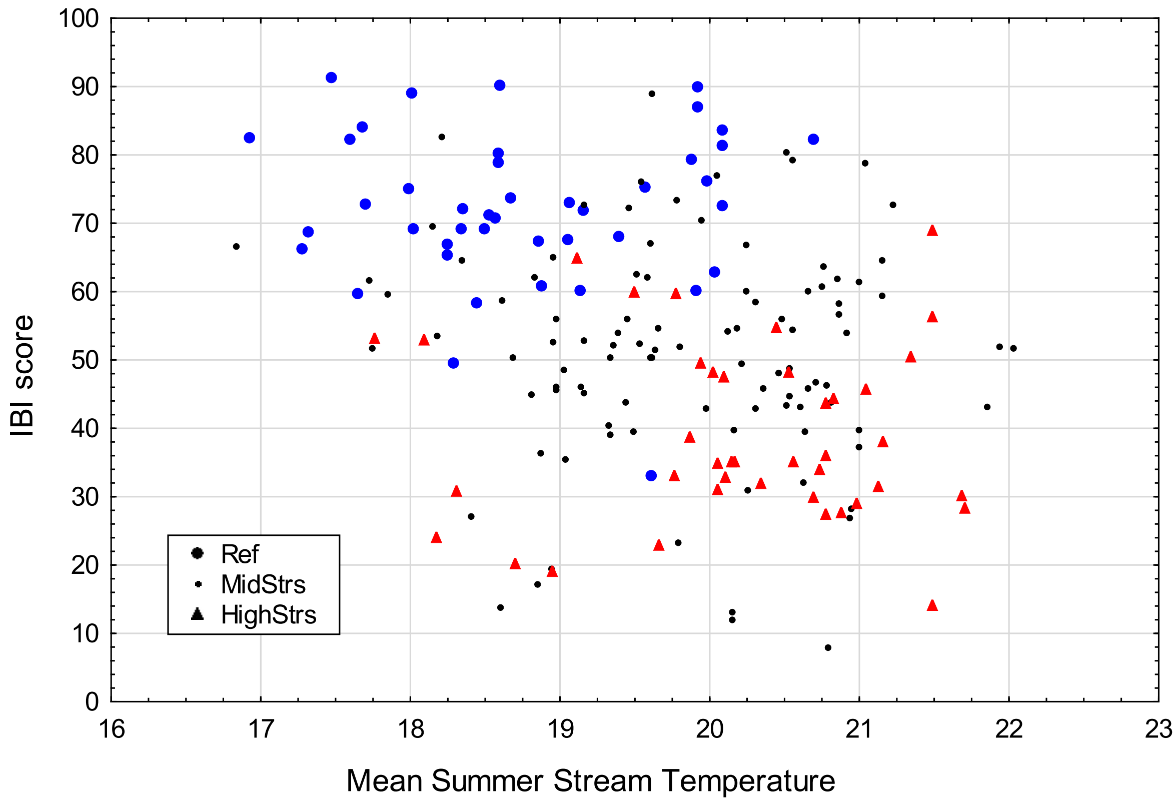


Figure 23. Mean Summer Stream Temperature (MSST) in relation to IBI scores, marked by disturbance category.

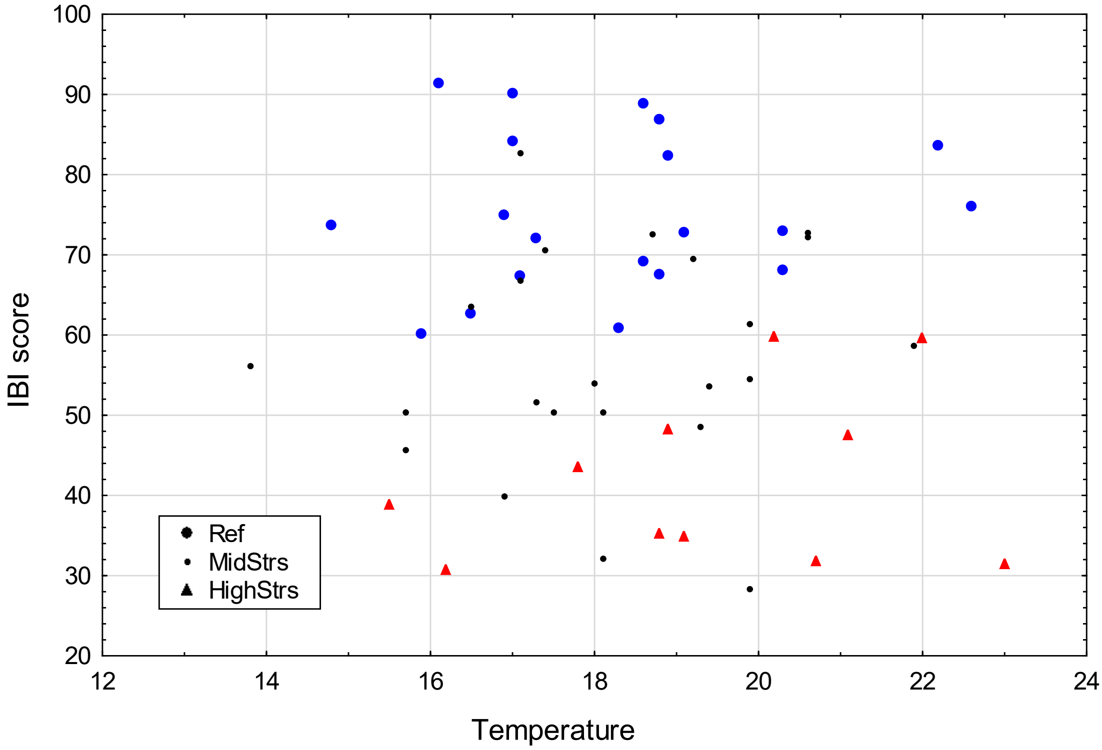


Figure 24. In situ (measured) stream temperature (where available), in relation to IBI scores, marked by disturbance category.

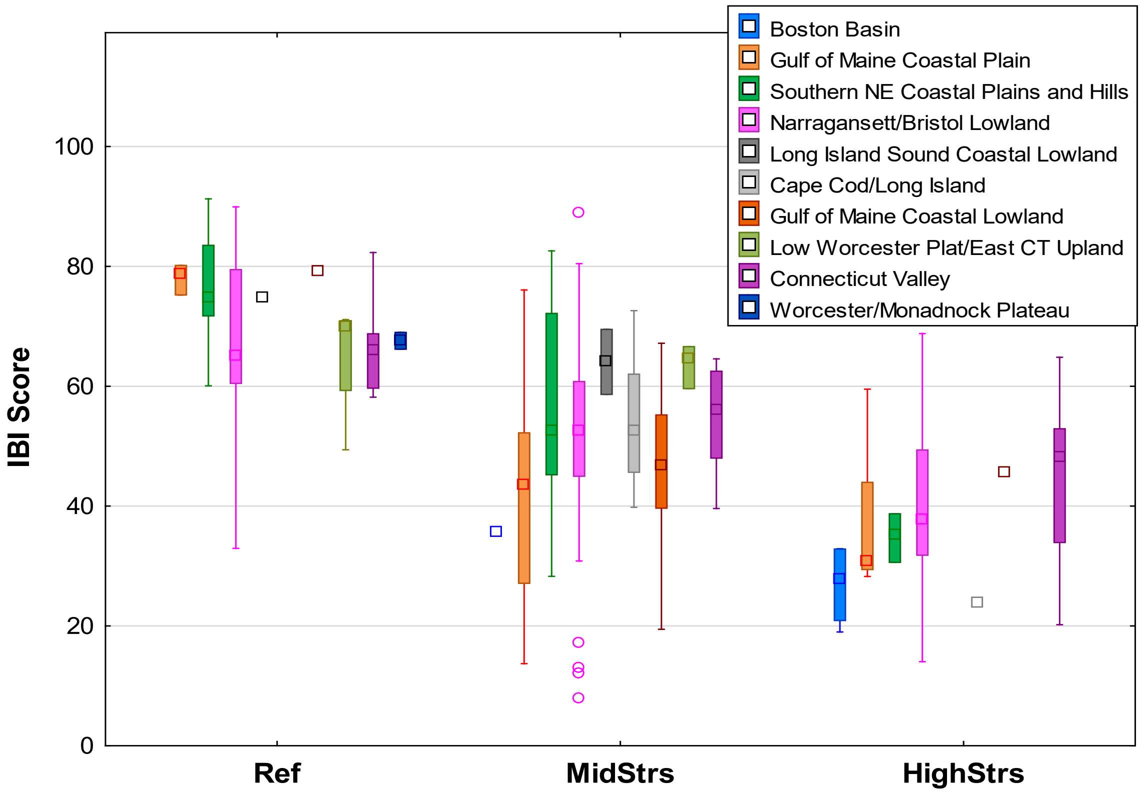


Figure 25. IBI score distributions (medians, interquartile ranges, non-outlier ranges, and outliers) in Level 4 ecoregions and disturbance categories; reference (Ref), intermediate (MidStrs), and stressed (HighStrs).

The IBI scores do not show a strong relationship with percentage of water and wetland in the watershed within reference sites (Figure 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.

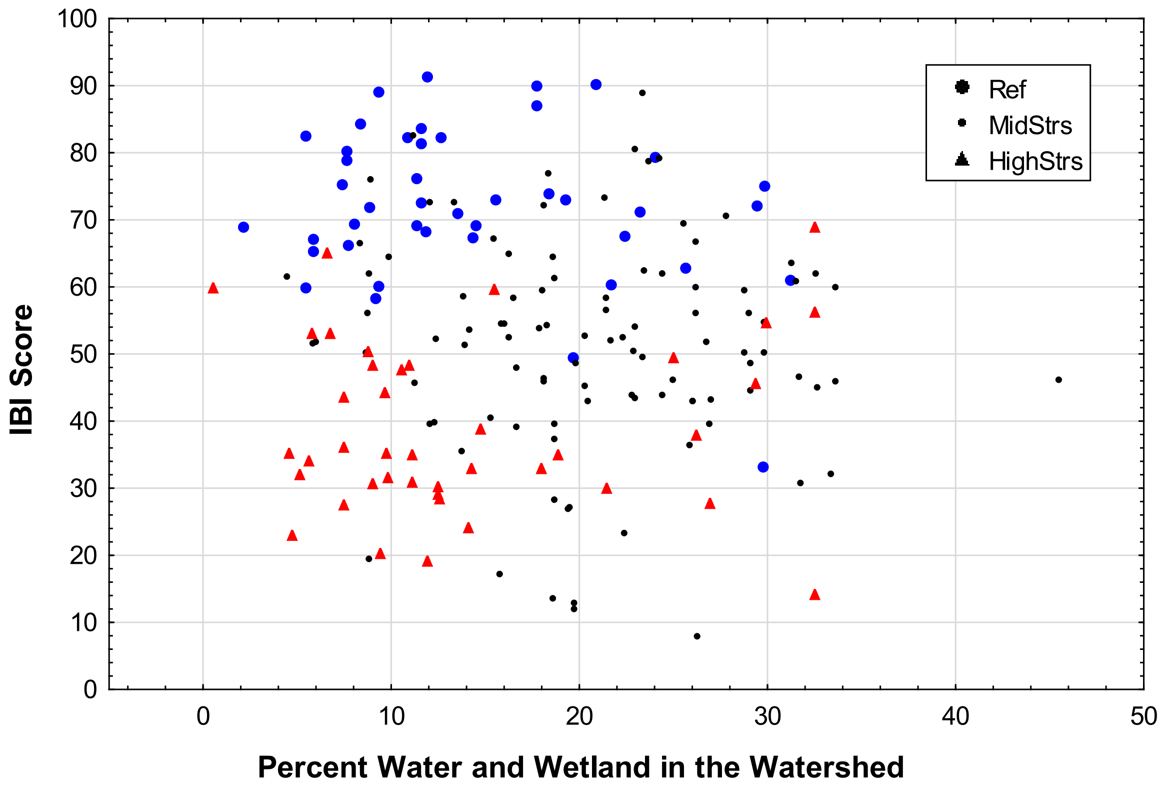


Figure 26. Percent water and wetland in the watershed in relation to IBI scores, marked by disturbance category.

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 B.

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 (0.94), indicating that the adjustment is applicable along the whole index gradient. The revised index is 5.4 points higher than the calibration index, on average. 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 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 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, it is likely that the minimum value is not representative of acceptable reference conditions because a) the reference sites were defined with relative, not absolute, stressor criteria, b) there is variability in biological conditions, and c) there might be undetected stressors due to limited data availability. 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 IBI 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 10^{th} - 25^{th} 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 IBI. In our data set, using the revised traits and scoring formulae in Table 14, these percentiles correspond to IBI scores of 61.5 – 70 index points, respectively (Table 16). At these thresholds, stressed sites would be identified as impacted for 88 – 95% of the cases.

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 near the 10^{th} percentiles, at approximately 61.5 index points (Table 16).

The standard deviation of the reference index distribution was 11.7 index points. A threshold of 61.5 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 64.7 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	184	43			
Minimum	7.9	34.1			
5^{th} Percentile	26.0	57.2	5%	80.5	19.5
10^{th} Percentile	34.0	61.5	10%	87.8	12.2
15^{th} Percentile	40.8	66.7	15%	92.7	7.3
20^{th} Percentile	43.7	69.3	20%	95.1	4.9
Lower Quartile	46.4	70.1	25%	95.1	4.9
Mean	58.7	76.4			
Median	58.6	79.1			
Upper Quartile	72.2	84.4			
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 10^{th} 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 $y = 0.97 x + 6.74$ ($r^2 = 0.94$). If the regression equation is applied to the suggested calibration index threshold, the interpolated revised index threshold would be 64.9 index points.

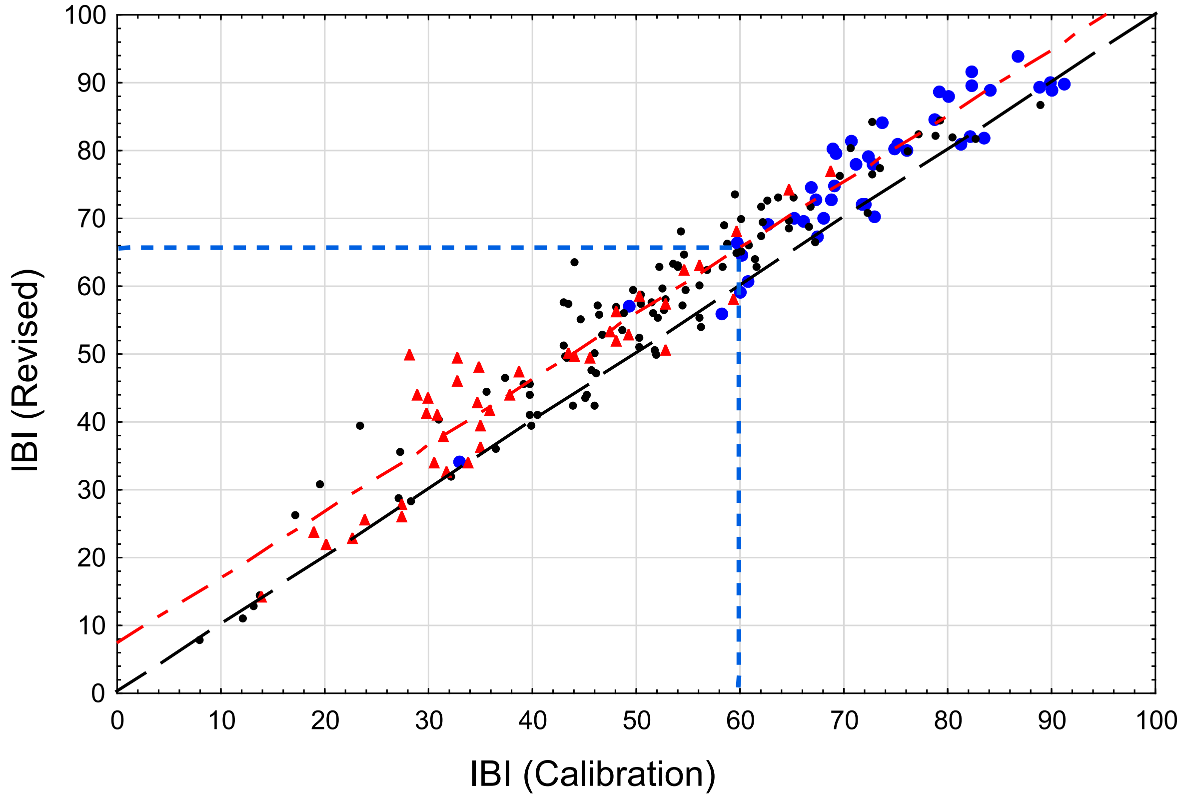


Figure 27. IBI values comparing calibration data and revised data, showing the unity line (black dashed), regression line (red dashed), and the central threshold at 60 in calibration data and 65 in revised data. The regression equation is $y = 0.97x + 6.74$ ($r^2 = 0.94$).

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 61.5 – 70 index points. If the balance of errors and the 10th percentile are given greater weight because they recognize the 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 61.5 index points. For simplicity in application and communication and taking a conservative approach to identifying impacts, a general threshold of 62 index points is recommended.

Secondary Thresholds

As demonstrated in the 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 62 index points. A simple bisection of the unimpacted index range would suggest a threshold of 81 index points, half-way between the general threshold and the maximum of the index scale. Similarly, 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.

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 an equal probability between groups can describe a potential threshold that would evenly divide the Exceptional and Satisfactory index values and also the Moderately Degraded and Severely Degraded index values. Based on proportional odds logistic regression, a threshold between Exceptional and Satisfactory conditions was identified at 80 index points. The threshold between Moderately Degraded and Severely Degraded conditions was identified at 38 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.

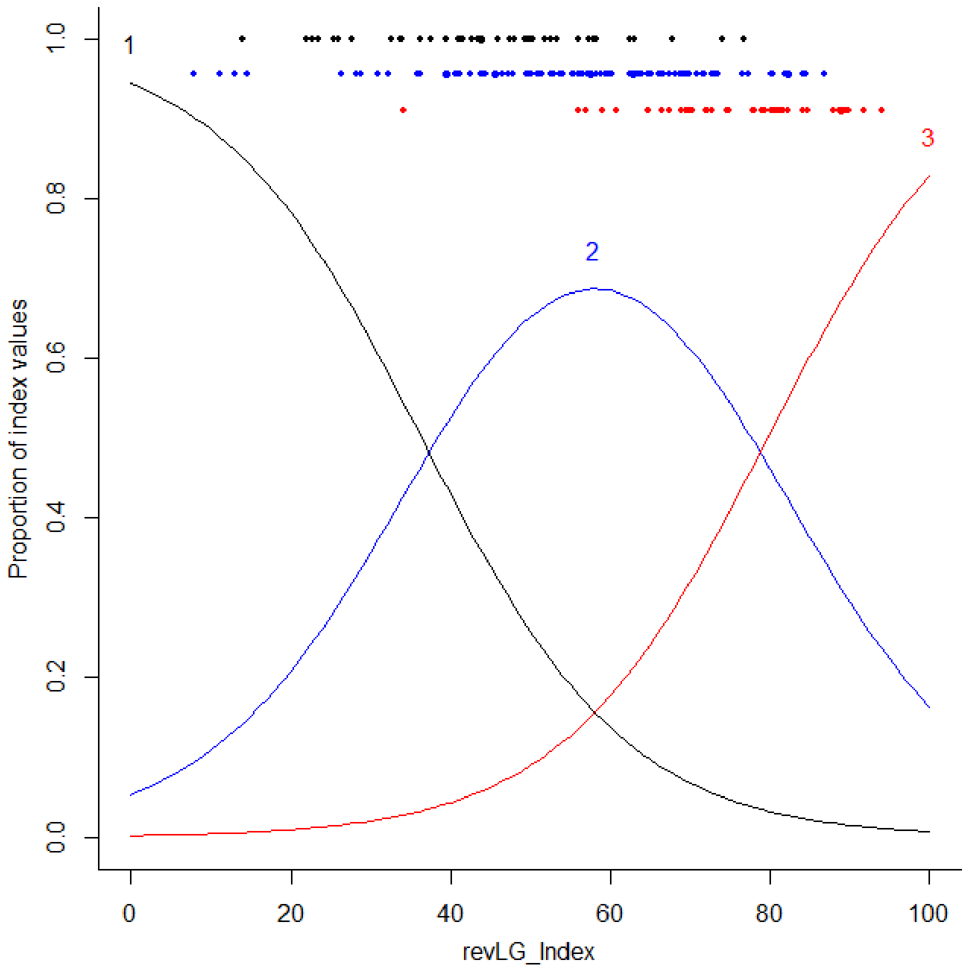


Figure 28. Proportional odds logistic regression graph, showing probability of membership in the highly stressed (1), moderately stressed (2), and reference (3) disturbance categories. Actual data points for the revised index are plotted at the top of the graph.

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 before they are applicable in biological assessment programs.

Table 17. Threshold ranges and recommended index 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">80-81</td><td colspan="2">62-70</td><td colspan="3">31-38</td></tr>
<tr><td>Recommended index threshold</td><td colspan="3">81</td><td colspan="2">62</td><td colspan="3">38</td></tr>
</table>

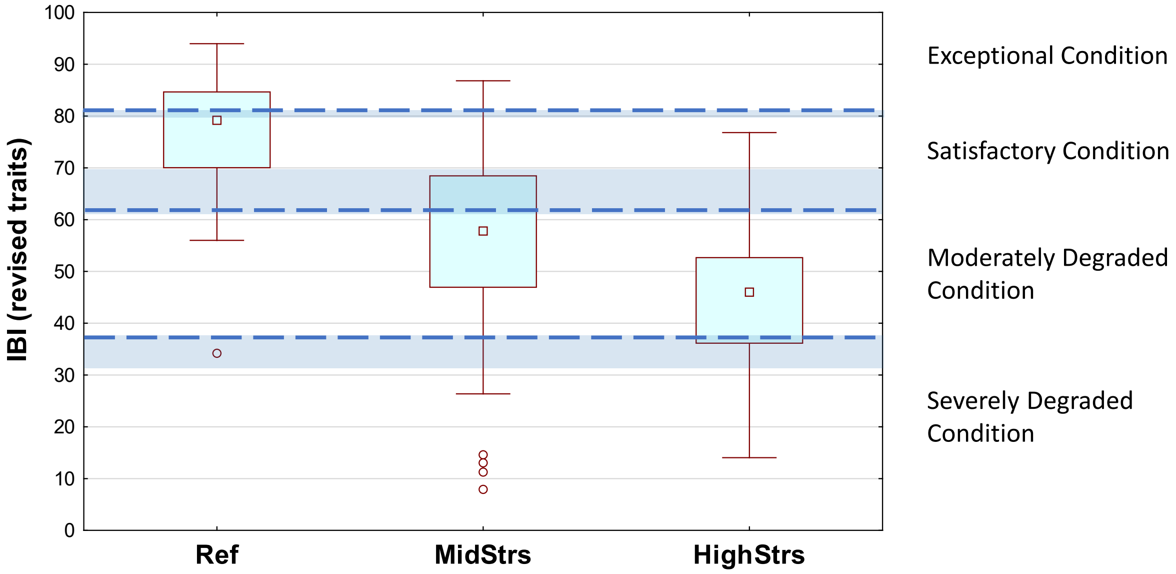


Figure 29. Low gradient index distributions plotted by disturbance category and showing recommended thresholds (dashed lines) and threshold ranges (shaded bars) to describe index values associated with narrative condition categories.

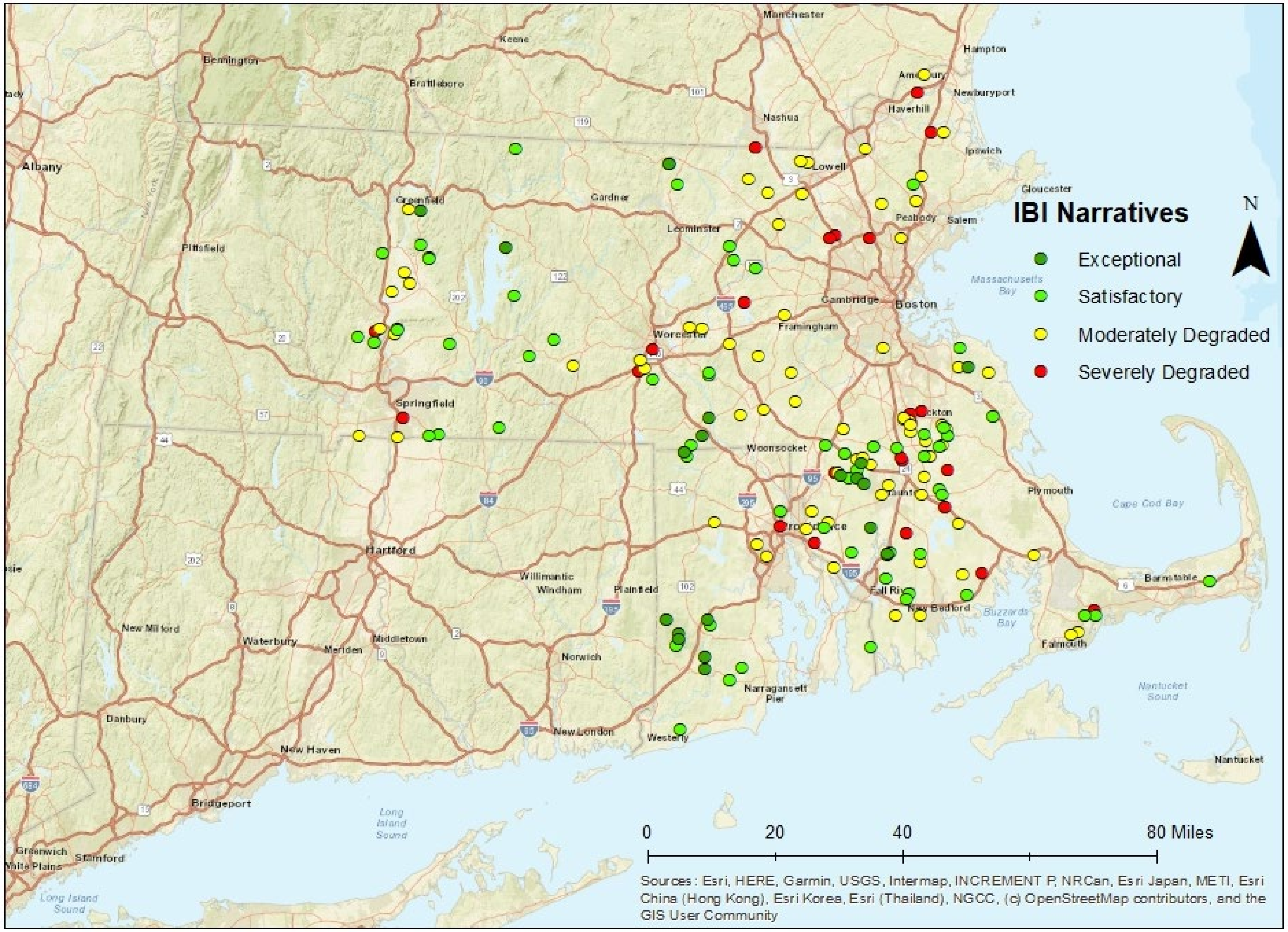


Figure 30. Low gradient, multihabitat sites color-coded by biological condition category based on the recommended IBI thresholds in Table 17.

7 IBI Application

The low gradient, multihabitat IBI improves MassDEP's diagnostic ability to identify degradation in biological integrity and water quality. It can be applied to low gradient streams in the broader Southeast New England region as well (including watersheds in southern Cape Cod, Narragansett Bay and Buzzards Bay). 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 available through a free R-based tool (referred to as a Shiny app). The Shiny app was developed to calculate the low gradient IBI, as well as MassDEP's 100- and 300-count riffle habitat IBIs. The IBI calculator can be accessed via this weblink:

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

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: Massachusetts and Rhode Island
- Stream type: perennial, freshwater, wadeable low gradient, 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 across Massachusetts. 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 explored potential thresholds for four biological condition categories (Exceptional Condition, Satisfactory Condition, Moderately Degraded, and Severely Degraded). These categories can be used in the Consolidated Assessment and Listing Methodology (CALM) to interpret the narrative biological criteria in the SWQS (Massachusetts Division of Watershed

Management Watershed Planning Program 2016). The thresholds proposed in this report are preliminary and subject to further review, refinement, and approval by MassDEP before they are applicable in biological assessment programs. Moving ahead, in addition to further exploring potential IBI thresholds, MassDEP 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., and 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., 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., 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.

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.

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.

Massachusetts Division of Watershed Management Watershed Planning Program. 2016. Massachusetts Consolidated Assessment and Listing Methodology (CALM) Guidance Manual for the 2016 Reporting Cycle. Available online: https://www.mass.gov/files/documents/2016/10/wy/2016calm.pdf

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* 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 New England Interstate Water Pollution Control Commission, 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.

Appendix A

Crosswalk of MassDEP vs SNEP major macroinvertebrate habitat types

Table A1. Crosswalk of MassDEP and SNEP major macroinvertebrate habitat types

MassDEP 2013-2019	**SNEP**
leaf pack	wood jabs*
snags	wood jabs
other (coarse substrates)	hard bottom jabs
other (non-riffle kick)	hard bottom jabs
other (runs with cobble)	hard bottom jabs
other, bottom kicks	hard bottom jabs
other, coarse substrates	hard bottom jabs
other, cobble kicks	hard bottom jabs
other, run	hard bottom jabs
riffle cobbles	hard bottom jabs
riffles	hard bottom jabs
runs	hard bottom jabs
root mats and submerged macrophytes	undercut banks/overhanging vegetation
overhanging vegetation/stream bank	undercut banks/overhanging vegetation
stream banks	undercut banks/overhanging vegetation
submerged macrophytes	submerged vegetation
submerged macrophytes and root mats	submerged vegetation

*leaf packs are typically associated with snags/wood

Appendix B

Taxa tolerance analyses

B1 Background

Taxon 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 help inform macroinvertebrate tolerance value assignments related to sensitivity to stressors in low gradient streams in Massachusetts (MA) and Rhode Island (RI), we ran taxa tolerance analyses on four variables that capture anthropogenic disturbance: the Indices of Watershed and Catchment Integrity (IWI & ICI, respectively) (Thornbrugh et al. 2018, Johnson et al. 2019), percent urban and percent agricultural land use. We also ran analyses to better understand the relationship between taxon occurrence and drainage area, flowline slope, elevation and modeled summer stream temperature. The tolerance analyses were run on a regional dataset that included low gradient data from Massachusetts (MA), Rhode Island (RI), Connecticut (CT), Vermont (VT), and New York (NY). The regional scale allowed for a larger sample size than just the MA/RI dataset alone, which improved the robustness of the analyses and allowed tolerance assignments to be generated for more taxa. Biologists from MassDEP reviewed results from the analyses and assigned taxa to three tolerance categories: intolerant, intermediate, and highly tolerant. In this document, we describe the dataset, methods and results and conclude with recommendations on potential future analyses that could further improve our understanding of taxon-stressor relationships in low gradient streams.

B2 Data compilation

B2.1 Macroinvertebrates

The regional dataset was comprised of macroinvertebrate samples from low gradient, freshwater, wadeable, perennial streams in MA, RI, CT, VT and NY that were collected with each state's low gradient collection method (Table B1). Data from 541 sites that spanned nine Level 3 ecoregions were included in the analysis (Table B2, Figure B1).

Table B1. Summary of the regional macroinvertebrate collection methods being used in MA, RI, CT, VT and NY.

Method	Habitat	Effort	Gear	Reach length	Index period	Target # organisms	Taxonomic resolution
MassDEP RBP 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 major 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
Southern New England Program (SNEP) multihabitat (used in RI and at some MA sites)	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	100-m	July 1 – September 30	300	Lowest practical level
CT DEEP Standard Semi-Quantitative Low Gradient	Multiple habitat approach that focuses primarily on the most productive habitats (vegetation, woody debris, undercut banks/roots) but also includes, at minimized effort, the less productive fine sediment habitat (sand/silt)	20 jabs/sweeps (1 meter in length, followed by 2-3 sweeps through the suspended material. fixed number of two jabs/sweeps from fine sediments; the other eighteen are based on the percentage of most productive habitats present in sampling reach	Long handled, 500-micron mesh, D-frame net	100 meters		200	Lowest practical level

Table B1. continued...

Method	Habitat	Effort	Gear	Reach length	Index period	Target # organisms	Taxonomic resolution
NYSDEC Low gradient	Four habitats: bank, center channel substrate, woody debris/snags and macrophyte bed	Composite of two jab samples for each of the four habitats (8 samples in total). Consistent effort at each habitat for ~30 seconds (total of ~4 minutes for all samples and habitats). Alternating jabbing and sweeping is performed to catch dislodged macroinvertebrates.	Rectangular kick net (23 cm × 46 cm) with 800–900-µm mesh.	20 times wetted width at sample site	June-September	200	Lowest practical level
VT DEC Low gradient (Sweep Bottom Kick Net Sampling)	Debris dams, vegetation, or root wads. Used in wadeable low gradient streams with substrates dominated by silt or sand and velocities, where velocity is less than 0.2 fps and the depth is less than 1 meter	Four-point composite sample. A jab is performed by jabbing the net into debris dams, vegetation, or root wads, pulling back rapidly to dislodge animals, then sweeping forward again into the same area to scoop up dislodged animals. This jabbing and sweeping motion should be repeated several times at the same point and considered one of four jabs. All four jabs (from different points in reach) are then combined into a single composite sample	Mesh size 500 microns, 18" wide x 9" high		September–mid-October	300	Lowest practical (species whenever possible)

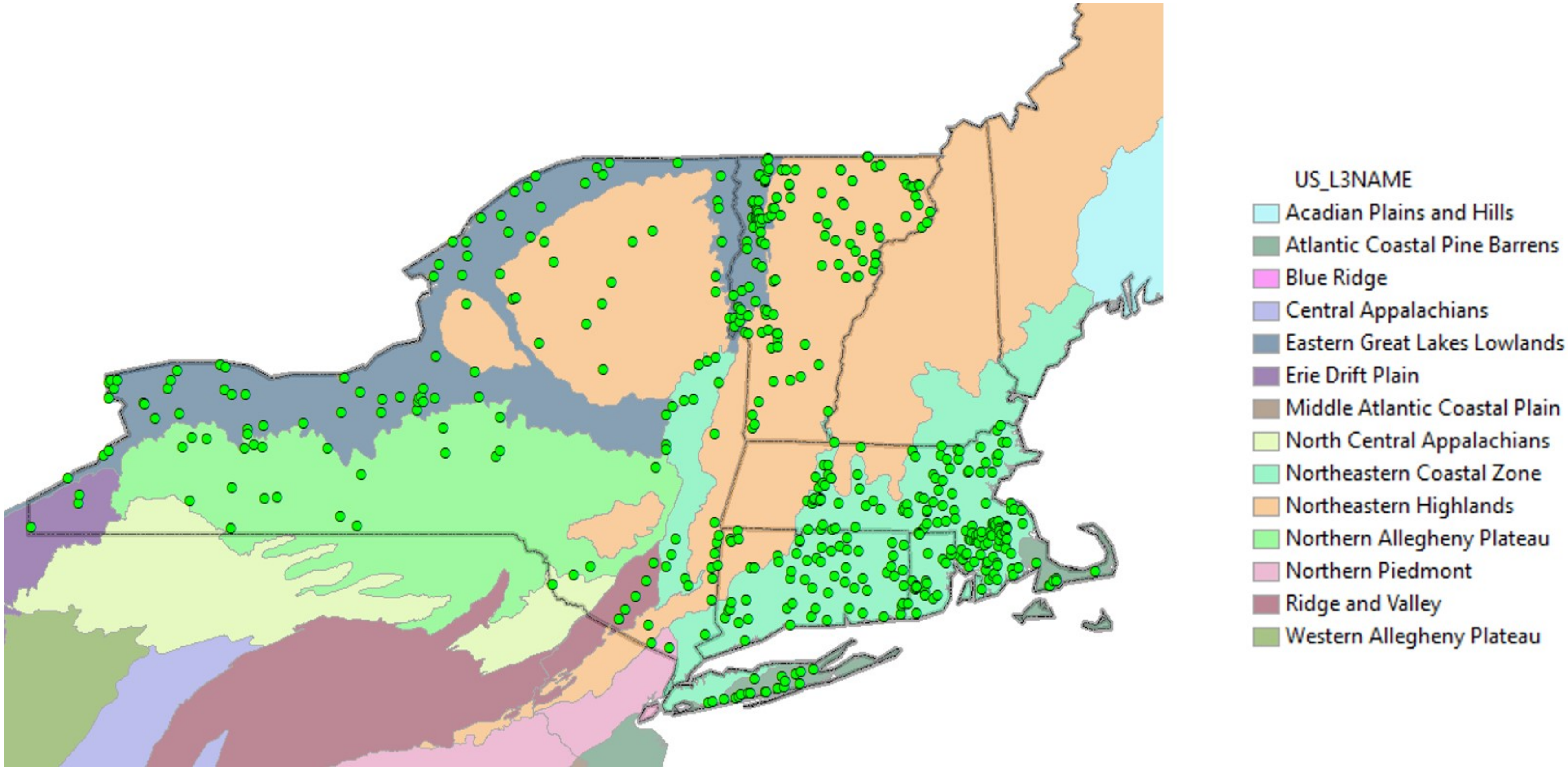


Figure B1. Locations of low gradient sites that were used in the regional tolerance analysis, against a Level 3 ecoregion backdrop.

Table B2. Number of sites in each Level 3 ecoregion.

US_L3NAME	CT	MA	NY	RI	VT	Total
Atlantic Coastal Pine Barrens		7	17			***24***
Eastern Great Lakes Lowlands			66		53	***119***
Erie Drift Plain			3			***3***
North Central Appalachians			2			***2***
Northeastern Coastal Zone	57	152	15	23	1	***248***
Northeastern Highlands	5	2	24		84	***115***
Northern Allegheny Plateau			22			***22***
Northern Piedmont			2			***2***
Ridge and Valley			6			***6***
Total	***62***	***161***	***157***	***23***	***138***	***541***

B2.2 Disturbance variables

We performed the tolerance analysis on four anthropogenic disturbance variables: ICI, IWI, percent urban and percent agricultural land use (Table B3). The data came from the USEPA Stream-Catchment (StreamCat) dataset[1] (Hill et al. 2016), which is associated with the National Hydrography Dataset (NHD) Plus Version 2 (NHDPlusV2) geospatial layer (McKay et al. 2012) via the unique identifiers for the stream segments (COMID) and local catchments (FEATUREID). First we used Geographic Information System software (ArcGIS 10.7.1) to spatially join the biological sampling sites with the NHDPlusV2 dataset. Then we joined the sites with StreamCat data via the NHDPlusV2 identifiers. This was done in a MS Access relational database.

We did several cursory quality control (QC) checks to evaluate whether the biological sampling sites were associated with the correct NHDPlusV2 flowlines. If NHDPlusV2 stream segments had waterbody names (referred to as 'GNIS_Names'), we checked those against the waterbody names of the sites and flagged mismatches for further evaluation. If exact drainage areas were available for the sites, we calculated differences between those and the estimated drainage areas from the StreamCat dataset[2] and flagged sites where differences seemed excessively large (based on our best professional judgment). Next we visually checked the flagged sites to try and determine whether they were associated with the incorrect flowline. One of the most common errors occurred when sites were located on small tributaries that were not captured in the 1:100K NHDPlusV2 dataset and the nearest flowline was a large mainstem. In the end, we excluded 46 sites from the analysis because they were clearly associated with the incorrect flowline.

StreamCat data are available at two spatial scales: local catchment (Cat) (which is defined as the landscape area draining to a single stream segment, excluding upstream contributions) and total watershed (Ws) (which includes the local catchment plus the accumulated area of all upstream catchments) (Figure B2). Three of the disturbance variables (ICI, percent urban and percent agricultural)

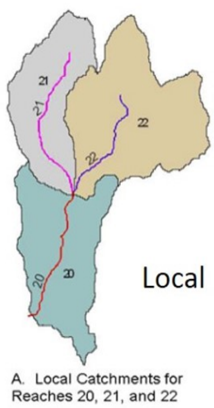
---

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

[2] 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; instead, a site is characterized based on the attributes that are associated with the catchment in which the site is located.

were at the local catchment scale, while the IWI was at the watershed scale. Because 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), there may be occasional inaccuracies in the attribute data. For example, if a site is located upstream of urban land cover, but the urban land cover is located within the local catchment, the urban land cover data will be (wrongly) associated with the site.

The dataset captured a wide range of disturbance. IWI and ICI scores, which are scaled from 0 (worst) to 1 (best), ranged from 0.16 to 0.92, with most sites falling in the middle of that range (0.4 to 0.7) (Table B4, Figure B3). Urban and agricultural land cover at most sites was < 10% (Figure B3), with median values of 4 and 6%, respectively (Table B4, Figure B3). Figure B4 shows the sites overlaid on the NLCD 2016 land cover geospatial layer.

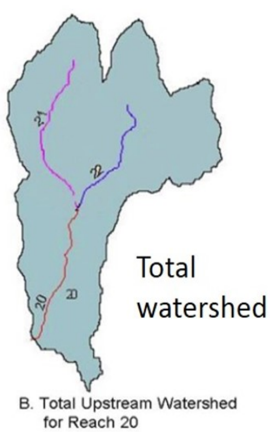


Local catchment
Definition: the landscape area draining to a single stream segment, excluding upstream contributions.

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 B2. USEPA's StreamCat metrics (Hill et al. 2016) cover two spatial scales: local catchment and total watershed.

Table B3. Disturbance variables that were included in the taxa tolerance analyses.

Metric (Abbrev)	Scoring scale	Description	Source
Index of Watershed Integrity version 2.1 (IWI_21)	0 (worst) to 1 (best)	Overall watershed condition at the total watershed scale. Scored based on six components: hydrologic regulation, regulation of water chemistry, sediment regulation, hydrologic connectivity, temperature regulation, and habitat provision	EPA StreamCat (Thornbrugh et al. 2018, Johnson et al. 2019)
Index of Catchment Integrity version 2.1 (ICI_21)	0 (worst) to 1 (best)	Overall watershed condition at the local catchment scale. Scored based on the six components listed above	EPA StreamCat (Thornbrugh et al. 2018, Johnson et al. 2019)
% Urban land use - local catchment scale, based on NLCD 2016 (pcUrb_local)	0 to 100%	% of catchment area classified as developed, low-intensity land use (NLCD 2011 class 22) + medium-intensity land use (NLCD 2011 class 23) + high-intensity land use (NLCD 2011 class 24)	EPA StreamCat (NLCD 2016 - Dewitz 2019)
% Agricultural land use - local catchment scale, based on NLCD 2016 (pcAg_local)	0 to 100%	% of catchment area classified as hay land use (NLCD 2011 class 81) + crop land use (NLCD 2011 class 82)	EPA StreamCat (NLCD 2016 - NLCD 2016 - Dewitz 2019)

Table B4. Summary statistics for the anthropogenic disturbance variables.

Variable	Valid N	Minimum	10th percentile	25th percentile	50th percentile	Mean	75th percentile	90th percentile	Maximum	Std.Dev.
Index of Catchment Integrity version 2.1 (ICI_21)	541	0.16	0.35	0.45	0.56	0.57	0.69	0.79	0.92	0.16
Index of Watershed Integrity version 2.1 (IWI_21)	541	0.16	0.36	0.46	0.57	0.57	0.69	0.79	0.92	0.16
% Urban land use - local catchment scale, based on NLCD 2016 (pcUrb_local)	541	0.00	0.35	1.25	4.44	15.69	19.65	51.35	98.91	22.71
% Agricultural land use - local catchment scale, based on NLCD 2016 (pcAg_local)	541	0.00	0.00	0.61	6.12	14.62	21.54	48.17	84.18	18.84

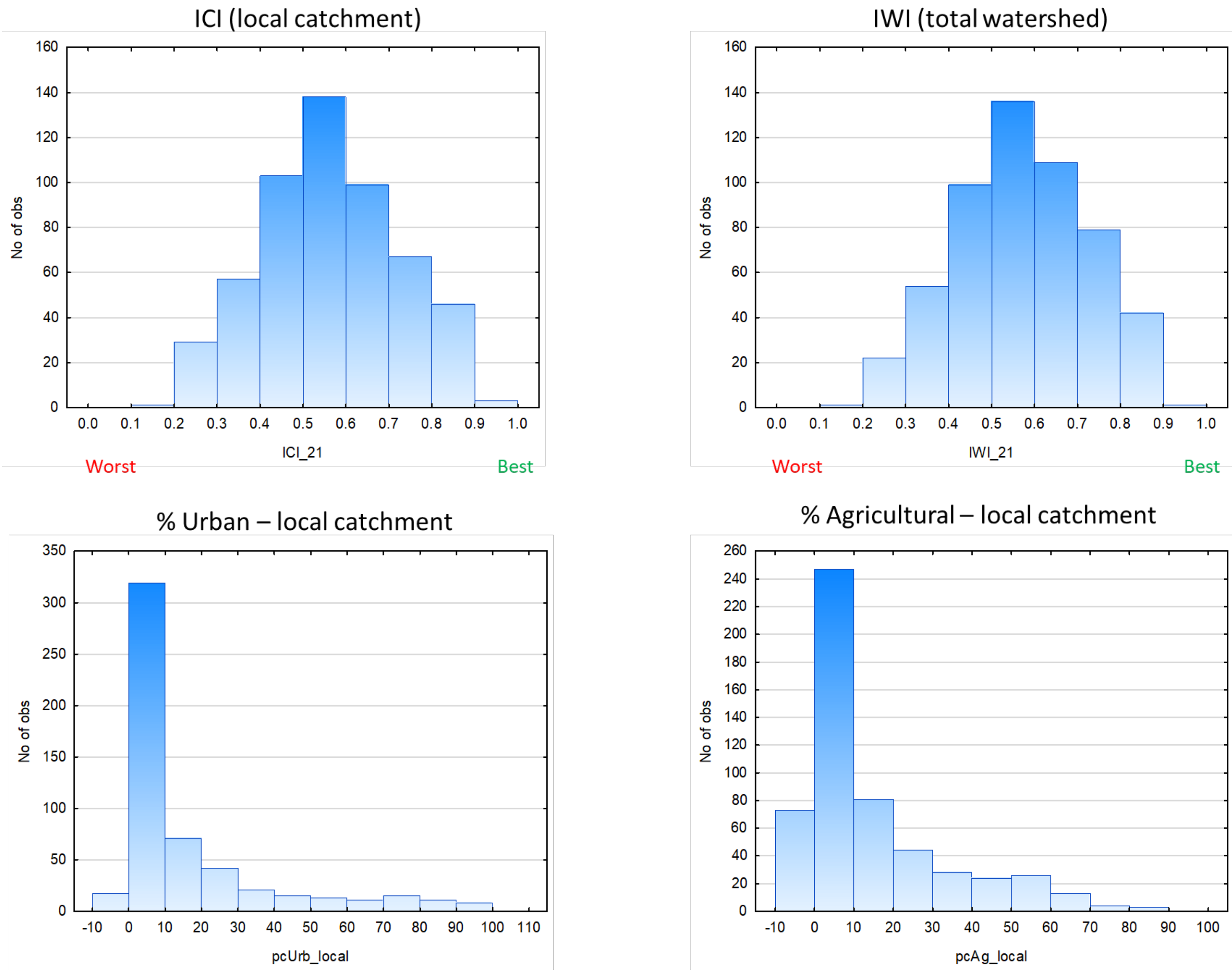


Figure B3. Histograms showing the distribution of sites across the disturbance gradient for each variable (broken into incremental 'bins').

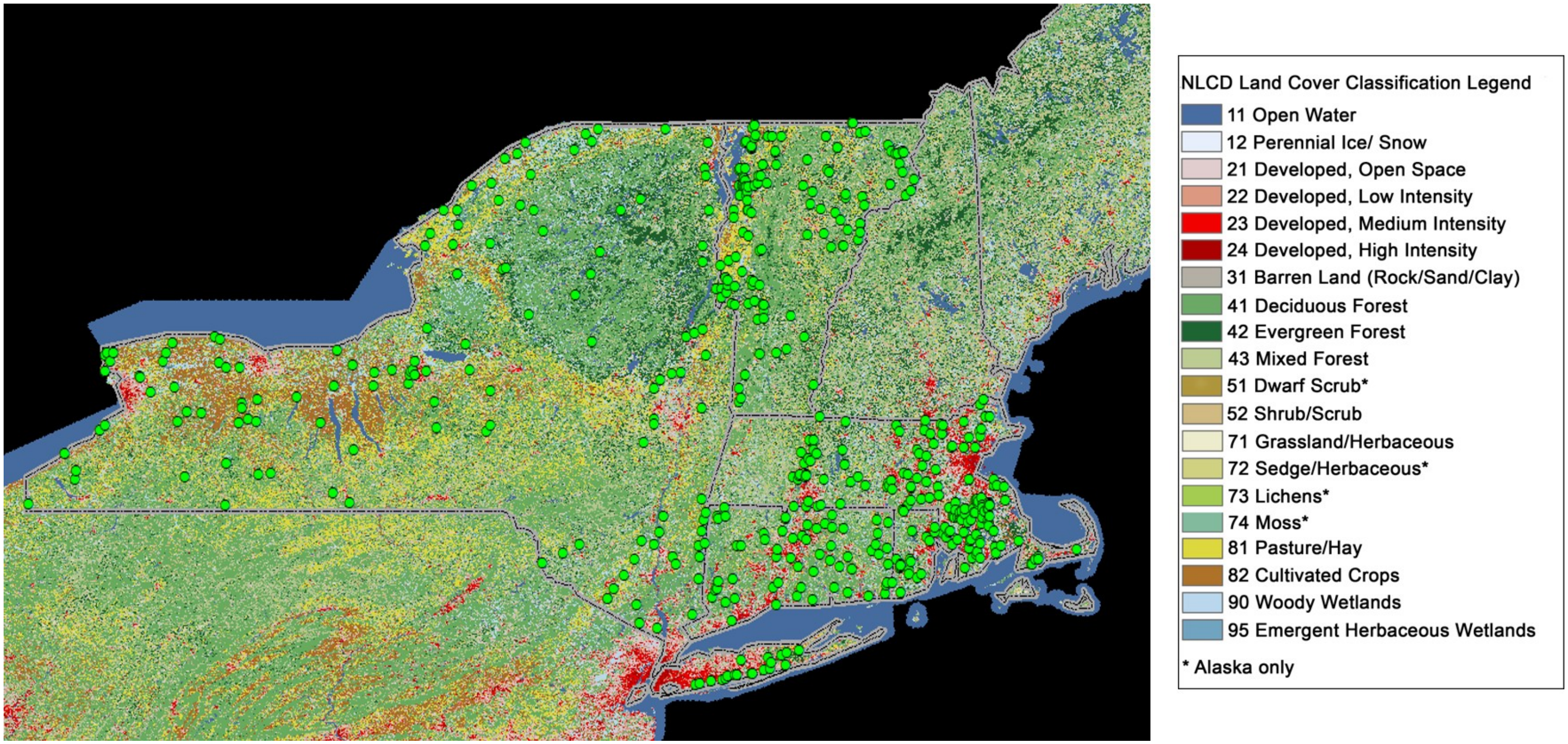


Figure B4. Sites overlaid on the NLCD 2016 land cover geospatial layer.

B2.3 Natural variables

A secondary analysis was performed on four natural variables: drainage area, flowline slope, elevation and modeled summer stream temperature (Table B5). Flowline slope was derived from the NHDPlusV2 attribute data. The source of the other variables was the USEPA Stream-Catchment (StreamCat) dataset (Hill et al. 2016). All four variables are known to influence distributions of macroinvertebrates along a longitudinal gradient (from headwaters to mouth) (Vannote et al. 1980). Most sites had drainage areas less than 100 km^2 (median = 21) and flowline slopes of less than 1% (median = 0.3) (Table B6, Figure B5). Elevation ranged from 7 to 609 meters (median = 111 meters). Most sites had summer stream temperatures in the transitional cool-warm range (18-21°C) (Table B6, Figure B5).

Table B5. Natural variables that were included in the taxa tolerance analyses.

Metric, units (Abbrev)	Description	Source
Drainage area, km2 (DrArea_km2)	Watershed area based on exact delineations where available; where not available, based on EPA StreamCat (estimate from NHDPlusV2 stream segment outlet, i.e., at the most downstream location of the vector line segment	exact delineation or EPA StreamCat estimate
Elevation - local catchment scale, meters (ElevCat)	Mean catchment elevation (m). Obtained from the NHDPlusV2 snapshot of the National Elevation Datasets (NED). Data are distributed through NHDPlusV2 website by HydroRegion.	EPA StreamCat
Flowline slope, % (pcSLOPE)	Slope of flowline (meters/meters) based on smoothed elevations; a value of -9998 means that no slope value is available. See NHDPlusV2 user guide for information about slope computation. Multiplied by 100 to convert to a percentage	NHDPlusV2 (McKay et al. 2012) \NHDPlusAttributes\ElevSlope
Summer stream temperature, °C (MSST_avg)	Modeled mean values for July-August; based on average of 2008, 2009, 2013 and 2014 values in the EPA StreamCat Dataset (which correspond with years of the National Rivers and Streams Assessment (NRSA))	EPA StreamCat (Hill et al. 2013)

Table B6. Summary statistics for the natural variables.

Variable	Valid N	Minimum	10th percentile	25th percentile	50th percentile	Mean	75th percentile	90th percentile	Maximum	Std.Dev.
Drainage area (km2) (DrArea_km2)	541	0.22	4.88	9.75	21.16	45.93	48.72	99.95	1235.12	91.97
Percent Flowline slope (pcSLOPE)	540	0.00	0.03	0.13	0.30	0.64	0.78	1.40	10.30	0.95
Elevation - local catchment scale (m) (ElevCat)	541	7.18	21.34	42.16	111.28	153.75	213.78	376.56	608.97	138.07
Summer stream temperature, degree Celsius (MSST_avg)	536	14.50	16.70	18.00	19.16	19.00	20.18	20.88	22.76	1.58

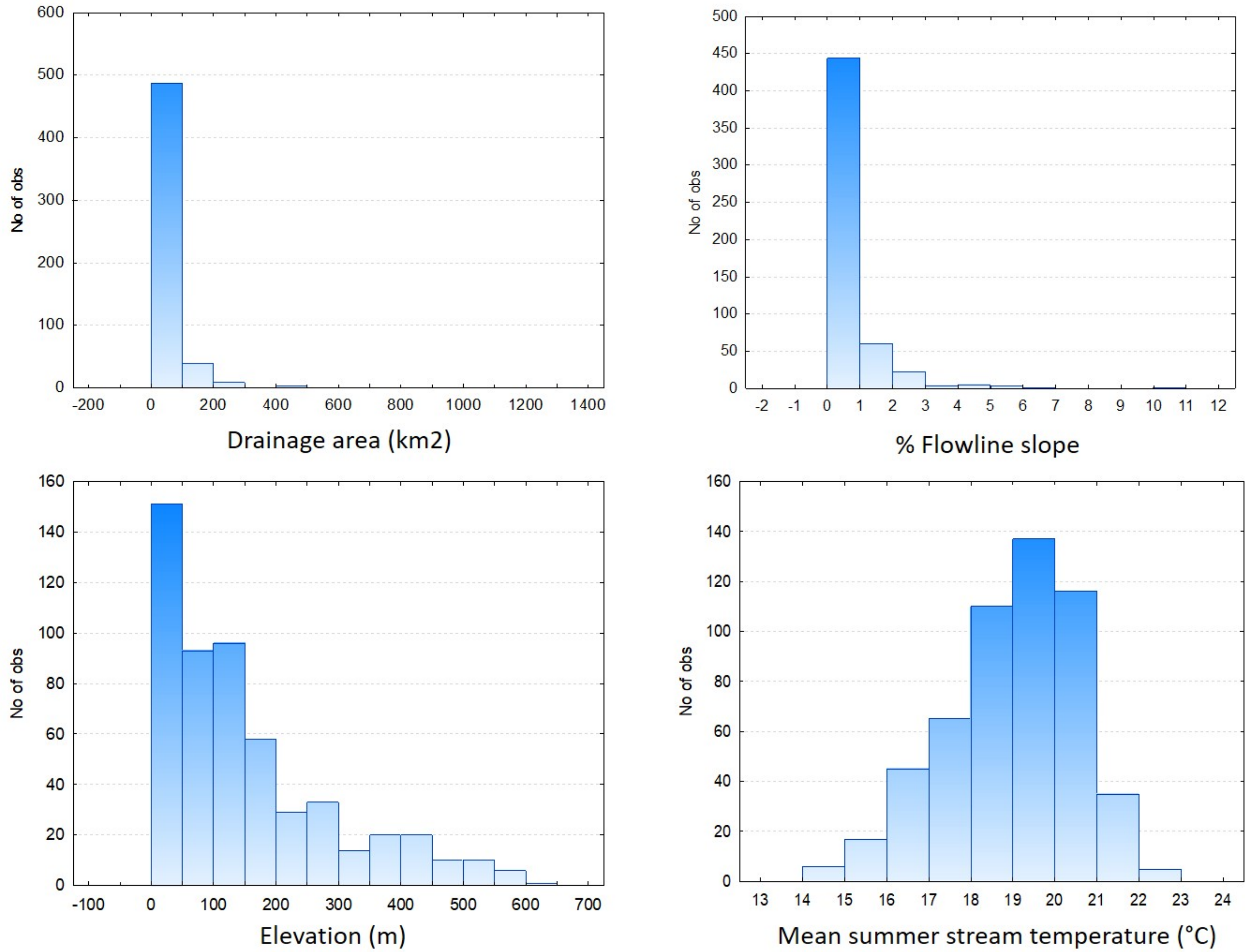


Figure B5. Histograms showing the distribution of sites across gradients for each variable (broken into incremental 'bins').

B3 Methods

B3.1 Data preparation

Data from 541 sites were included in the analysis. To prevent unequal weighting, only one sample per site (the one from the most recent sampling date) was included. To prepare the data, unique taxa names from each entity's dataset were composited into a single 'master' taxa list. We assigned a 'FinalID' after reconciling differences across entites stemming from misspellings and naming schemes (for example, some entities use 'grp' and others use 'group' – e.g., "Eukiefferiella devonica grp" vs. "Eukiefferiella devonica group"; for the FinalID, we changed all to 'group'). We did not delve into possible differences due to use of different taxonomic keys. For each taxon in each sample, we calculated relative abundance, which was used in the tolerance analysis (vs. straight abundance data).

We generated results for five levels of taxonomic resolution: species, genus, tribe, subfamily and family. Analyses were limited to taxa that occurred in at least 10 samples. Table B7 shows an example of how data for seven species of Polypedilum were collapsed to coarser levels of resolution for the genus, tribe, subfamily and family-level analyses. Because all seven species occurred at 10 or more sites, results were generated for each species. For the genus-level run (Polypedilum), the seven species were collapsed to genus-level (otherwise their counts would have been excluded from the coarser-level analyses). The species and genus-level identifications were further collapsed for the tribe, subfamily and family-level analyses (and combined with data for other Chironomini, Chironominae and Chironomidae taxa, as appropriate). Table B8 shows how many taxa within each major taxonomic group were assessed and at what level of taxonomic resolution.

Table B7. Example of how species-level data (in this case, for the midge Polypedilum) were collapsed to coarser levels of resolution for the genus, tribe, subfamily and family-level analyses.

TaxaID	Total # sites	Species	Genus	Tribe	Subfamily	Family
Polypedilum	438	Exclude	Polypedilum	Chironomini	Chironominae	Chironomidae
Polypedilum aviceps	107	Polypedilum aviceps	Polypedilum	Chironomini	Chironominae	Chironomidae
Polypedilum fallax group	104	Polypedilum fallax group	Polypedilum	Chironomini	Chironominae	Chironomidae
Polypedilum flavum	124	Polypedilum flavum	Polypedilum	Chironomini	Chironominae	Chironomidae
Polypedilum halterale group	50	Polypedilum halterale group	Polypedilum	Chironomini	Chironominae	Chironomidae
Polypedilum illinoense group	284	Polypedilum illinoense group	Polypedilum	Chironomini	Chironominae	Chironomidae
Polypedilum scalaenum group	92	Polypedilum scalaenum group	Polypedilum	Chironomini	Chironominae	Chironomidae
Polypedilum tritum	67	Polypedilum tritum	Polypedilum	Chironomini	Chironominae	Chironomidae

Table B8. Number of taxa within each major taxonomic group that were assessed, along with the level of taxonomic resolution.

Taxonomic Group	Family	Subfamily	Tribe	Genus	Species	Total
Amphipods & Isopods	4			5	5	***14***
Bivalvia	2			4	1	***7***
Chironomidae	1	5	10	73	37	***126***
Coleoptera	10	1	1	23	17	***52***
Decapoda	1			1	2	***4***
Diptera without Chironomidae	11	4		19	3	***37***
Ephemeroptera	10			24	15	***49***
Gastropoda	10	1		11	11	***33***
Megaloptera	2			3	1	***6***
Odonata	7			11	5	***23***
Plecoptera	9			9	2	***20***
Trichoptera	16	1		30	13	***60***
Water mites (Trombidiformes)	8			9		***17***
Worms and Leeches	6	2		11	9	***26***
Total	***97***	***14***	***11***	***233***	***121***	***476***

B3.2 Outputs

We used customized R code to generate weighted average optima (WAopt) and tolerance (WAtol) values for each taxon. The WAopt is a commonly used measure for estimating the central tendency of a taxon along an environmental gradient. The WA is calculated by multiplying taxon relative abundance (=the weighting factor) by the variable of interest (e.g., IWI) for each sample, summing the resulting numbers and dividing that by the sum of all the weights. The width of the bell shape is often called 'tolerance' which can also be used to characterize the environmental niche for species along the environmental gradient.

In addition to the WAopt and WAtol values, we generated histograms (Figure B6), relative abundance scatterplots (Figure B7) and cumulative distribution functions (CDFs) (Figure B8) to visualize the relationship between each taxon's occurrence and the environmental variables. The results provide information on where the taxa occur along stressor gradients and whether they increase or decrease in relative abundance with increasing or decreasing stress. Each output also included taxon distribution maps, with data points sized by relative abundance (such that locations with higher relative abundances had larger dots). Separate sets of output files were generated for each taxonomic group, and disturbance and natural variables were analyzed separately.

The WA optima and tolerance values for each taxon/variable were compiled into a MS Excel worksheet. The worksheet also included sample size. Taxa that occurred in fewer than 30 samples were flagged for low abundance[3] and their outputs were interpreted with caution. In addition to the numeric WAopt

[3] More specifically, those that occurred in 10 to 19 samples were flagged as 'very low' and those that occurred in 20-29 samples were flagged as 'low'.

values for each disturbance variable, the worksheet contained columns with categorical, relative rankings for each variable (five levels, ranging from worst to best, based on the criteria in Table B9).

Table B9. Five narrative rankings were assigned to each taxon for each disturbance variable, using the criteria below. Thresholds were based on statistics (the distributions of WAopt values in the dataset) and best professional judgment.

Category	ICI	IWI	PctUrb	PctAg
Worst	<0.50	<0.50	<5	<5
Worse	0.50-0.54	0.50-0.54	5-9.9	5-9.9
Intermediate	0.55-0.65	0.55-0.65	10-19.9	10-19.9
Better	0.66-0.79	0.66-0.79	20-29.9	20-24.9
Best	≥0.80	≥0.80	≥30	≥25

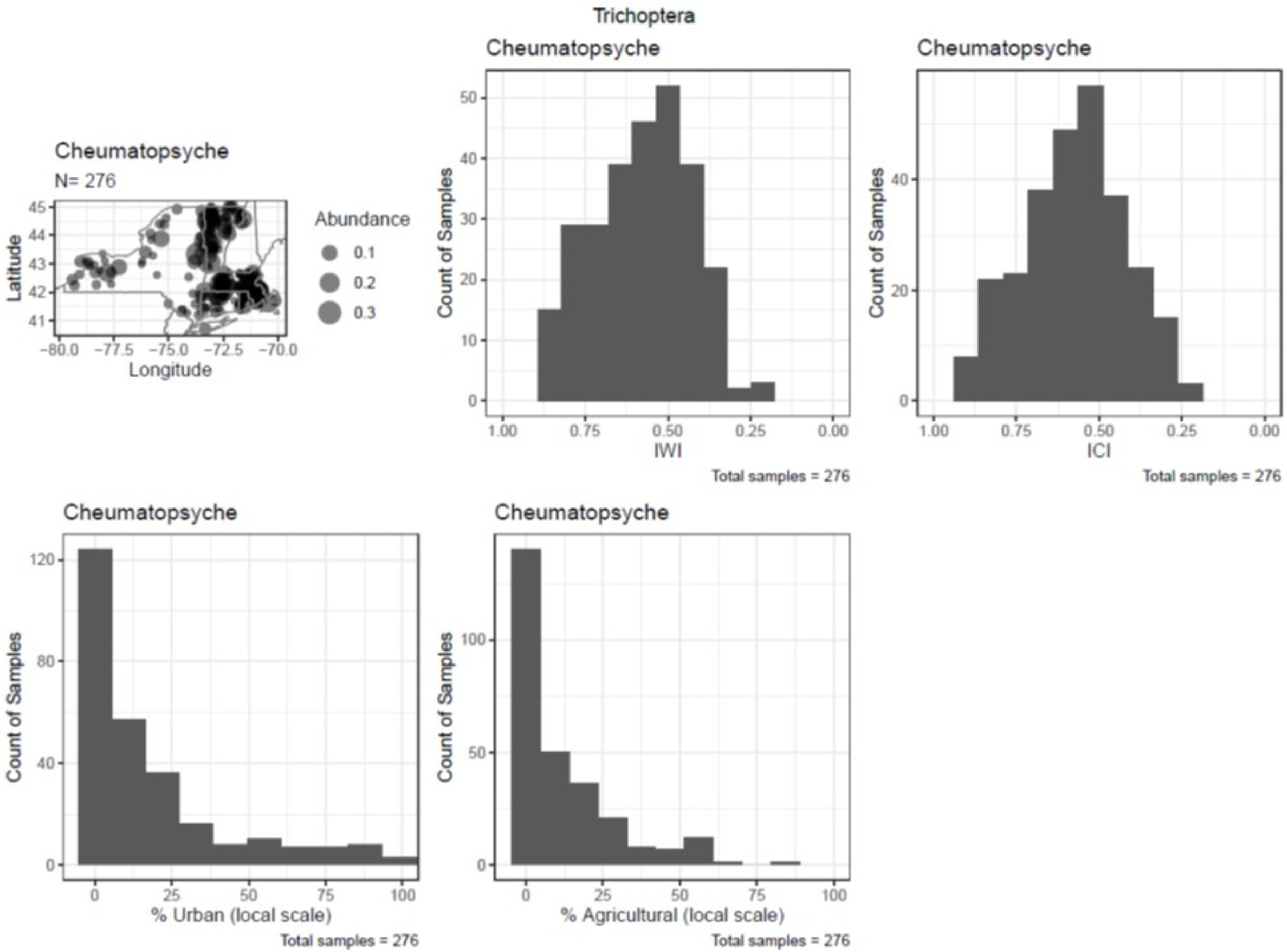


Figure B6. Example of a histogram plot.

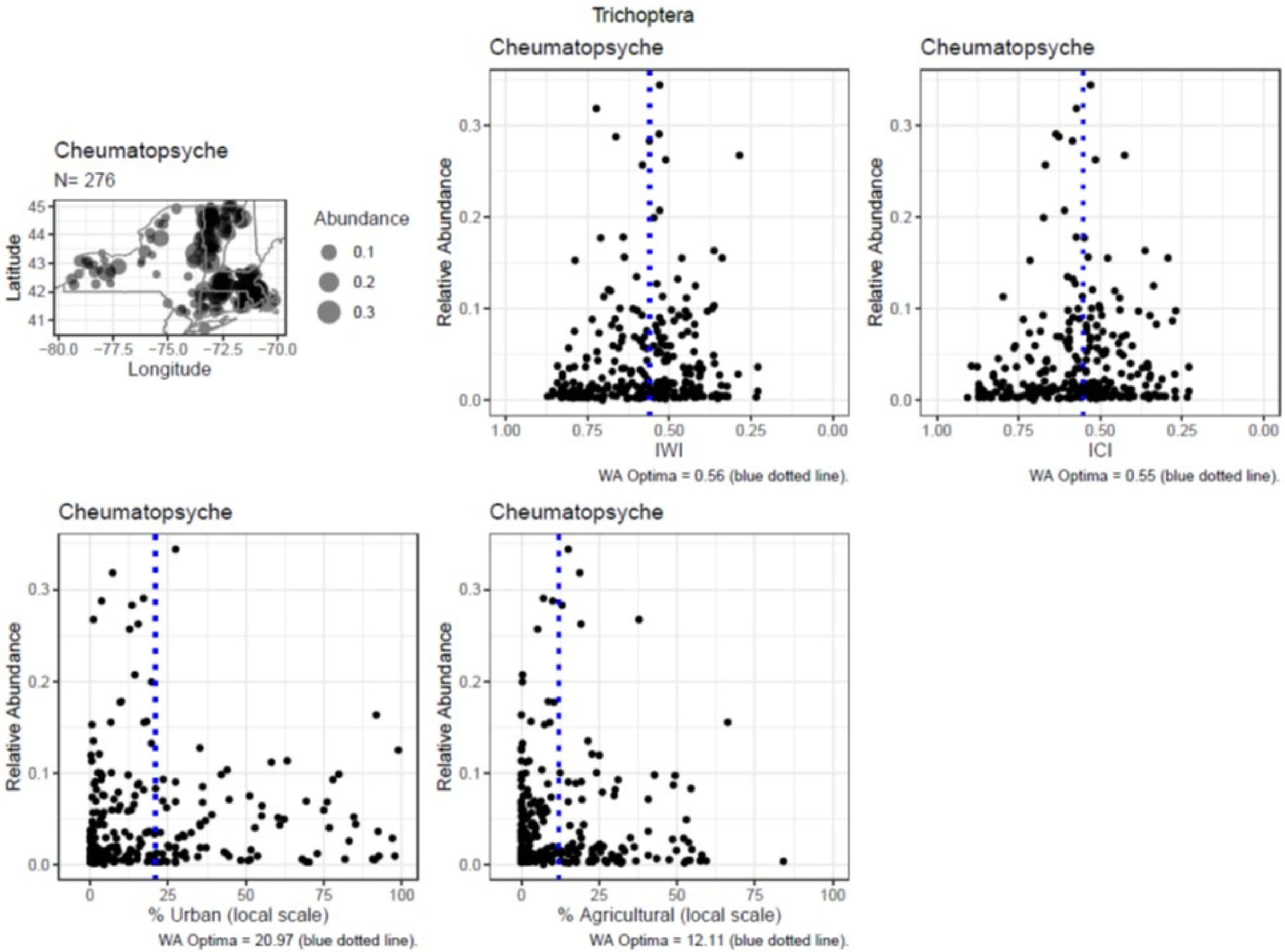


Figure B7. Example of a relative abundance scatterplot. The blue vertical dashed line equals the weighted average optima value.

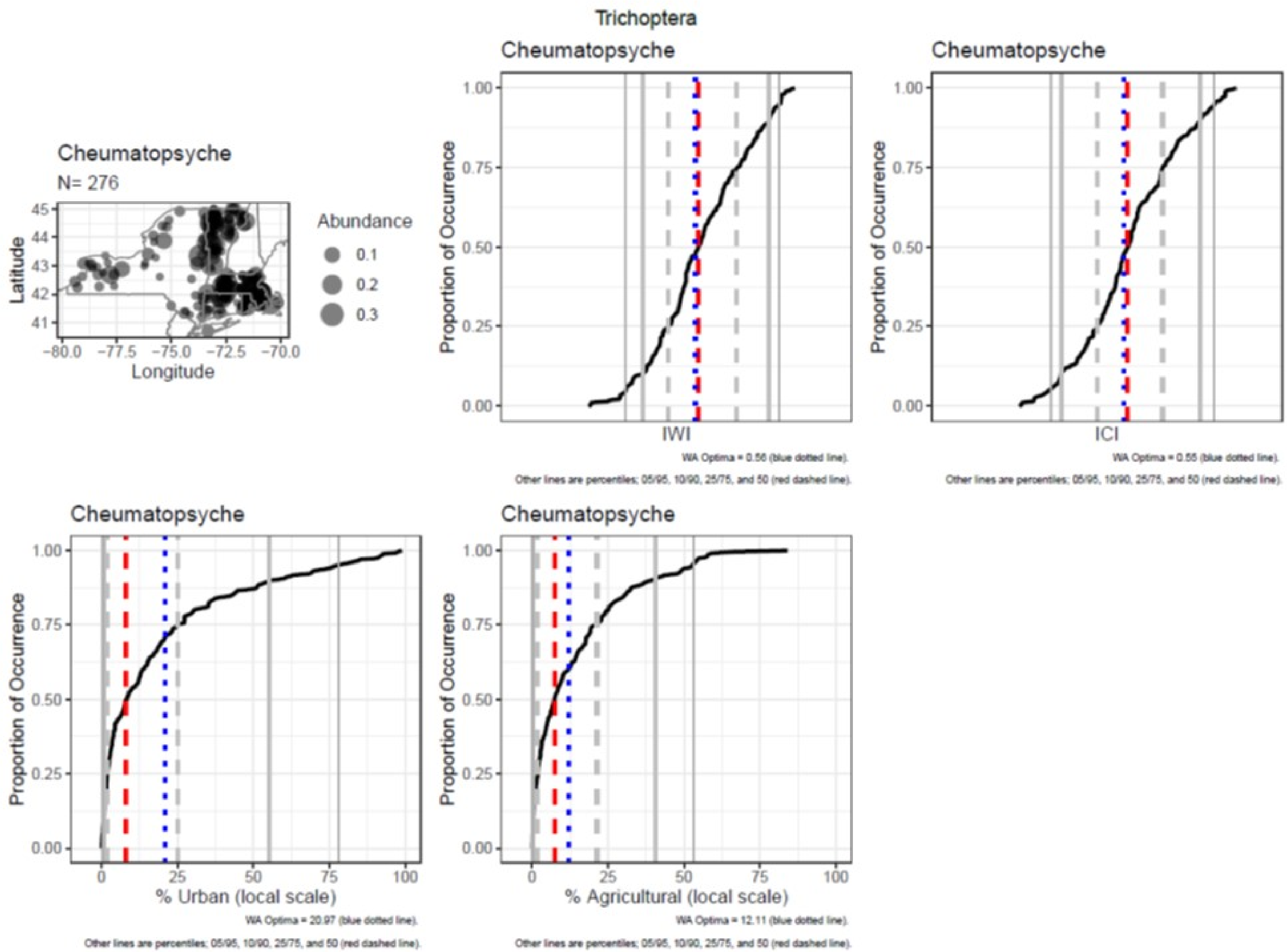


Figure B8. Example of a cumulative distribution function (CDF) plot. The blue vertical dashed line equals the weighted average optima value.

B3.3 Intepretation of results

Biologists from MassDEP who were experienced at assessing macroinvertebrate assemblages reviewed the Excel worksheet and assigned taxa to three tolerance categories: intolerant (numeric value = 2), intermediate (numeric value = 5) and tolerant (numeric value = 8). The worksheet was limited to taxa that occurred in the MA and RI dataset. The review process focused on the disturbance variables, not the natural variables.

The biologists considered multiple lines of evidence when making taxa tolerance assignments, including: 1) WAopt and WAtol values and rankings; 2) distribution across the stressor gradients as shown by the scatterplots, CDFs and histograms; 3) sample size (the more samples the taxon occurred in, the more confident we were in the results); and 4) personal experience and best professional judgment (BPJ). When assigning taxa to the three tolerance categories, the reviewers looked for patterns like those shown in Figure B9. Intolerant taxa occurred mostly (and in higher relative abundance) at sites with the lowest levels of disturbance. Intermediate taxa were generally ubiquitous and most prevalent in the middle of the disturbance gradient. Tolerant taxa tended to occur throughout the stressor gradient and generally increased in relative abundance as stress levels increased. Some taxa showed differing sensitivities to the four disturbance variables. In these situations, the reviewers generally made their assignments based on the 'worst' results (for example, if a taxon was found to be tolerant to stressors associated with urban land cover but not to agricultural land cover, the taxon was generally assigned to the 'tolerant' category).

When interpreting results, it was important for the reviewers to consider both the plots and the WAopt values since WAopt values were sometimes influenced by outliers (see example in Figure B10). The outliers could be either legitimate or incorrect. Potential reasons for erroneous outliers include: the disturbance variable was incorrect (perhaps because the StreamCat data were not based on exact watershed delineations), or the taxon was misidentified. Reasons for the outliers were not investigated. When interpreting results, the reviewers took note of outliers but focused more on the dispersal of data points across the rest of the gradient.

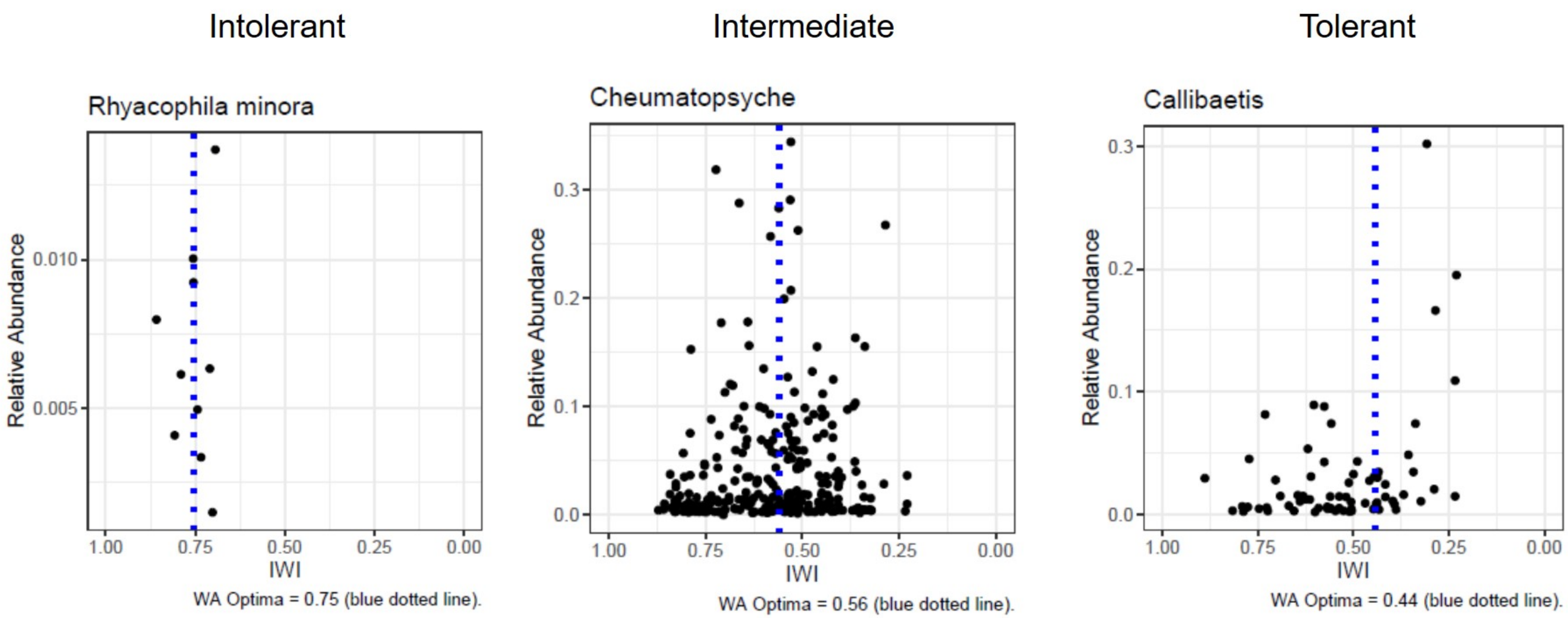


Figure B9. Examples of taxon-response patterns for taxa that were categorized as intolerant, intermediate tolerant and tolerant. The IWI scoring scale ranges from 0 (worst) to 1 (best).

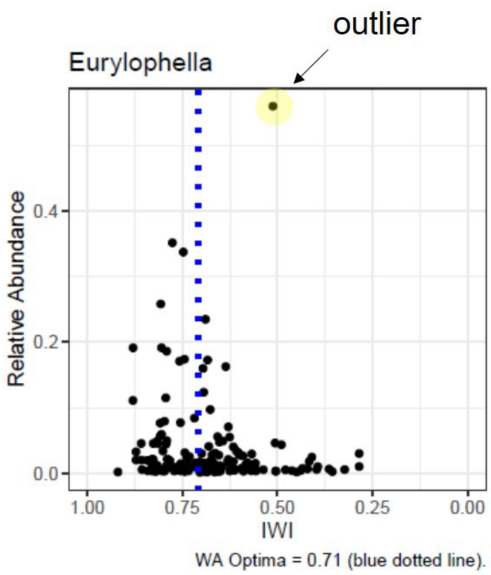


Figure B10. Example of a situation where a taxon's WAopt value was influenced by an outlier.

B4 Results

Table B10 shows the number of taxa in each tolerance category, by taxonomic group. Most taxa were placed in the intermediate group (257 of the 331 taxa that were assessed). The worms/leeches and Chironomidae had the most taxa in the tolerant group, while Plecoptera had the most intolerant taxa, followed by Ephemeroptera and Chironomidae (Table B10). The full set of results (including the plots and worksheet that the reviewers used) are available upon request (contact Jen.Stamp@tetratech.com).

Table B10. Distribution of taxa across tolerance categories, broken into taxonomic groups.

Taxonomic Group	# Intolerant	# Intermediate	# Tolerant	*Total #*
Amphipods & Isopods	0	2	5	***7***
Bivalvia	0	3	2	***5***
Chironomidae	6	80	15	***101***
Coleoptera	1	26	0	***27***
Diptera without Chironomidae	1	21	2	***24***
Ephemeroptera	6	25	1	***32***
Gastropoda	0	14	5	***19***
Megaloptera	0	5	0	***5***
Odonata	2	16	1	***19***
Plecoptera	7	8	0	***15***
Trichoptera	3	42	0	***45***
Water mites (Trombidiformes)	0	6	3	***9***
Worms and Leeches	0	9	14	***23***
Total	***26***	***257***	***48***	***331***

B5 Conclusions

We used low gradient stream macroinvertebrate data provided by regional partners and the StreamCat dataset to examine relationships between taxa occurrence and anthropogenic disturbance. Results helped inform macroinvertebrate tolerance value assignments related to sensitivity to stressors in low gradient streams. The tolerance values were then used to calculate tolerance-based metrics, one of which is included in MassDEP's low gradient Index of Biological Integrity (IBI) (% Tolerant taxa).

While the taxa tolerance analysis described here was an important step forward, more work remains to be done. If resources permit, recommendations for possible future work include:

- Running a similar analysis on data collected from riffle habitats in higher gradient, rocky bottom streams, and then comparing results with the low gradient outputs. This will help biologists better understand differences in the structure and function of macroinvertebrate assemblages in low vs. higher gradient streams, which in turn will improve the ability of biomonitoring programs to identify degradation in biological integrity and water quality.
- Rerunning the low gradient analyses with:
  - New data that MA, RI, CT, NY and VT have collected since the time of the analysis
  - (Possibly) data from low gradient streams in Maine and New Hampshire (caveat: first we'd need to evaluate the suitability of rock basket data for this type of analysis)
  - Environmental data based on exact watershed delineations. Doing exact watershed delineations with the USGS StreatStats stream layer may allow for inclusion of the 46 sites that had to be excluded because they did not match with the NHDPlusV2 flowlines
  - Running an additional set of plots based on Generalized Additive Models (GAM) (see examples in Yuan 2006)
- Working with a group of regional biologists on reviewing results, and through that process, developing better guidance on how to interpret results.
- Developing a regional Biological Condition Gradient (BCG) model for low gradient streams, to go along with the existing New England high gradient streams BCG model (Stamp and Gerritsen 2009)

B6 Literature Cited

Dewitz, J., 2019, National Land Cover Database (NLCD) 2016 Products: U.S. Geological Survey data release, https://doi.org/10.5066/P96HHBIE.

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

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

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

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

Stamp, J. and J. Gerritsen. 2009. Methods and Assessment Comparability Among State and Federal Biological Monitoring Protocols. Prepared for New England Interstate Water Pollution Control Commission. Prepared by Tetra Tech, Inc.

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

Vannote, R.L., Minshall, G., Cummins, K., Sedell, J., & Gushing, C.E. 1980. The River Continuum Concept. Canadian Journal of Fisheries and Aquatic Sciences, 37, 130-137.

Yuan, Lester. 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.

Acknowledgments

State partners provided low gradient data that were used in the regional taxa tolerance analyses. Contributors included Allyson Yarra, James Meek and Robert Nuzzo from Massachusetts Department of Environmental Protection (MassDEP), Katie DeGoosh and Jane Sawyers from Rhode Island Department of Environmental Management, 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 your assistance.

Appendix C

C1 Candidate Metrics

C2 Summary of input metrics in existing low gradient IBIs

C3 Non-target taxa

C4 Exclusion criteria for redundant taxa

C1 Candidate Metrics

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

Metric Name	Category	Description
nt_total	RICH	number of taxa - total
nt_Amph	RICH	number of taxa - Order Amphipoda
pi_Amph	COMP	percent individuals - Order Amphipoda
pt_Amph	RICH	percent taxa - Order Amphipoda
nt_Isop	RICH	number of taxa - Order Isopoda
pi_Isop	COMP	percent individuals - Order Isopoda
pt_Isop	RICH	percent taxa - Order Isopoda
pi_AmphIsop	COMP	percent individuals - Orders Amphipoda & Isopoda
pi_Baet	COMP	percent individuals - Family Baetidae
nt_Bival	RICH	number of taxa - Class Bivalvia
pi_Bival	COMP	percent individuals - Class Bivalvia
pt_Bival	RICH	percent taxa - Class Bivalvia
pi_Caen	COMP	percent individuals - Family Caenidae
nt_Chiro	RICH	number of taxa - Family Chironomidae
pi_Chiro	COMP	percent individuals - Family Chironomidae
pt_Chiro	RICH	percent taxa - Family Chironomidae
nt_Coleo	RICH	number of taxa - Order Coleoptera
pi_Coleo	COMP	percent individuals - Order Coleoptera
pt_Coleo	RICH	percent taxa - Order Coleoptera
nt_COET	RICH	number of taxa - Order Coleoptera, Odonata, Ephemeroptera & Trichoptera
pi_COET	COMP	percent individuals - Order Coleoptera, Odonata, Ephemeroptera & Trichoptera
pt_COET	RICH	percent taxa - Order Coleoptera, Odonata, Ephemeroptera & Trichoptera
nt_CruMol	RICH	number of taxa - Crustacea & Mollusca
pi_CruMol	COMP	percent individuals - Crustacea & Mollusca
nt_Dipt	RICH	number of taxa - Order Diptera
pi_Dipt	COMP	percent individuals - Order Diptera
pt_Dipt	RICH	percent taxa - Order Diptera
nt_Ephem	RICH	number of taxa - Order Ephemeroptera
pi_Ephem	COMP	percent individuals - Order Ephemeroptera
pt_Ephem	RICH	percent taxa - Order Ephemeroptera
pi_EphemNoCae	COMP	percent individuals - Order Ephemeroptera, excluding Family Caenidae
pi_EphemNoCaeBae	COMP	percent individuals - Order Ephemeroptera, excluding Families Caenidae & Baetidae

nt_EPT	RICH	number of taxa - Orders Ephemeroptera, Plecoptera & Trichoptera (EPT)
pi_EPT	COMP	percent individuals - Orders Ephemeroptera, Plecoptera & Trichoptera (EPT)
pt_EPT	RICH	percent taxa - Orders Ephemeroptera, Plecoptera & Trichoptera (EPT)
nt_Gast	RICH	number of taxa - Class Gastropoda
pi_Gast	COMP	percent individuals - Class Gastropoda
pt_Gast	RICH	percent taxa - Class Gastropoda
pi_Hydro	COMP	percent individuals - Family Hydropsychidae
nt_Insect	RICH	number of taxa - Class Insecta
pi_Insect	COMP	percent individuals - Class Insecta
pt_Insect	RICH	percent taxa - Class Insecta
nt_Mega	RICH	number of taxa - Order Megaloptera
pi_Mega	COMP	percent individuals - Order Megaloptera
pt_Mega	RICH	percent taxa - Order Megaloptera
nt_NonIns	RICH	number of taxa - Class not Insecta
pi_NonIns	COMP	percent individuals - Class not Insecta
pt_NonIns	RICH	percent taxa - Class not Insecta
nt_Odon	RICH	number of taxa - Order Odonata
pi_Odon	COMP	percent individuals - Order Odonata
pt_Odon	RICH	percent taxa - Order Odonata
nt_OET	RICH	number of taxa - Orders Odonata, Ephemeroptera & Trichoptera (OET)
pi_OET	COMP	percent individuals - Orders Odonata, Ephemeroptera & Trichoptera (OET)
pt_OET	RICH	percent taxa - Orders Odonata, Ephemeroptera & Trichoptera (OET)
nt_Oligo	RICH	number of taxa - Class Oligochaeta
pi_Oligo	COMP	percent individuals - Class Oligochaeta
pt_Oligo	RICH	percent taxa - Class Oligochaeta
nt_Pleco	RICH	number of taxa - Order Plecoptera
pi_Pleco	COMP	percent individuals - Order Plecoptera
pt_Pleco	RICH	percent taxa - Order Plecoptera
nt_POET	RICH	number of taxa - Orders Plecoptera, Odonata, Ephemeroptera & Trichoptera (POET)
pi_POET	COMP	percent individuals -Orders Plecoptera, Odonata, Ephemeroptera & Trichoptera (POET)
pt_POET	RICH	percent taxa - Orders Plecoptera, Odonata, Ephemeroptera & Trichoptera (POET)
nt_Trich	RICH	number of taxa - Order Trichoptera
pi_Trich	COMP	percent individuals - Order Trichoptera
pt_Trich	RICH	percent taxa - Order Trichoptera

pi_TricNoHydro	COMP	percent individuals - Order Trichoptera, excluding Family Hydropsychidae
pi_SimBtri	COMP	percent individuals - Families Simuliidae & Baetis tricaudatus
pi_dom01	RICH	percent individuals - most dominant taxon [max(N_TAXA)]
pi_dom02	RICH	percent individuals - two most dominant taxa
pi_dom03	RICH	percent individuals - three most dominant taxa
pi_dom04	RICH	percent individuals - four most dominant taxa
pi_dom05	RICH	percent individuals - five most dominant taxa
x_Shan_2	RICH	Shannon Wiener Diversity Index (log base 2) - x_Shan_Num/log(2)
x_D	RICH	Simpson's Index
x_Evenness	RICH	Evenness=x_Shan_e/log(nt_total)
x_Becks	TOLER	Becks Biotic Index = 2*[C1Taxa]+[C2Taxa] (see footnote)
x_HBI	TOLER	Hilsenhoff Biotic Index (references the TolVal field)
nt_tv_intol	TOLER	number of taxa - tolerance value - intolerant ≤ 3
pi_tv_intol	TOLER	percent individuals - tolerance value - intolerant ≤ 3
pt_tv_intol	TOLER	percent taxa - tolerance value - intolerant ≤ 3
nt_tv_toler	TOLER	number of taxa - tolerance value -tolerant ≥ 7
pi_tv_toler	TOLER	percent individuals - tolerance value -tolerant ≥ 7
pt_tv_toler	TOLER	percent taxa - tolerance value -tolerant ≥ 7
nt_ffg_col	FFG	number of taxa - Functional Feeding Group (FFG) - collector-gatherer (CG)
pi_ffg_col	FFG	percent individuals - Functional Feeding Group (FFG) - collector-gatherer (CG)
pt_ffg_col	FFG	percent taxa - Functional Feeding Group (FFG) - collector-gatherer (CG)
nt_ffg_filt	FFG	number of taxa - Functional Feeding Group (FFG) - collector-filterer (CF)
pi_ffg_filt	FFG	percent individuals - Functional Feeding Group (FFG) - collector-filterer (CF)
pt_ffg_filt	FFG	percent taxa - Functional Feeding Group (FFG) - collector-filterer (CF)
nt_ffg_pred	FFG	number of taxa - Functional Feeding Group (FFG) - predator (PR)
pi_ffg_pred	FFG	percent individuals - Functional Feeding Group (FFG) - predator (PR)
pt_ffg_pred	FFG	percent taxa - Functional Feeding Group (FFG) - predator (PR)
nt_ffg_scrap	FFG	number of taxa - Functional Feeding Group (FFG) - scraper (SC)
pi_ffg_scrap	FFG	percent individuals - Functional Feeding Group (FFG) - scraper (SC)
pt_ffg_scrap	FFG	percent taxa - Functional Feeding Group (FFG) - scraper (SC)
nt_ffg_shred	FFG	number of taxa - Functional Feeding Group (FFG) - shredder (SH)
pi_ffg_shred	FFG	percent individuals - Functional Feeding Group (FFG) - shredder (SH)
pt_ffg_shred	FFG	percent taxa - Functional Feeding Group (FFG) - shredder (SH)
nt_habit_burrow	HABIT	number of taxa - Habit - burrowers (BU)
pi_habit_burrow	HABIT	percent individuals - Habit - burrowers (BU)

pt_habit_burrow	HABIT	percent taxa - Habit - burrowers (BU)
nt_habit_climb	HABIT	number of taxa - Habit - climbers (CB)
pi_habit_climb	HABIT	percent individuals - Habit - climbers (CB)
pt_habit_climb	HABIT	percent taxa - Habit - climbers (CB)
nt_habit_cling	HABIT	number of taxa - Habit - clingers (CN)
pi_habit_cling	HABIT	percent individuals - Habit - clingers (CN)
pt_habit_cling	HABIT	percent taxa - Habit - clingers (CN)
nt_habit_sprawl	HABIT	number of taxa - Habit - sprawlers (SP)
pi_habit_sprawl	HABIT	percent individuals - Habit - sprawlers (SP)
pt_habit_sprawl	HABIT	percent taxa - Habit - sprawlers (SP)
nt_habit_swim	HABIT	number of taxa - Habit - swimmers (SW)
pi_habit_swim	HABIT	percent individuals - Habit - swimmers (SW)
pt_habit_swim	HABIT	percent taxa - Habit - swimmers (SW)
nt_volt_multi	VOLT	number of taxa - multivoltine (MULTI)
pi_volt_multi	VOLT	percent individuals - multivoltine (MULTI)
pt_volt_multi	VOLT	percent taxa - multivoltine (MULTI)
nt_volt_semi	VOLT	number of taxa - semivoltine (SEMI)
pi_volt_semi	VOLT	percent individuals - semivoltine (SEMI)
pt_volt_semi	VOLT	percent taxa - semivoltine (SEMI)
nt_volt_uni	VOLT	number of taxa - univoltine (UNI)
pi_volt_uni	VOLT	percent individuals - univoltine (UNI)
pt_volt_uni	VOLT	percent taxa - univoltine (UNI)
nt_ti_cc	TEMP	number of taxa - thermal indicator - cold/cool
pi_ti_cc	TEMP	percent individuals - thermal indicator - cold/cool
pt_ti_cc	TEMP	percent taxa - thermal indicator - cold/cool
nt_ti_w	TEMP	number of taxa - thermal indicator - warm
pi_ti_w	TEMP	percent individuals - thermal indicator - warm
pt_ti_w	TEMP	percent taxa - thermal indicator - warm

C2 Summary of input metrics in existing low gradient IBIs

Vermont DEC (in progress; personal communication Aaron Moore)

Hybrid low gradient (HLG)

1. Density
2. EOT Richness
3. BCG intolerant richness
4. BCG intolerant COTE %
5. Modified EOT/EOT+Chiro
6. PMA-O
7. Amphipoda+Isopoda %
8. Biotic Index
9. PPCS-F
10. Shr%/CF+Shr%

Soft/slow low gradient (SLG)

1. Density
2. EOT Richness
3. BCG intolerant richness
4. BCG intolerant COTE %
5. Modified EOT/EOT+Chiro
6. PMA-O
7. Amphipoda+Isopoda %
8. Biotic Index
9. PPCS-F
10. Modified EOT Density

New York State DEC (in progress; personal communication Gavin Lemly)

Provisional IBIs by regions for low-gradient streams for three regions:

Great Lakes

- rich_family: decrease with stress
- pct_dom1_order: increase with stress
- shannon_family: decrease with stress
- rich_scraper: decrease with stress

Adirondacks

- pct_insecta: decrease with stress
- rich_mollusca_amphipoda_fa: increase with stress
- rich_intolerant: decrease with stress
- rich_et_macro_genspecies: decrease with stress

Hudson Valley+Southern Tier:

- pct_rich_cote_family: decrease with stress
- pct_et: decrease with stress
- pct_filterer: decrease with stress
- shannon_genus: decrease with stress

Location	Index Identifier	Metrics	Metric response to stress	Citation	Remarks
Coastal plain region of the 6 states (New Jersey, Delaware, Maryland, Virginia, North Carolina, and South Carolina)	Coastal Plain Macroinvertebrate Index (CPMI)	# of taxa # of EPT taxa % Ephemeroptera Hilsenhoff Biotic Index (HBI) % clingers	**# taxa**: decrease; 45% overall assessment accuracy **# EPT taxa**: decrease; high assessment accuracy (84% overall); correlated with Ephem and Trichop metrics; historic reliability. **% Ephemeroptera**: decrease; 57% overall-assessment accuracy; lower redundancies with HBI & # EPT; high redundancy with % EPT metrics already selected **HBI**: increase; high assessment accuracy (80%overall); strongly correlated with other tolerance metrics; historic reliability. **% clingers**: decrease; not redundant with TT and %E metrics already selected, moderately redundant with the HBI and EPT metrics	Maxted et al. 2000	Oct/Nov sampling period. Accurately identified 86% of impaired sites overall (varied 83-100% across the 3 regions classified). 90% CI for the 5 core metrics were ±6.0 taxa for TT, ±2.5 taxa for EPT, ±8.9% for %E, ±0.28 units for the HBI, ±13.8% for %CL, and ±3.1 units for the CPMI.

Location	Index Identifier	Metrics	Metric response to stress	Citation	Remarks
Florida	Stream Condition Index (SCI)	# total taxa # EPT taxa # Chironomidae taxa Florida Index % dominant taxa % Diptera % gatherers % filterers	**# total taxa**: decrease **# EPT taxa**: decrease **# Chironomidae taxa**: decrease **Florida Index**: decrease **% dominant taxa**: increase **% Diptera**: increase **% gatherers**: variable **% filtere**rs: decrease; "filter feeders are also thought to be sensitive in low-gradient streams (Wallace et al. 1977)."	Barbour et al. 1996	Summer index sampling period (Jul-Sep). 3 classified regions: panhandle, peninsular Florida, & the northeastern portion of Florida Scores (5, 3, or 1) developed for 8 metrics to allow aggregation into an index

Location	Index Identifier	Metrics	Metric response to stress	Citation	Remarks
Choctawhatchee-Pea Rivers watershed, AL	Invertebrate community index (ICI)	# EPT taxa # Trichoptera taxa # Diptera taxa # Crustacea + Mollusca % Dominant taxa % Ephemeroptera % Diptera % Chironominae to chironomids Family Biotic Index (FBI) % Shredders	# EPT taxa: decrease # Trichoptera taxa: decrease # Diptera taxa: decrease # Crustacea + Mollusca: decrease % Dominant taxa: increase % Ephemeroptera: decrease % Diptera: increase % Chironominae to chironomids: decrease Family Biotic Index (FBI): increase % Shredders: decrease	Bennet et al. 2004	Within the coastal plains ecoregion in southeast Alabama; low elevation and loosely compacted, sandy soils. 34 wadeable first through sixth-order streams; plus for validation 7 additional least impacted and 8 impacted streams. 49 sites sampled once during April and May 2001. The 10 selected metrics (of 38 tested) had significant correlations with one or more physiochemical variables. ICI calculated by summing the 10 metric scores from 34 sites; ranged from 18 to 56 out of a possible score of 60. The ICI was not always capable of discriminating between artificially enriched sites and good quality sites

Location	Index Identifier	Metrics	Metric response to stress	Citation	Remarks
Central Valley, CA	Central Valley IBI	collector richness predator richness percent EPT taxa percent clinger taxa Shannon diversity	collector richness: decrease predator richness: decrease percent EPT taxa: decrease percent clinger taxa: decrease Shannon diversity: decrease *Note: these expectations deduced from the scoring ranges presented in Table 2 of paper.*	Rehn et al. 2008	Perennial streams on the valley floor In the Central Valley, minimally disturbed reference sites no longer available. Most streams are highly alteredby human activities such as urbanization, agriculture and water diversions. 80 metrics evaluated; metric criteria: 1) sufficient range for scoring; 2) responsiveness to land use and reach-scale disturbance variables (as data allowed); 3) good discrimination between reference and test sites; 4) lack of correlation with other responsive metrics. Lack of intolerant and shredder taxa in Valley floor streams. Final IBI more strongly related to reach-scale physical habitat variables than to water chemistry or land use variables. The final 5 IBI metrics did not vary between spring and fall samples and did not require seasonal adjustments in scoring.

Location	Index Identifier	Metrics	Metric response to stress	Citation	Remarks
Austria	Multimetric index (for A01 - Mid-sized (low-gradient) streams in the Hungarian Plains)	# of total families # of EP # taxa # of Plecoptera (abundance) [%] EP # individuals [%] EP # taxa Saprobic index # of sensitive taxa [%] Shredder Diversity (Margalef)	# of total families: decrease # of EP # taxa: decrease # of Plecoptera (abundance): [%] EP # individuals: decrease [%] EP # taxa: decrease Saprobic index: increase # of sensitive taxa: decrease [%] Shredder: decrease Diversity (Margalef): decrease	Ofenböck et al. 2004	Stressor – organic pollution

Location	Index Identifier	Metrics	Metric response to stress	Citation	Remarks
Willow Creek, Nebraska	Composite Biotic Index (CBI)	Percent dominance EPT index (i.e. # EPT taxa) EPT abund/EPT + chironomid abundance Scraper abund/filterer abund Taxa richness Hilsenhoff index	Not specified in paper	Whiles et al. 2002	Developed with metrics used previously by the NDEQ during their statewide stream survey (NDEQ 1991). CBI scores actually are based on a “reference condition” for Nebraska rather than the reference stream (site 4) in our basin Corrected metrics for stream size (based on discharge) using relationships generated from a prior investigation; i.e. metrics were scored 1, 3, or 5 based on regression equations generated by the NDEQ (1991) that divided scatter plots of stream size vs metric scores into thirds

Location	Index Identifier	Metrics	Metric response to stress	Citation	Remarks
Netherlands	Multimetric index (for slow-running streams)	See metrics listed in Table 2 copied below from paper.	complex	Vlek et al. 2004	Included metrics that indicated the different classes (from 5 (high quality) to 1 (low quality); final index equation combined these; for slow running streams: $$S = \frac{T_1 * \frac{1}{2} + T_2 * \frac{1}{2} + T_3 * \frac{1}{2} + T_4 \frac{1}{4}}{n_1 * \frac{1}{2} + n_2 * \frac{1}{2} + n_3 * \frac{1}{2} + n_4 * \frac{1}{4}}$$ Where: S, final score; T1, sum of scores for the individual metrics indicating class 1; T2, sum of scores for the individual metrics indicating class 2; etc. And n1, number of indices indicating class 1; etc. Validation showed that 54% of the streams were classified correctly

References – existing low gradient IBIs

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

Bennett, H.H., M.W. Mullen, P.M. Stewart, J.A. Sawyer, and E.C. Webber. 2004. Development of an invertebrate community index for an Alabama coastal plain watershed. Journal of the American Water Resources Association 40(1): 43-51. https://doi.org/10.1111/j.1752-1688.2004.tb01008.x|.

Lemley, G.M., B.T. Duffy, Z.M. Smith, A.J. Smith, and R.G. Smith. 2019. Assessing Low Gradient Streams in New York State: Developing a Macroinvertebrate IBI. PowerPoint Presentation, February 28th, 2019. New York Department of Environmental Conservation.

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

Mazor, R.D., K.C. Schiff, K.J. Ritter, A. Rehn, and P. Ode. 2009. Bioassessment tools in novel habitats: An evaluation of indices and sampling methods in low-gradient streams in California. Environmental Monitoring and Assessment 167:91–104.

Ofenböck. T., O. Moog, J. Gerritsen, and M. Barbour. 2004. A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. Hydrobiologia 516:251–268.

Rehn, A.C., J.T. May, and P.R. Ode. 2008. An Index of Biotic Integrity (IBI) for Perennial Streams in California's Central Valley. Surface Water Ambient Monitoring Program, California Water Boards, Technical Report, 33 pp.

Vlek, H.E., Piet F. M. Verdonschot & Rebi C. Nijboer. 2004. Towards a multimetric index for the assessment of Dutch streams using benthic macroinvertebrates. Hydrobiologia 516:173–189.

Whiles, M.R., B.L. Brock, A.C. Franzen, and S.C. Dinsmore. 2004. Stream Invertebrate Communities, Water Quality, and Land-Use Patterns in an Agricultural Drainage Basin of Northeastern Nebraska, USA. Environmental Management 26(5):563–576. DOI: 10.1007/s002670010113.

C3 Non-target taxa

The following non-target taxa were excluded from metric calculations:

ORDER	TAXAID	NONTARGET
Hemiptera	Belostoma	TRUE
Hemiptera	Belostomatidae	TRUE
Hemiptera	Corixidae	TRUE
Hemiptera	Gerridae	TRUE
Hemiptera	Gerris	TRUE
Hemiptera	Microvelia	TRUE
Hemiptera	Neoplea striola	TRUE
Hemiptera	Notonecta	TRUE
Sarcoptiformes	Oribatida	TRUE
Hemiptera	Pleidae	TRUE
Hemiptera	Ranatra	TRUE
Hemiptera	Rhagovelia	TRUE
Hemiptera	Veliidae	TRUE

For the purposes of IBI calculations, "macroinvertebrate" is defined to include:

- all aquatic Annelida;
- all aquatic Mollusca;
- aquatic macro Crustacea (except as noted below);
- all aquatic Arachnida except for Oribatid mites (which are not truly aquatic); and
- the aquatic life stages of Insecta except Hemiptera and adult Coleoptera other than Elmidae.

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

C4 Exclusion criteria for redundant taxa

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

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

We used the 'markExcluded' function in the BioMonTools R package (https://github.com/leppott/BioMonTools) to mark redundant taxa in the low gradient samples prior to metric calculations. Redundant taxa were identified on a sample-by-sample basis and excluded from the richness calculations.

Redundant taxa were identified based on the following steps:

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

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

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

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

TAXA LIST

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

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

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

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

Terminology

- Target Rank = intended level of taxonomy for identification, e.g., genus. Typically, specified in the project's SOP but can be adjusted during the OTU process.
- Parent or Parent Taxon = a taxon that occurs in the data in addition to other taxa in the same group that are identified to a more specific level. For example, the family Baetidae may occur in the data in addition to genera within the family Baetidae. In this case the name Baetidae is a parent to the other taxa within the family. Parents do not have to be only a single rank above the child taxon. That is, the class and order ranks are parents of any family ranks within them.
- Child or Children Taxa = a taxa or taxon that occurs in the data in addition to individuals identified to a coarser level. For example, the genera Baetis and Procloeon may occur in addition to the family Baetidae (of which the 2 genera listed are a member). In this case Baetis and Procloen are children of Baetidae.

Rule Development

For each sample:

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

Requirements

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

Procedures

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

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

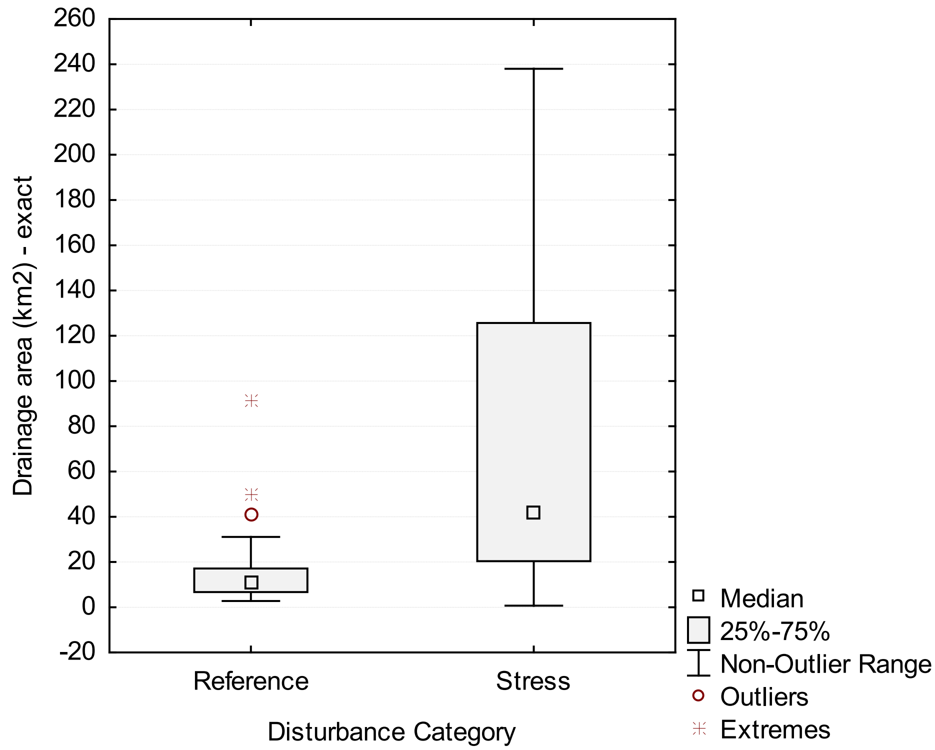
Appendix D

Characterization of reference vs. stressed sites

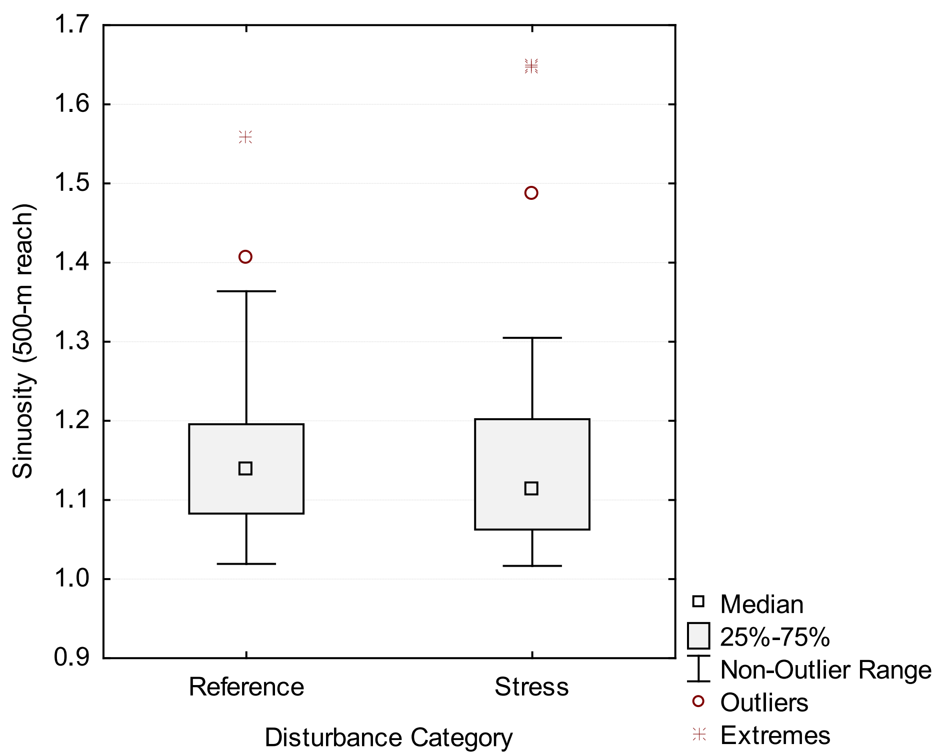
Natural and disturbance variables

Drainage area, sinuosity, slope, elevation, baseflow, temperature, precipitation, ICI, IWI, land cover statistics, RBP habitat assessment score, macroinvertebrate jab allocations, % sediment composition

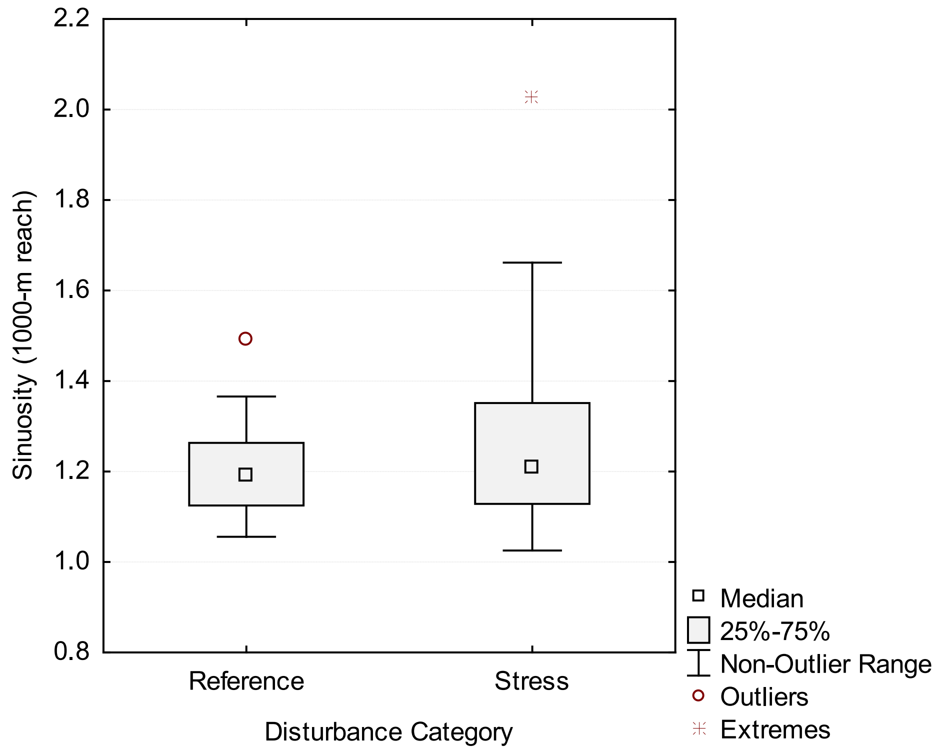
MassDEP & Tetra Tech/SNEP sites
41 reference sites, 41 stressed sites



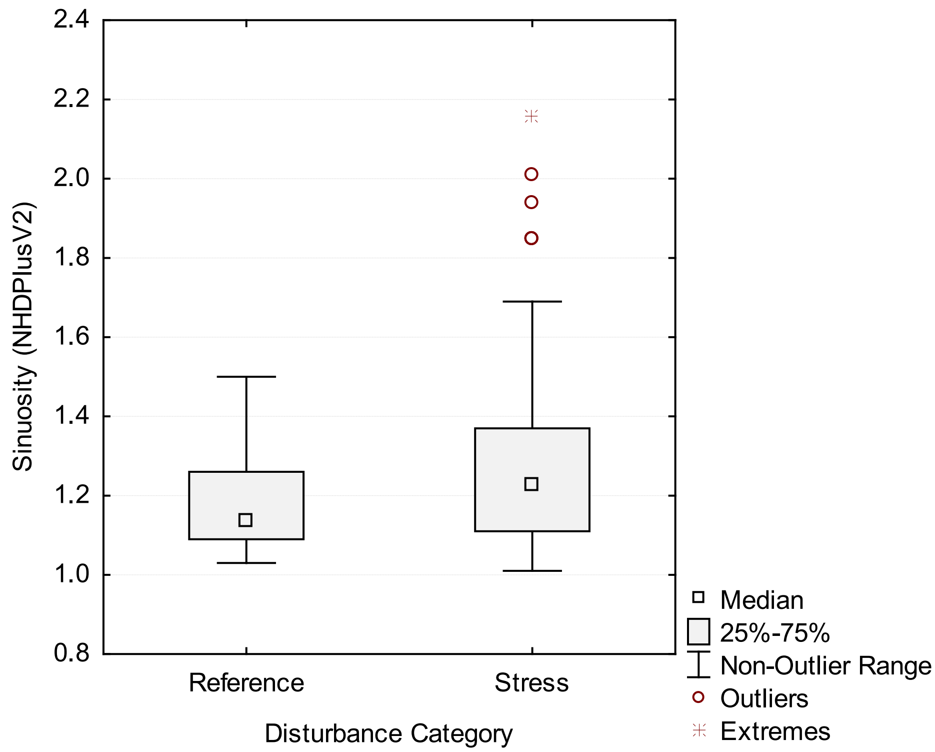
260
240
220
200
180
160
140
120
100
80
60
40
20
0
-20
Drainage area (km2) - exact
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes



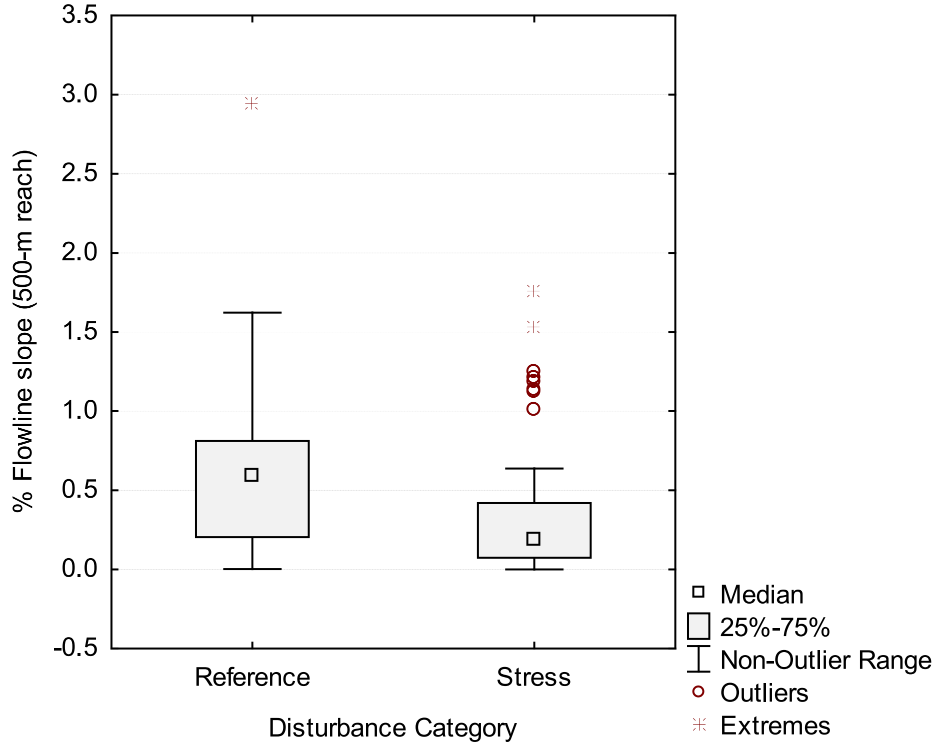
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
Sinuosity (500-m reach)
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

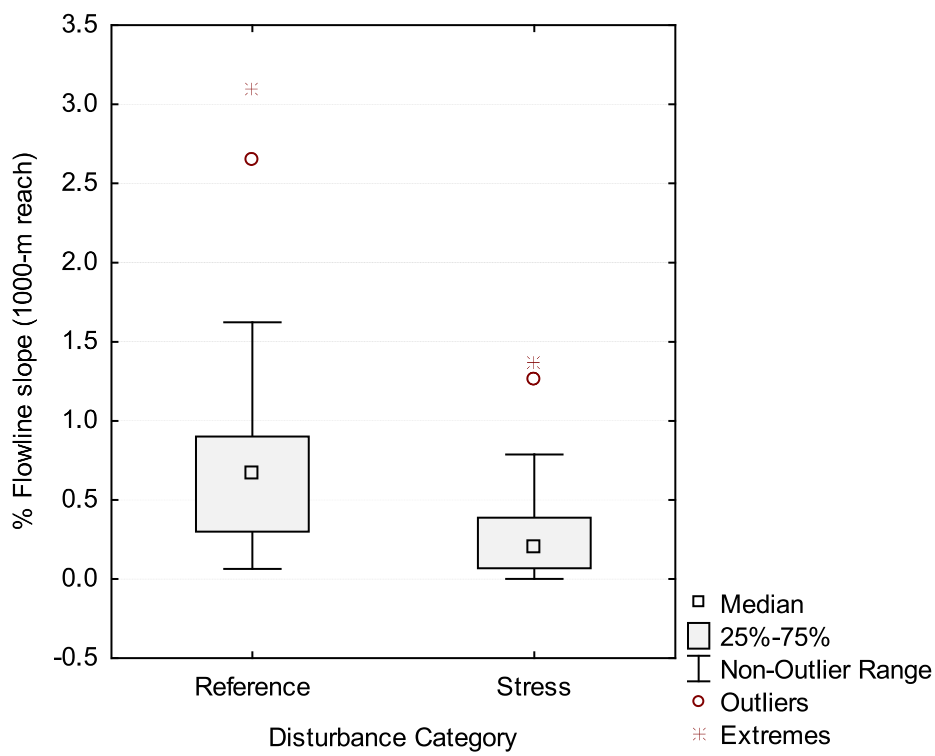


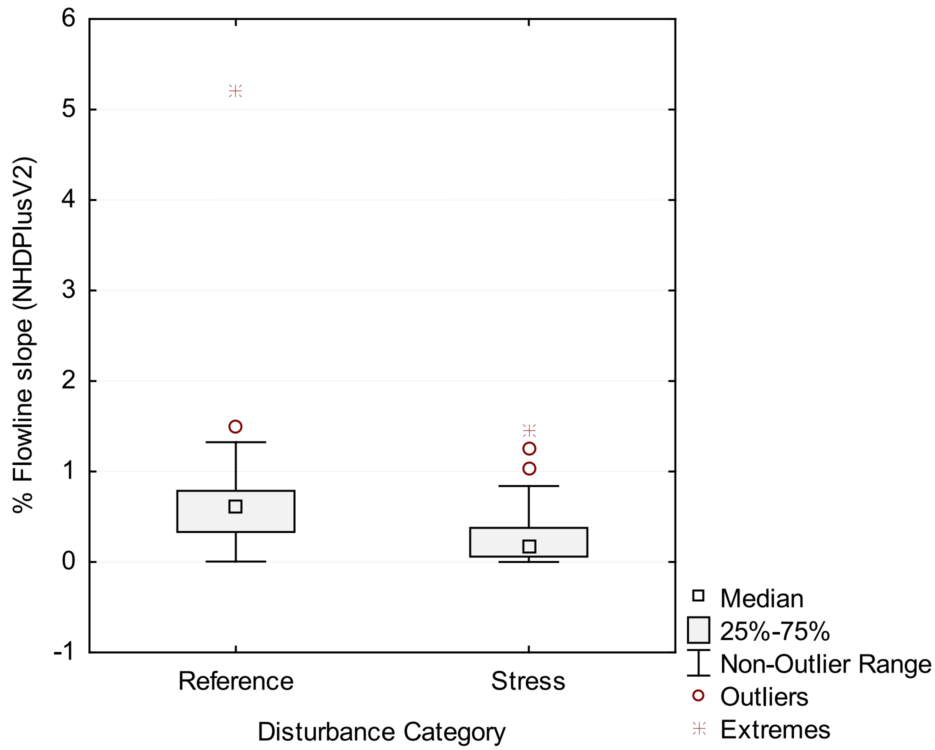
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
Sinuosity (1000-m reach)
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

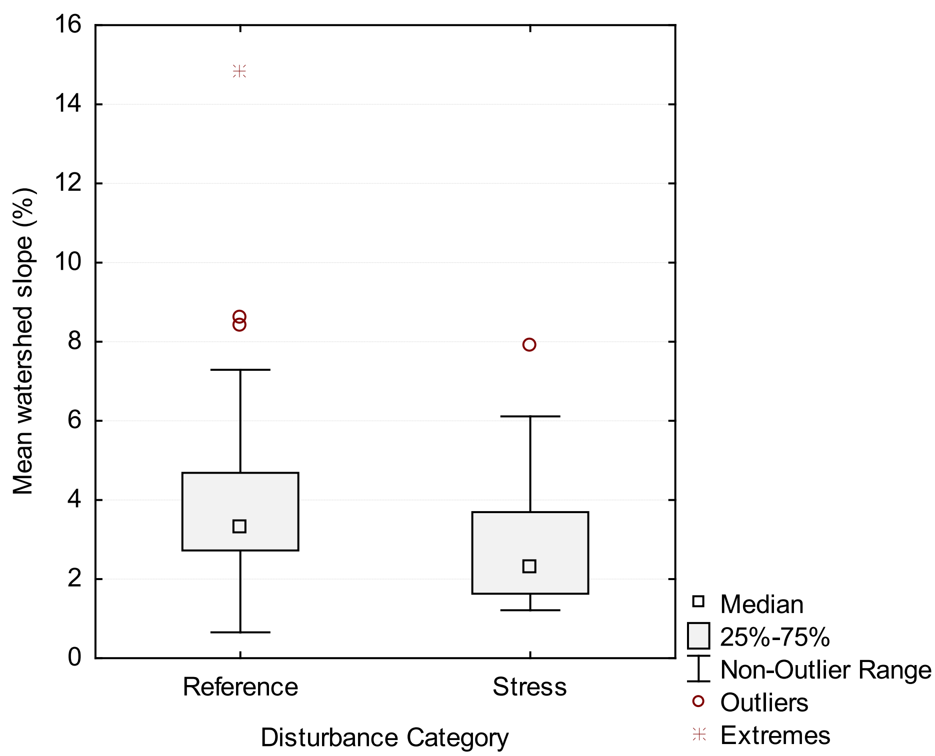


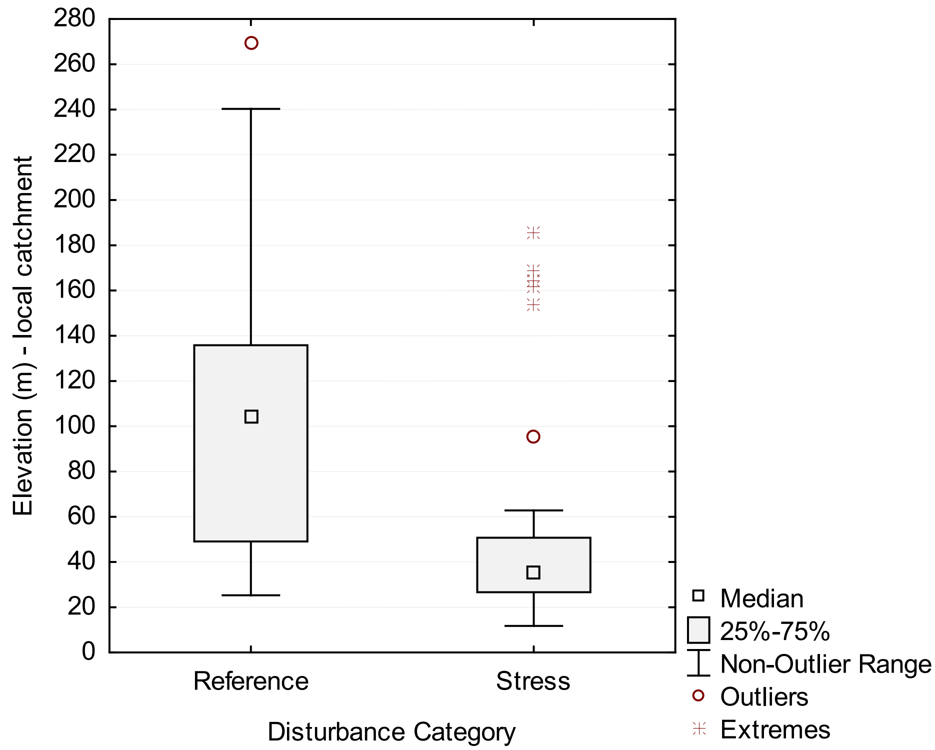
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
Sinuosity (NHDPlusV2)
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

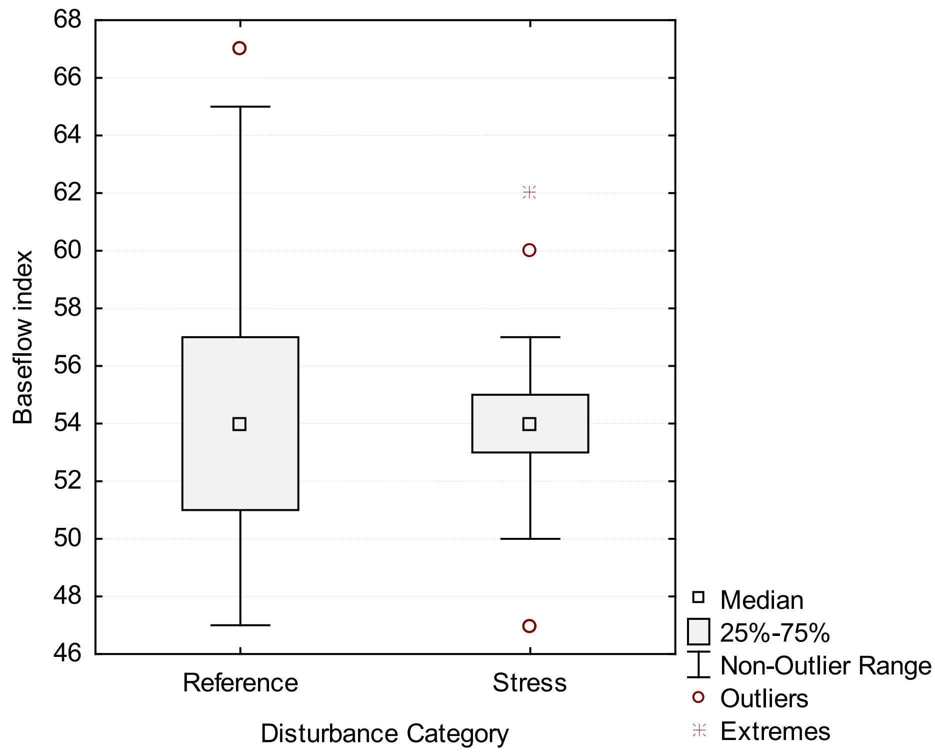

3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
% Flowline slope (500-m reach)
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

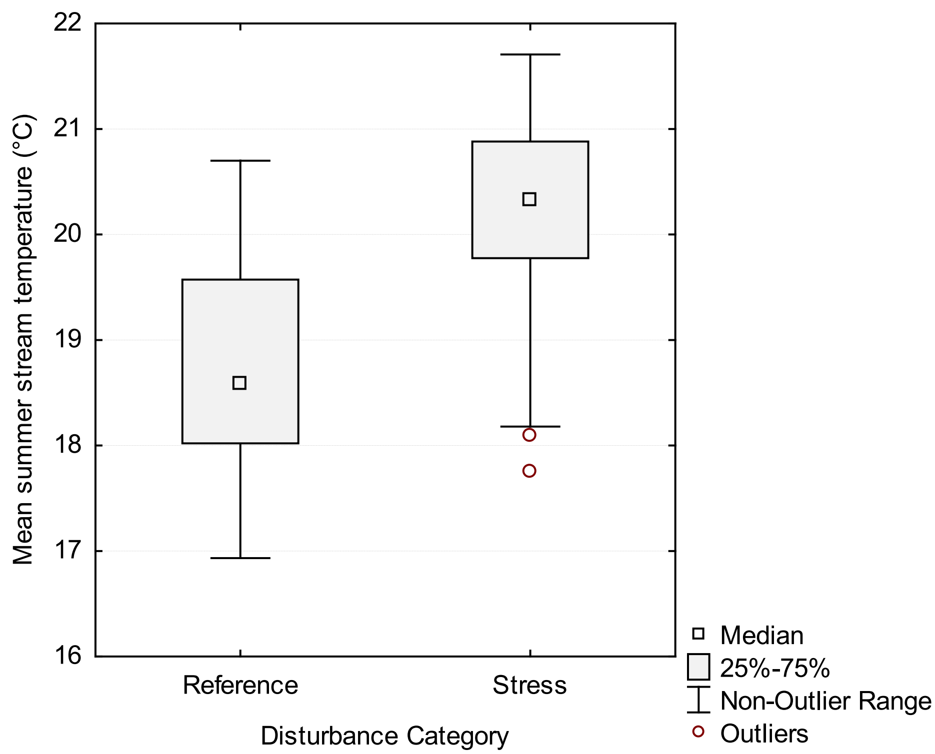

3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
% Flowline slope (1000-m reach)
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes


6
5
4
3
2
1
0
-1
% Flowline slope (NHDPlusV2)
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

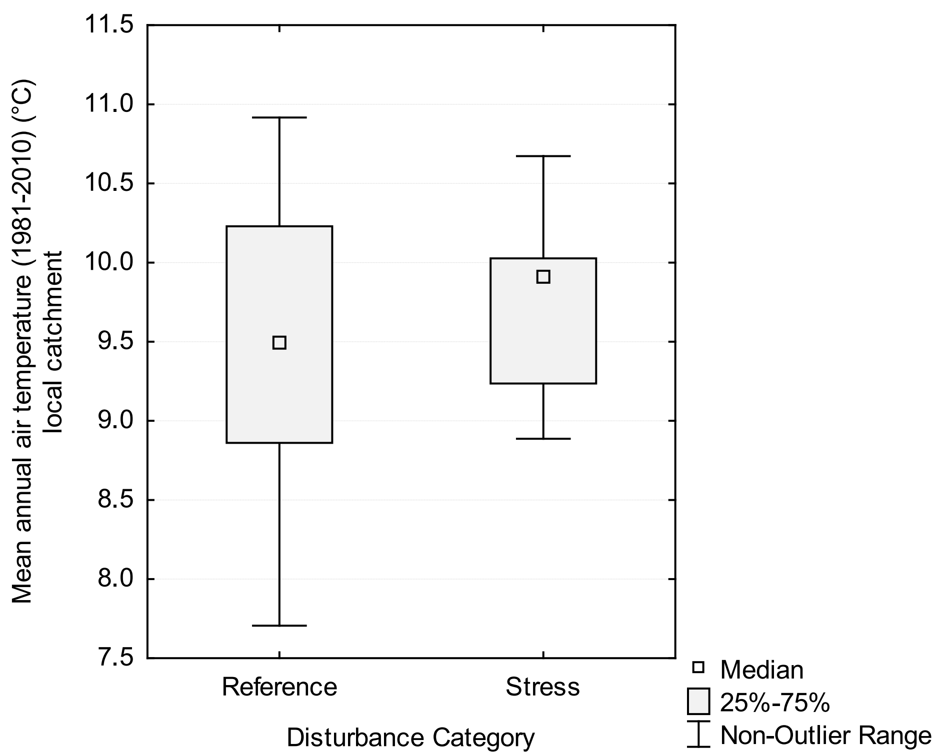

16
14
12
10
8
6
4
2
0
Mean watershed slope (%)
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes


Elevation (m) - local catchment
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

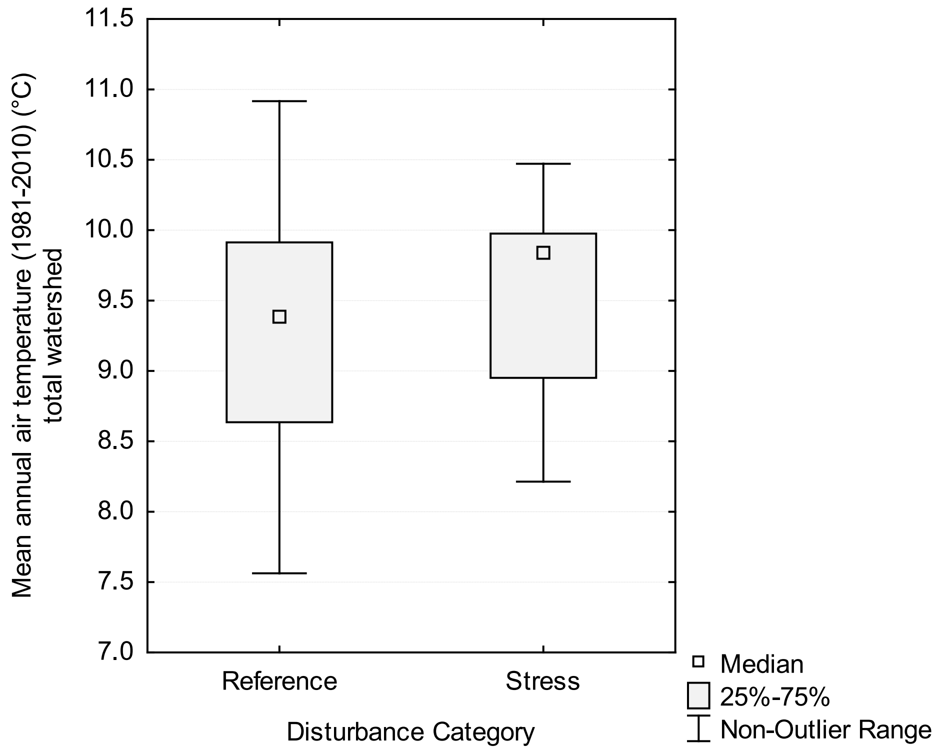

Baseflow index
68
66
64
62
60
58
56
54
52
50
48
46
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

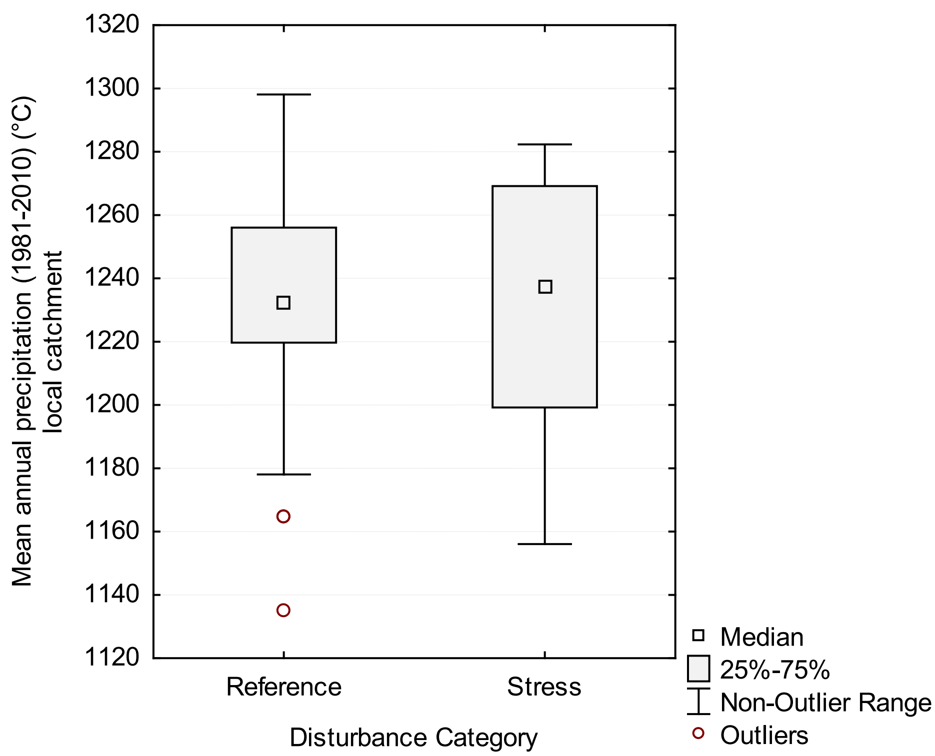


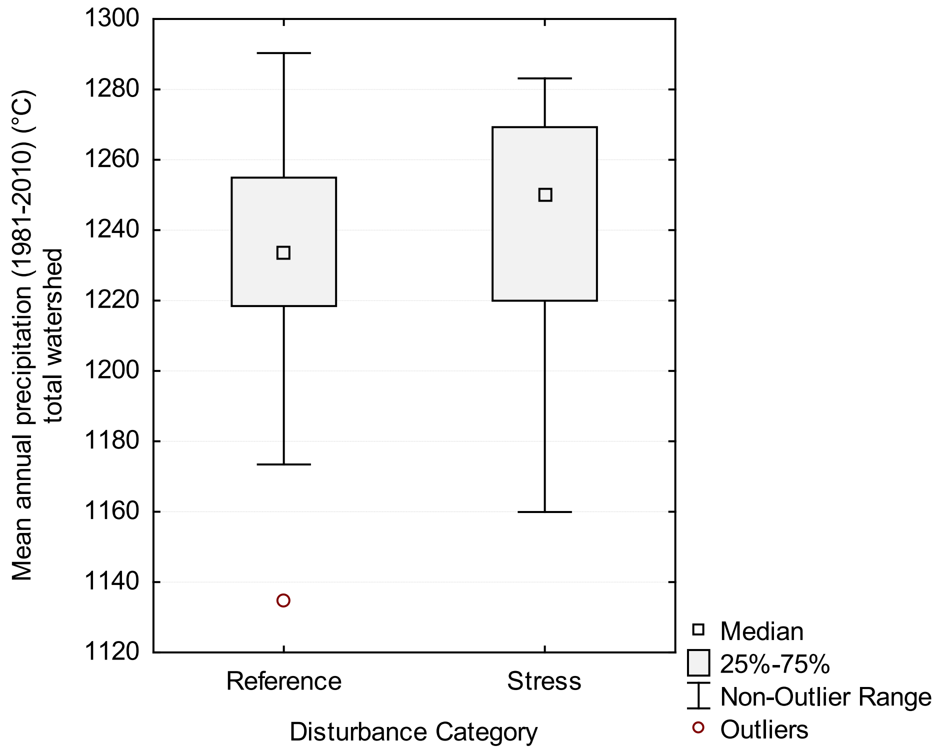
Mean summer stream temperature (°C)
22
21
20
19
18
17
16
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers



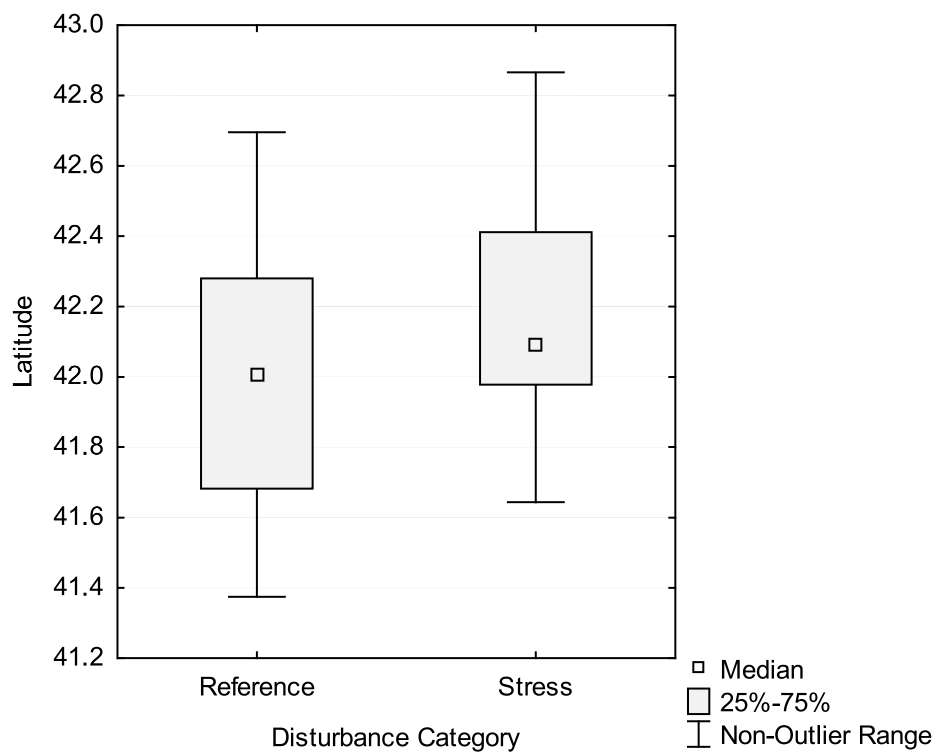
Mean annual air temperature (1981-2010) (°C)
local catchment
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range


Mean annual air temperature (1981-2010) (°C)
total watershed
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range

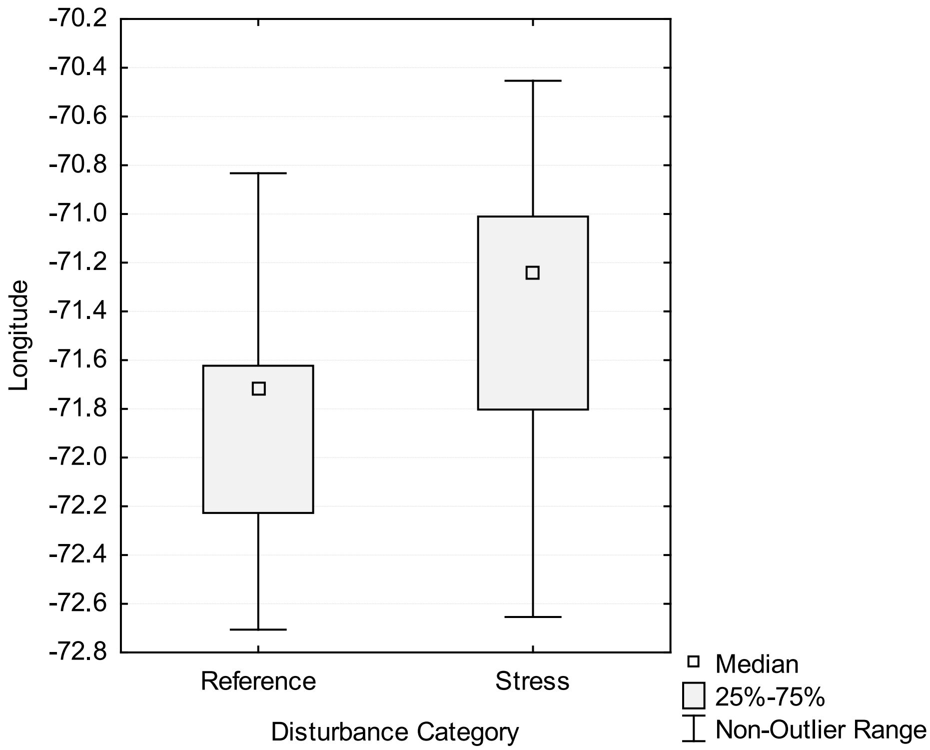

Mean annual precipitation (1981-2010) (°C)
local catchment
1320
1300
1280
1260
1240
1220
1200
1180
1160
1140
1120
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers



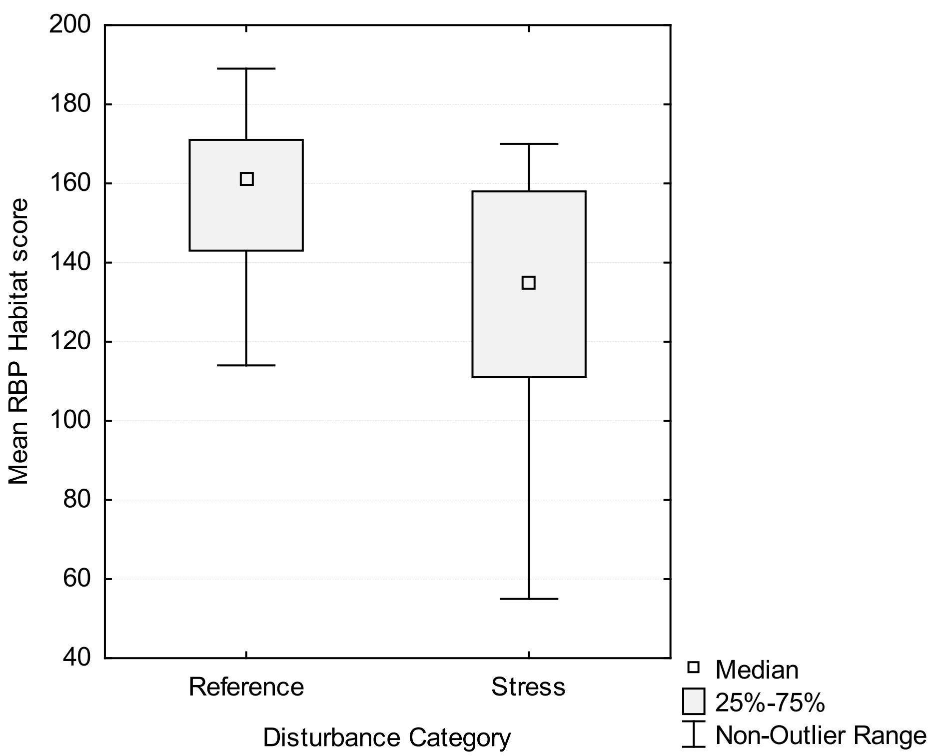
1300
1280
1260
1240
1220
1200
1180
1160
1140
1120
Mean annual precipitation (1981-2010) (°C)
total watershed
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers



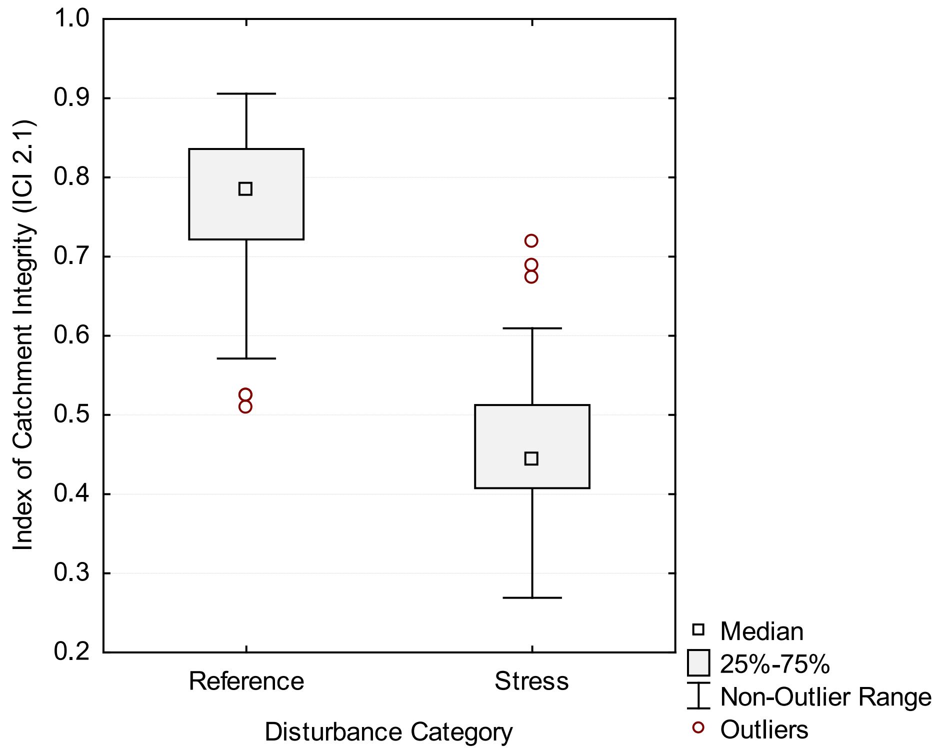
43.0
42.8
42.6
42.4
42.2
42.0
41.8
41.6
41.4
41.2
Latitude
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range

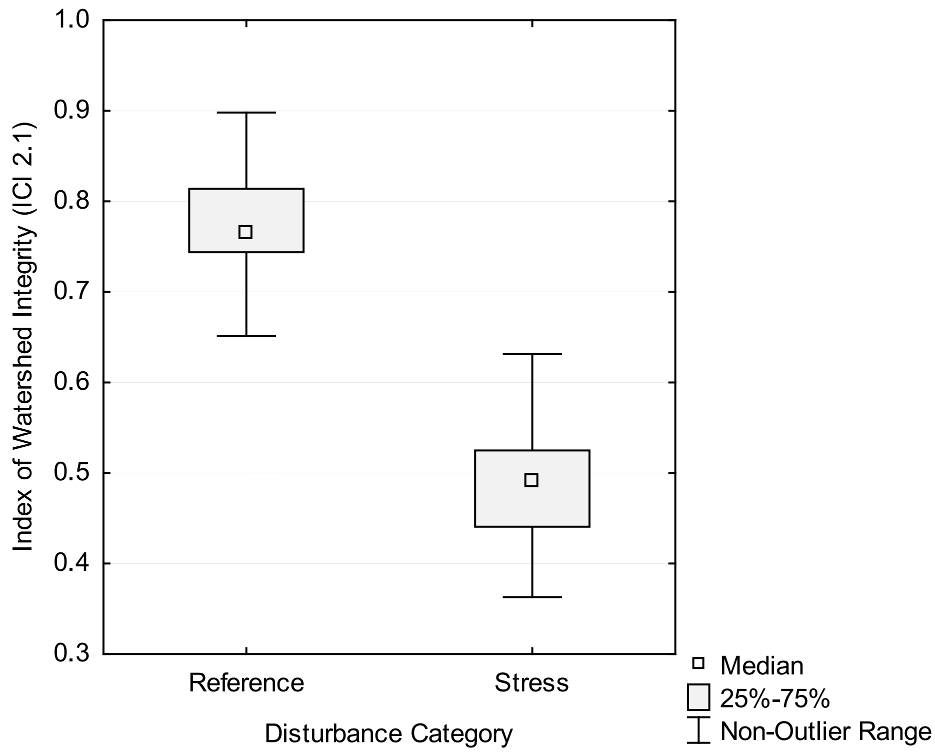


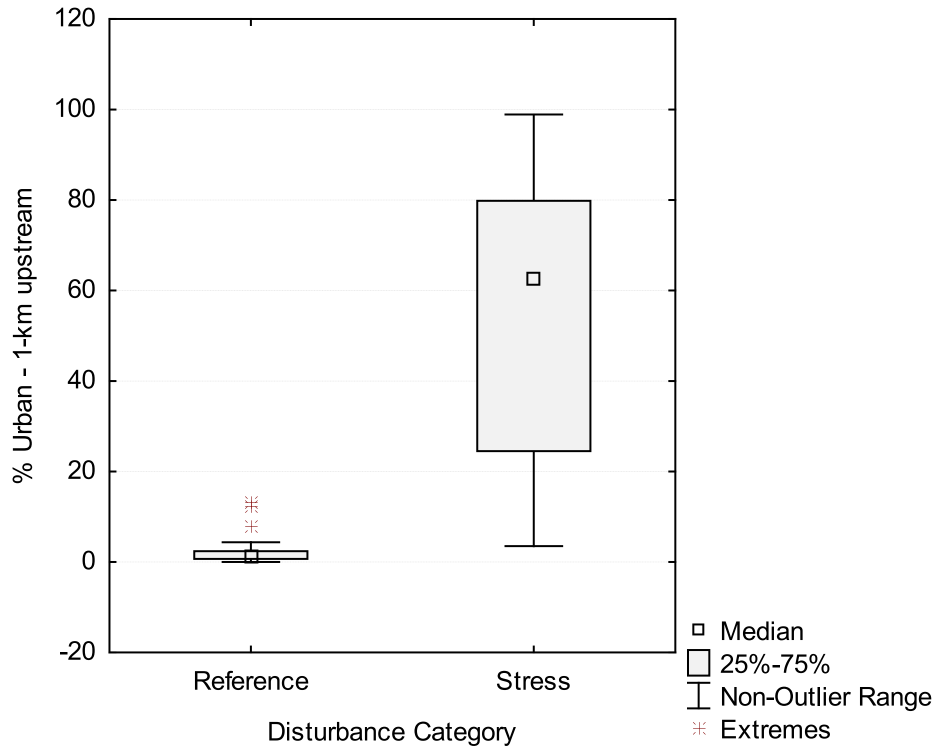
-70.2
-70.4
-70.6
-70.8
-71.0
-71.2
-71.4
-71.6
-71.8
-72.0
-72.2
-72.4
-72.6
-72.8
Longitude
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range

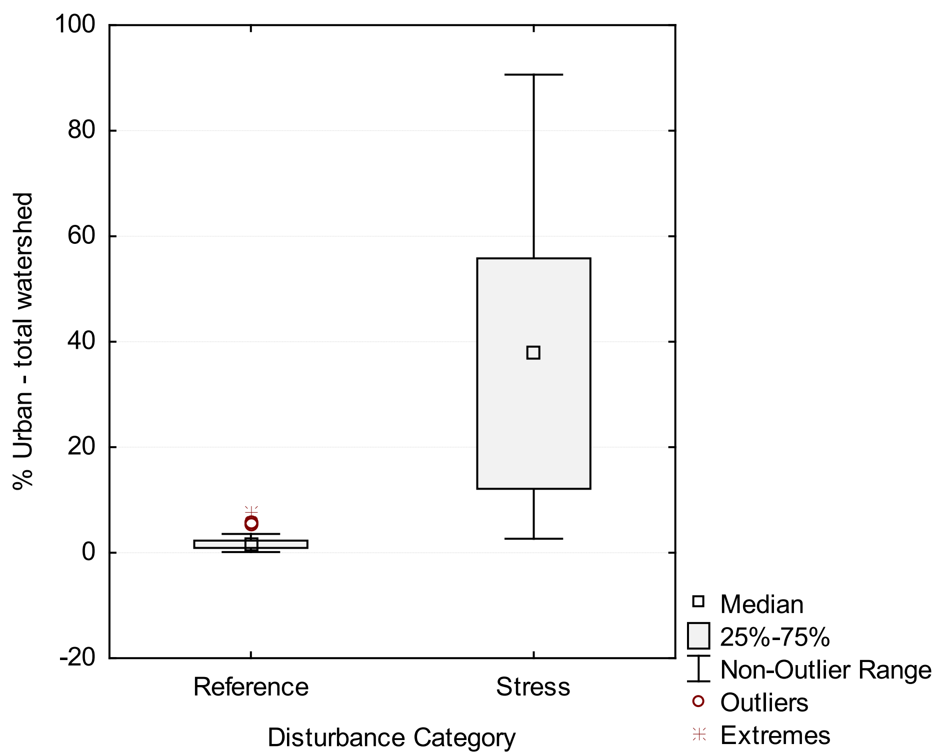


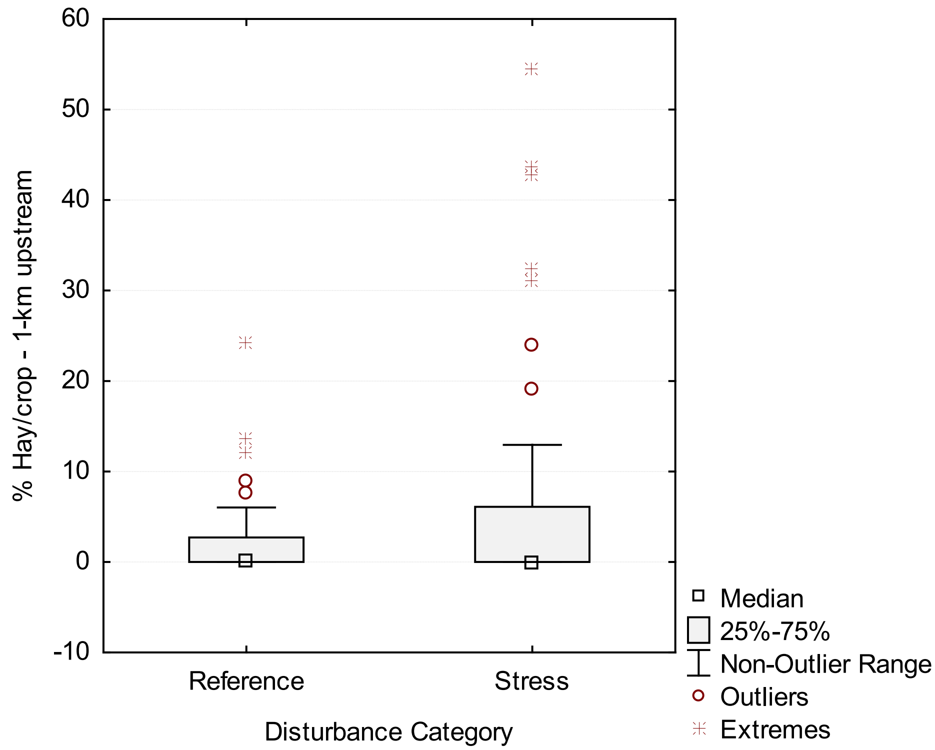
200
180
160
140
120
100
80
60
40
Mean RBP Habitat score
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range

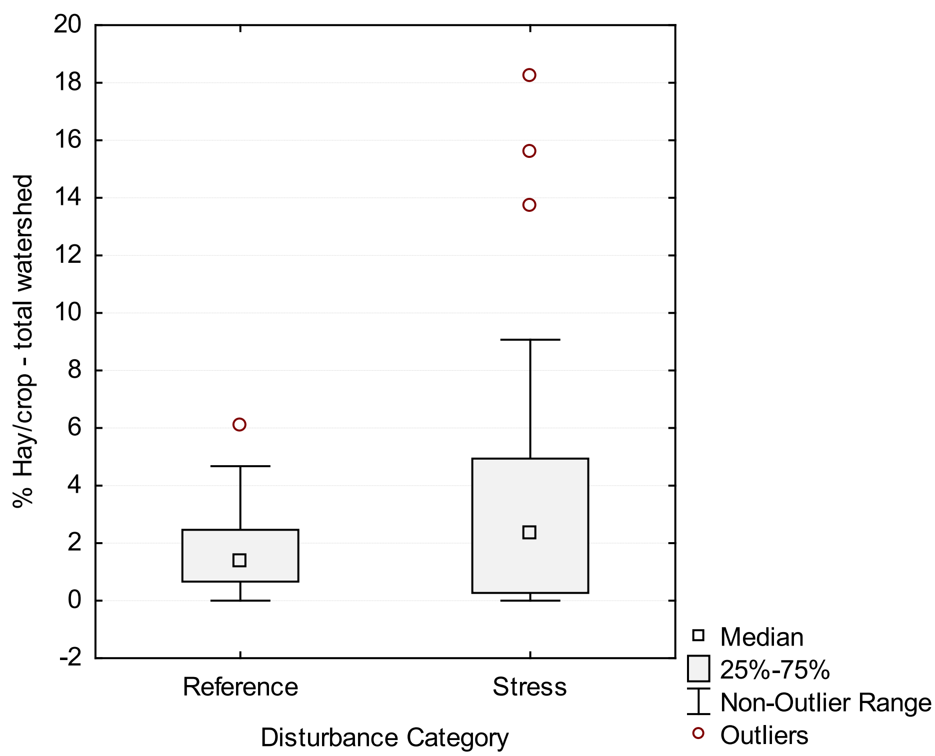

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Index of Catchment Integrity (ICI 2.1)
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers

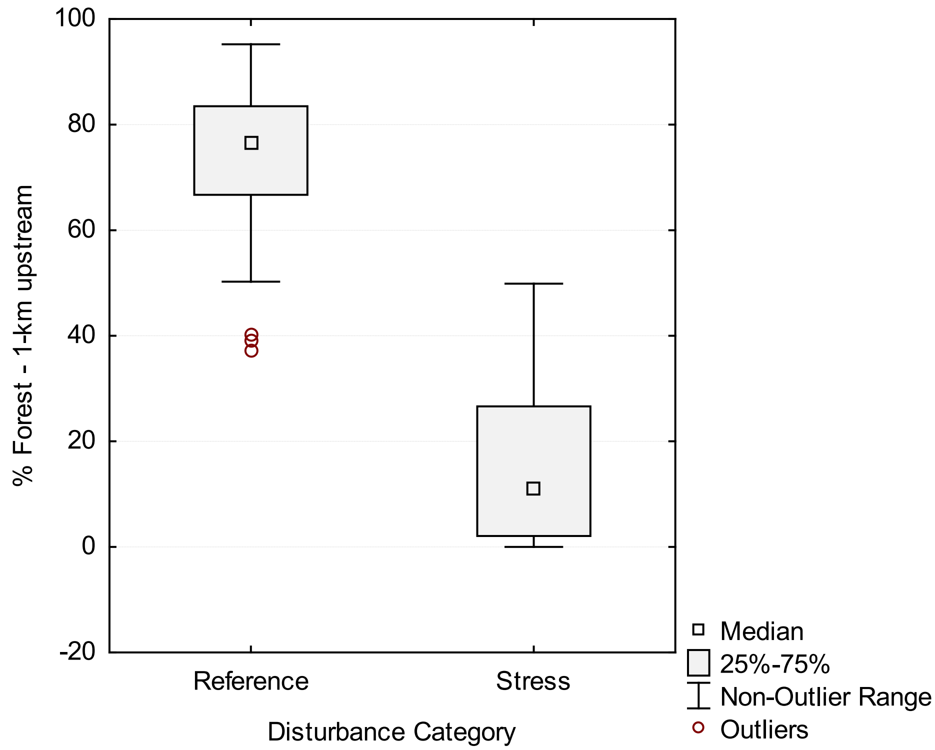

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Index of Watershed Integrity (ICI 2.1)
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range

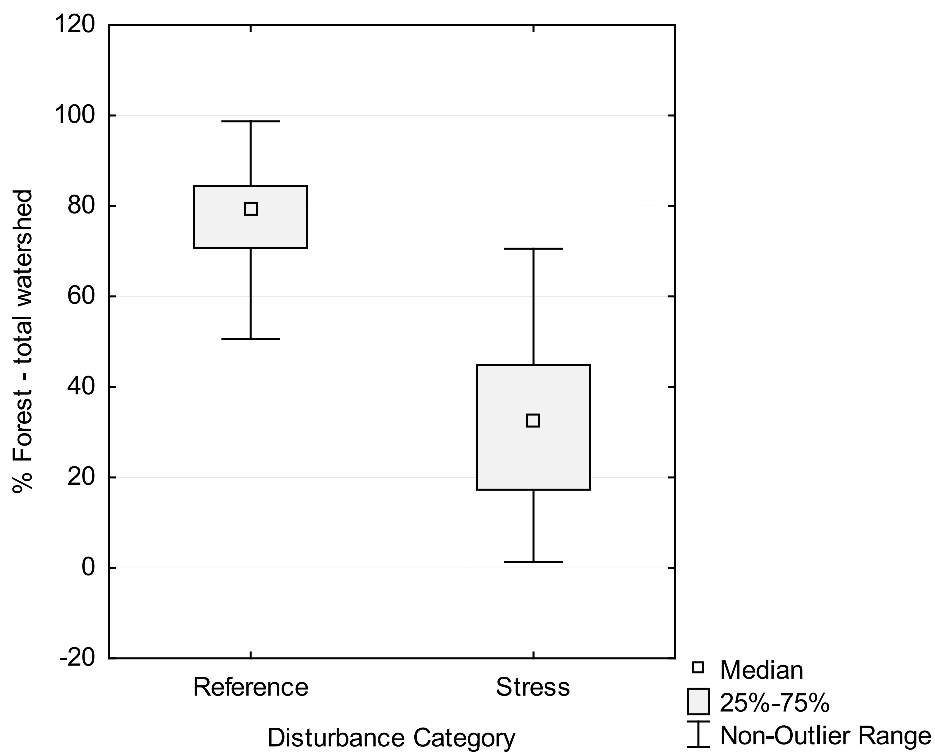

120
100
80
60
40
20
0
-20
% Urban - 1-km upstream
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Extremes

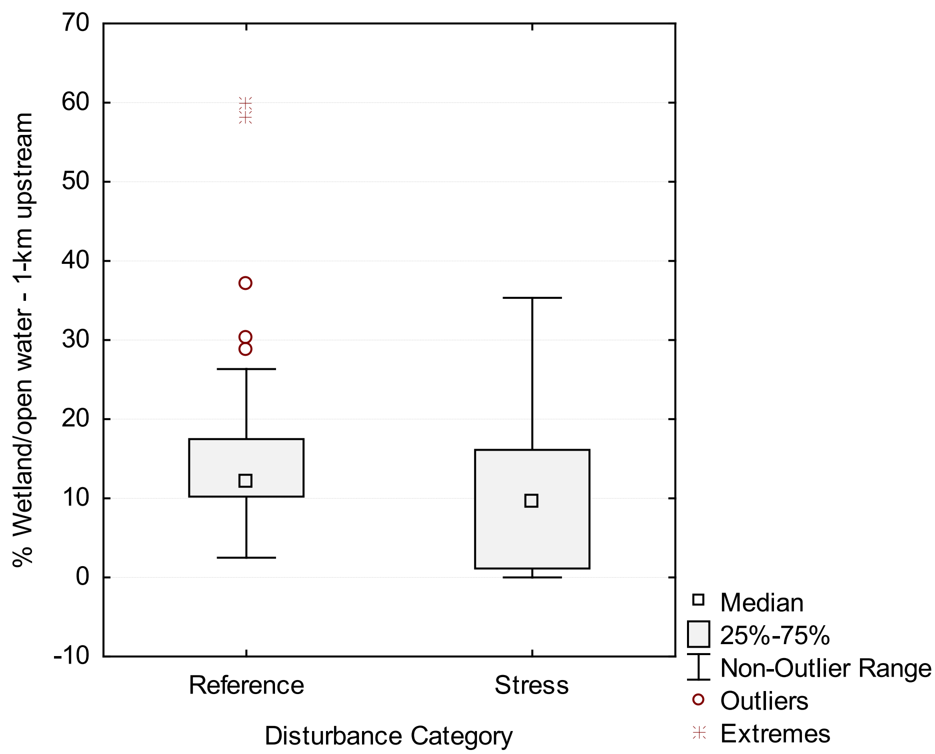

100
80
60
40
20
0
-20
% Urban - total watershed
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

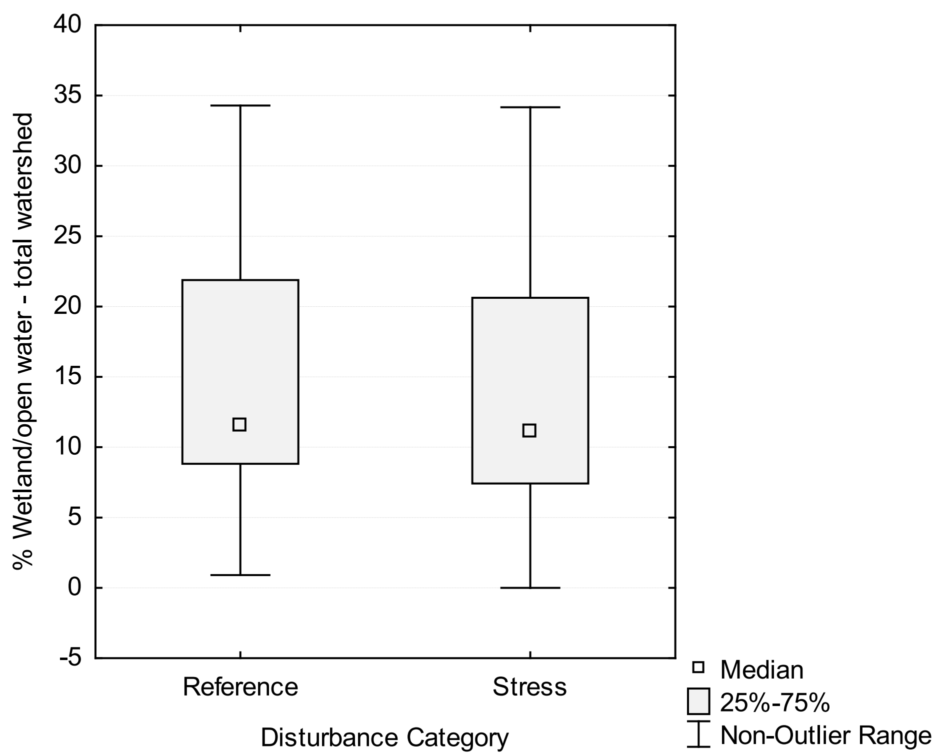

60
50
40
30
20
10
0
-10
% Hay/crop - 1-km upstream
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

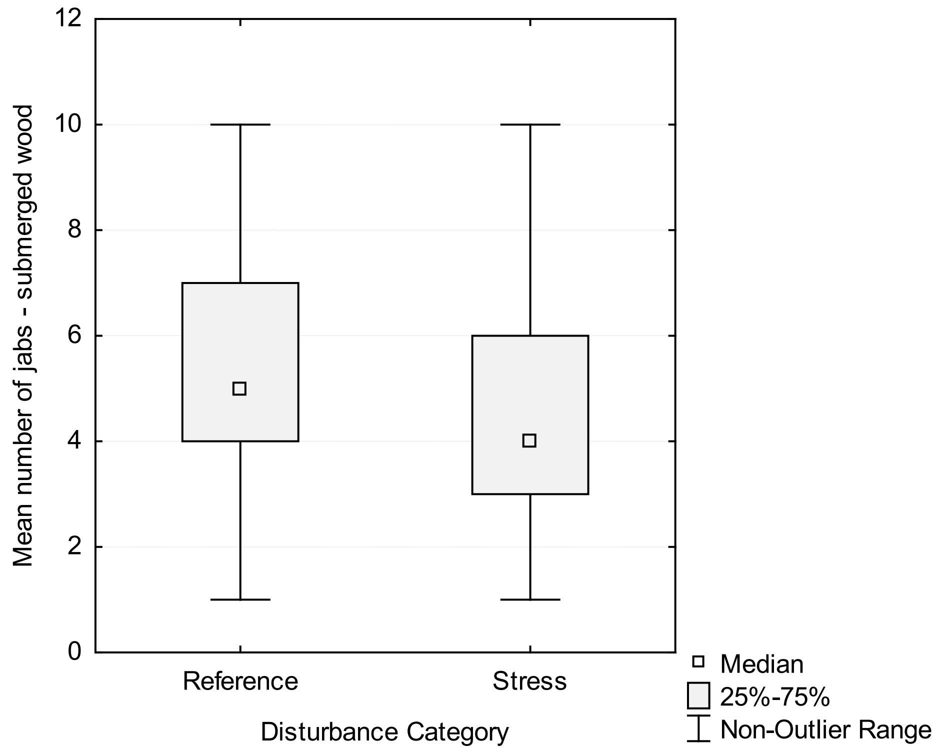

20
18
16
14
12
10
8
6
4
2
0
-2
% Hay/crop - total watershed
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers

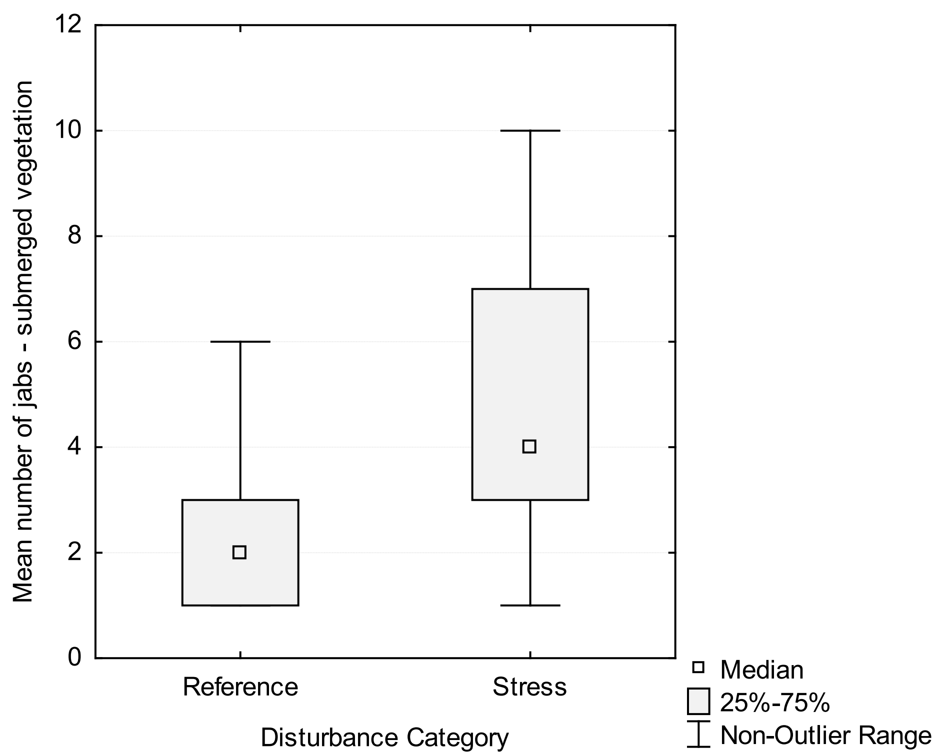

100
80
60
40
20
0
-20
% Forest - 1-km upstream
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers

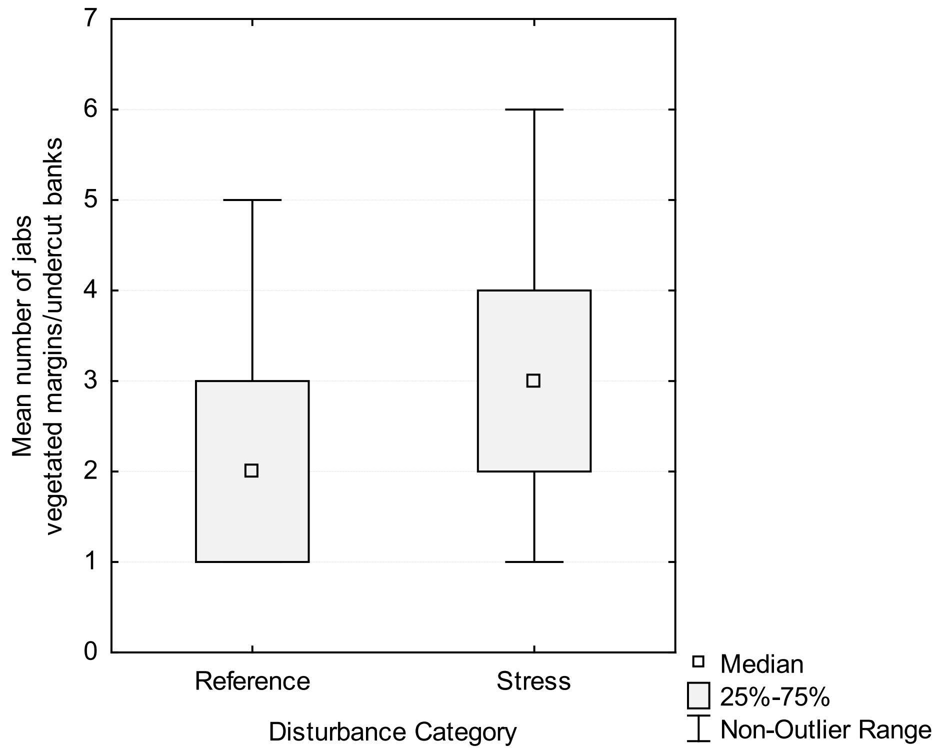

120
100
80
60
40
20
0
-20
% Forest - total watershed
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range


70
60
50
40
30
20
10
0
-10
% Wetland/open water - 1-km upstream
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

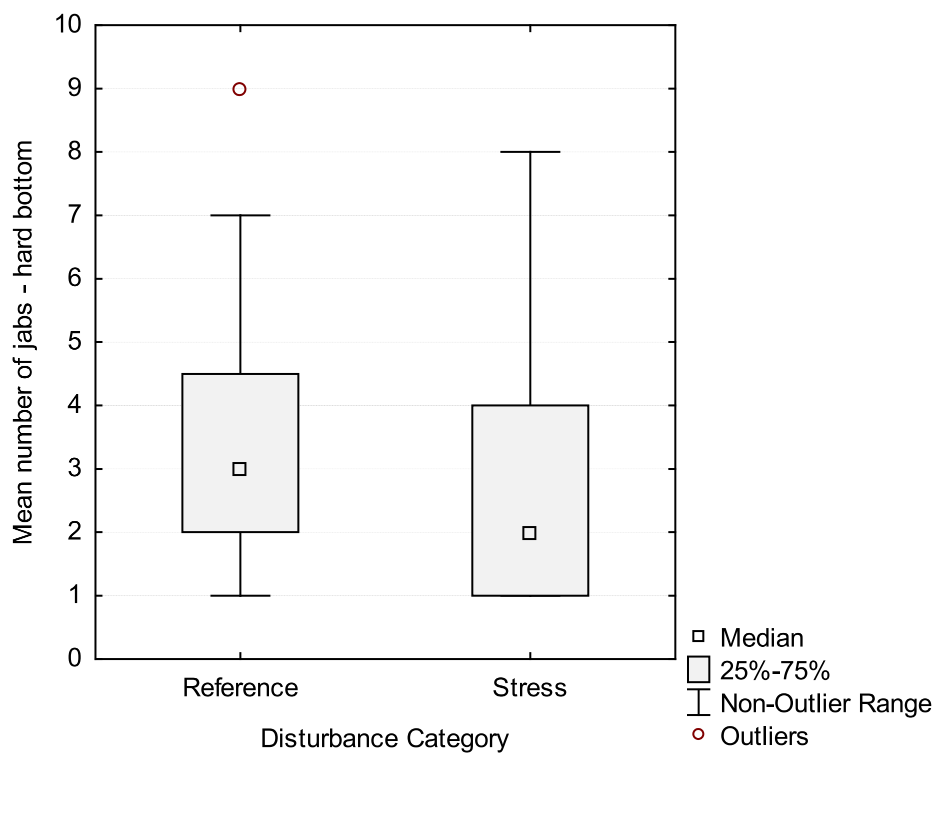

40
35
30
25
20
15
10
5
0
-5
% Wetland/open water - total watershed
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range


12
10
8
6
4
2
0
Mean number of jabs - submerged wood
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range

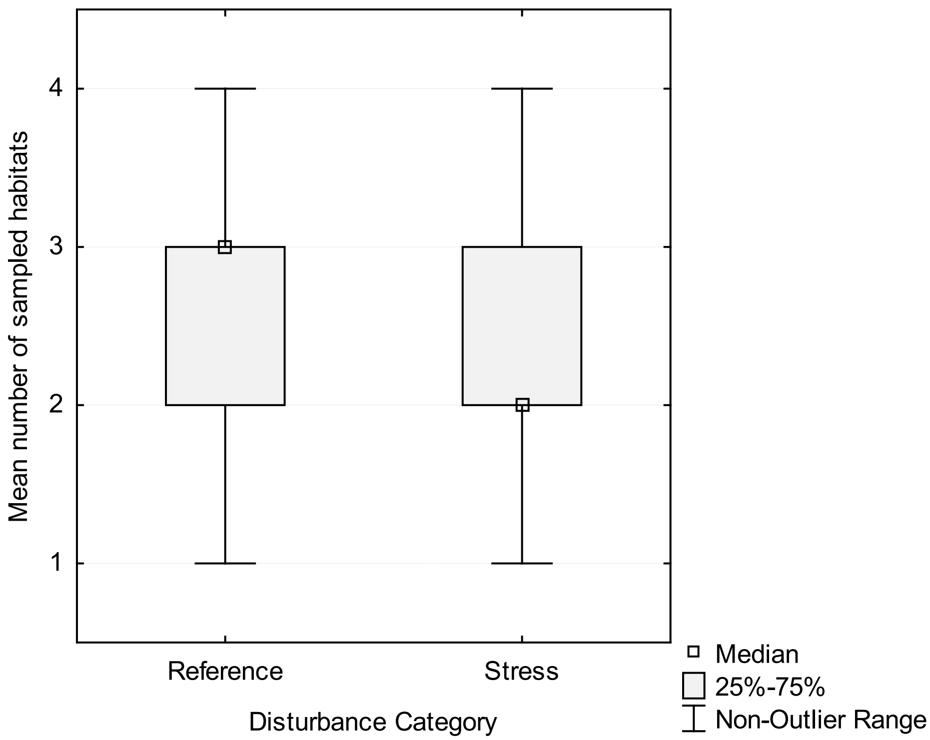

12
10
8
6
4
2
0
Mean number of jabs - submerged vegetation
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range



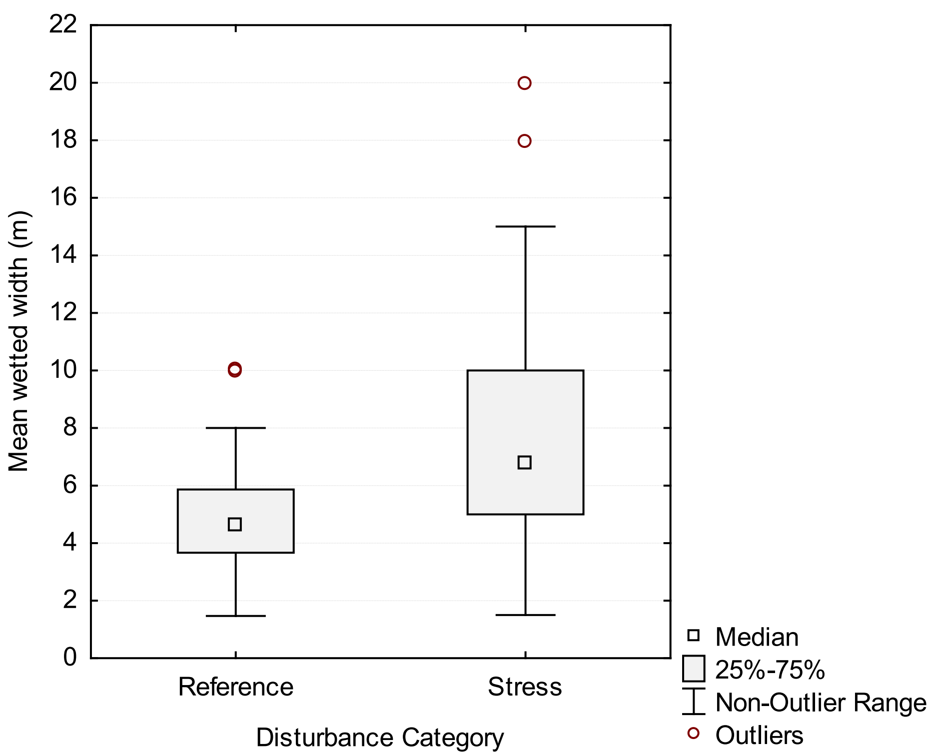
Mean number of jabs
vegetated margins/undercut banks
7
6
5
4
3
2
1
0
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range



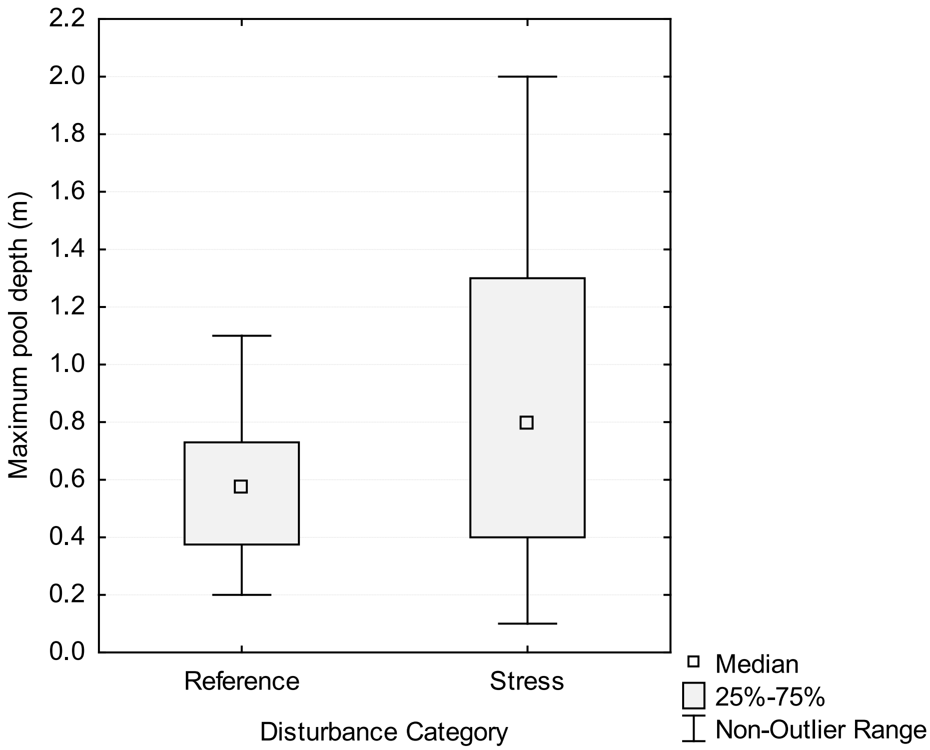
Mean number of jabs - hard bottom
10
9
8
7
6
5
4
3
2
1
0
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers

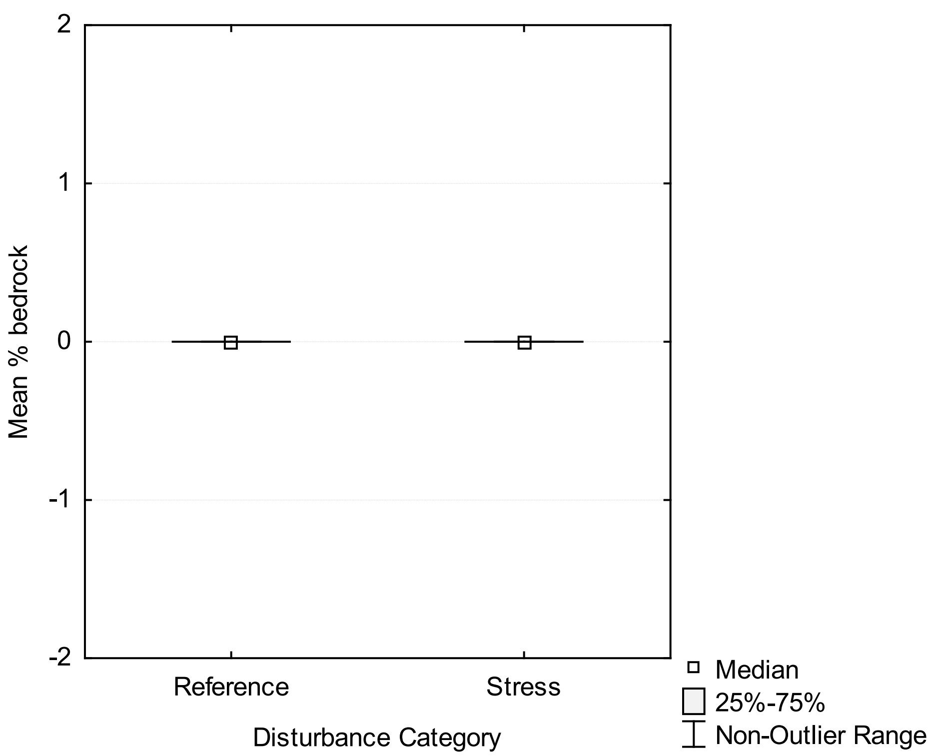


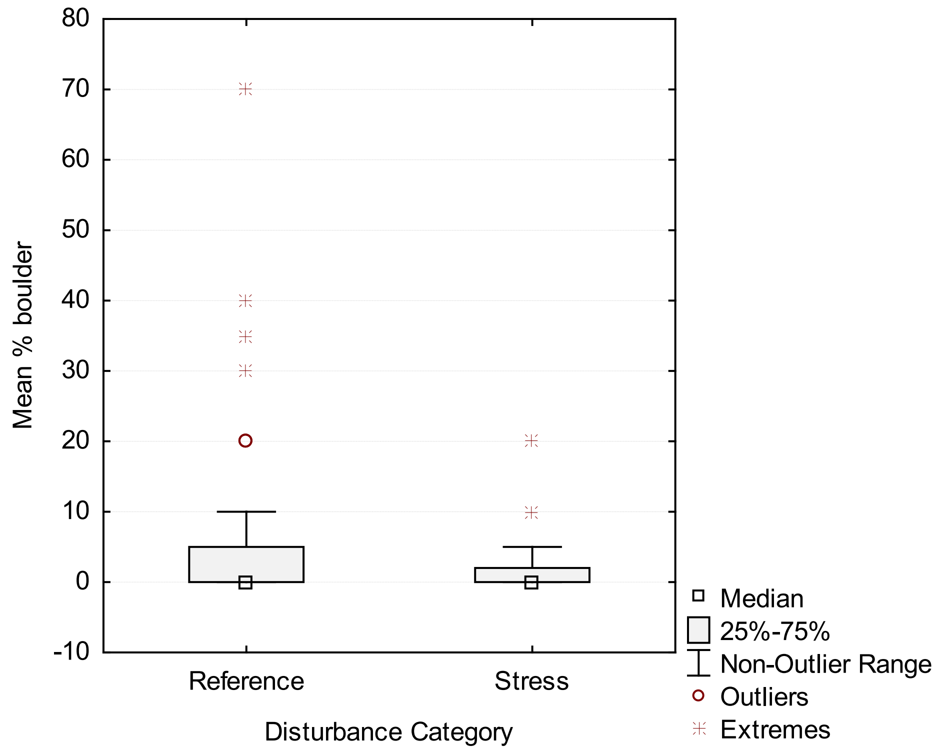
Mean number of sampled habitats
4
3
2
1
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range

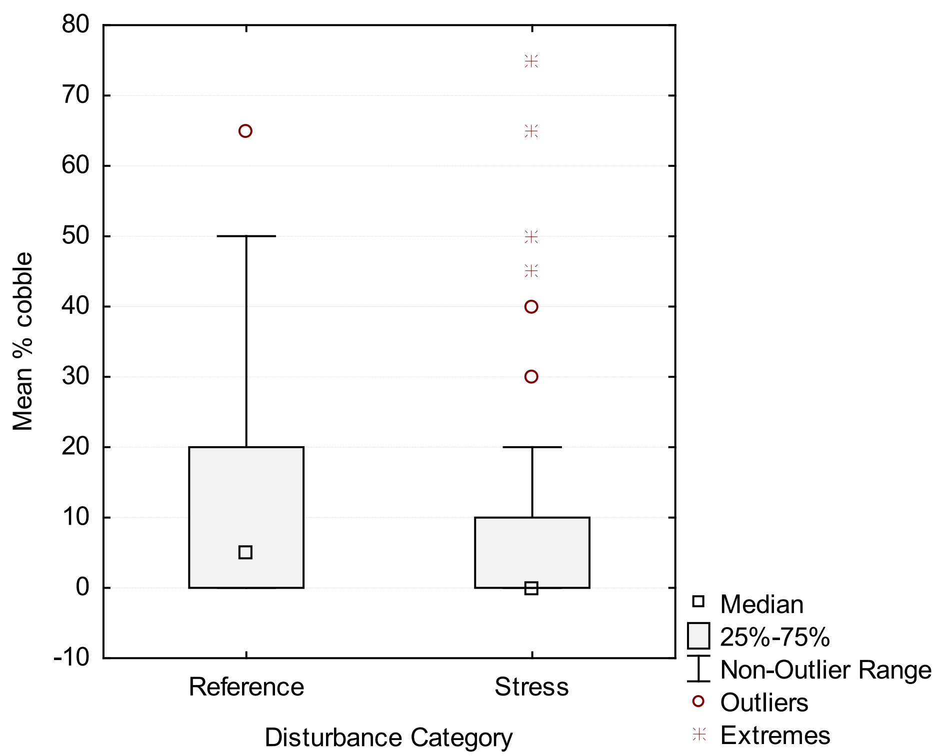


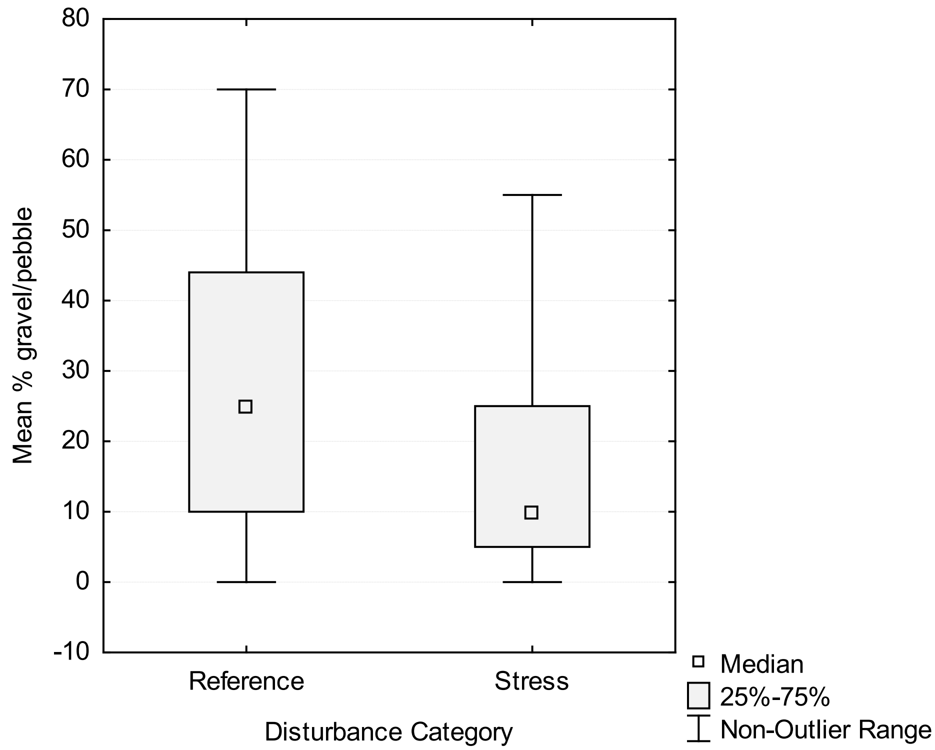
Mean wetted width (m)
22
20
18
16
14
12
10
8
6
4
2
0
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers


2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Maximum pool depth (m)
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range

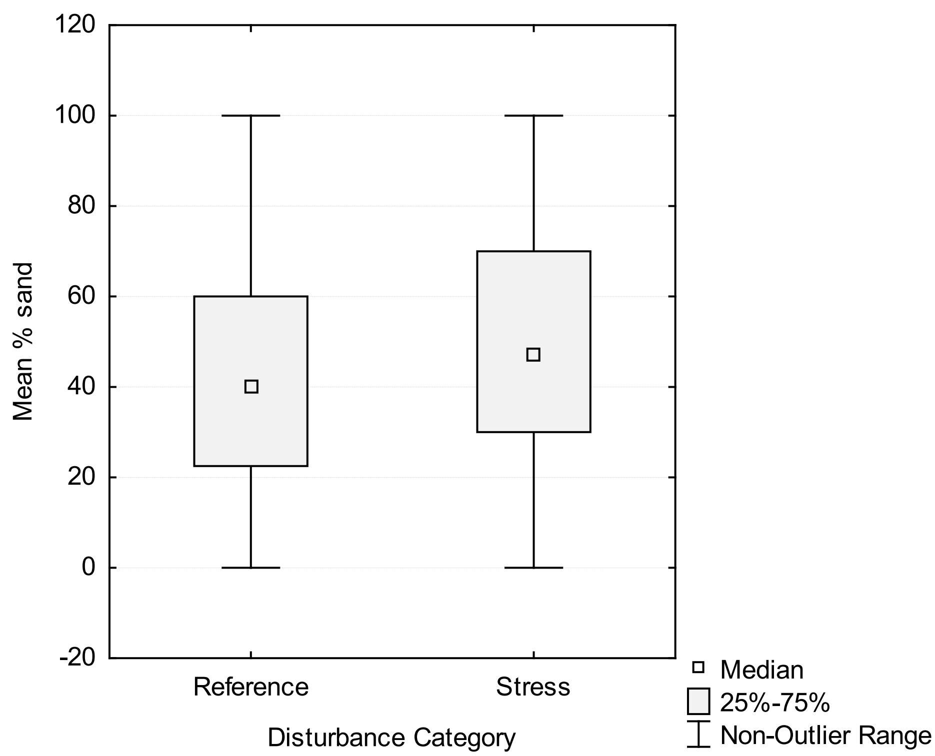

2
1
0
-1
-2
Mean % bedrock
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range


80
70
60
50
40
30
20
10
0
-10
Mean % boulder
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

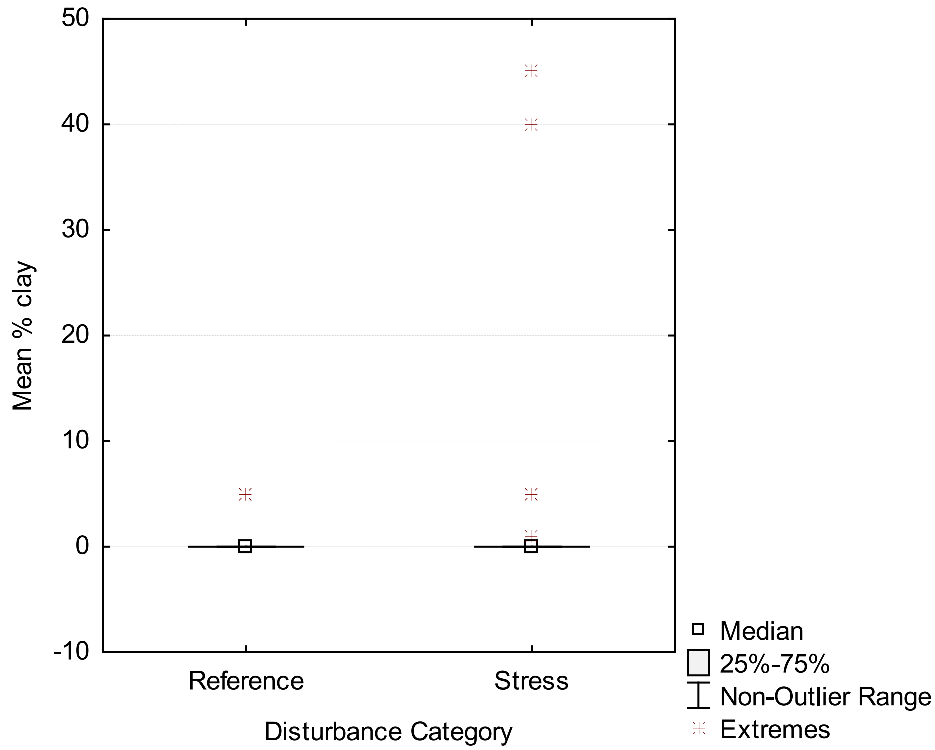

80
70
60
50
40
30
20
10
0
-10
Mean % cobble
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

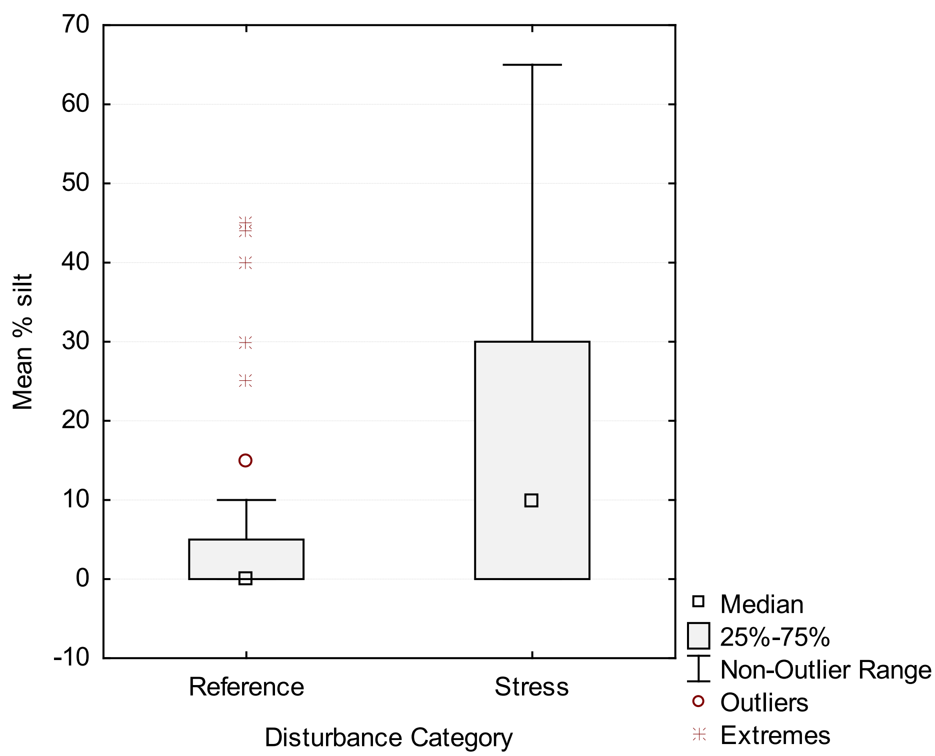


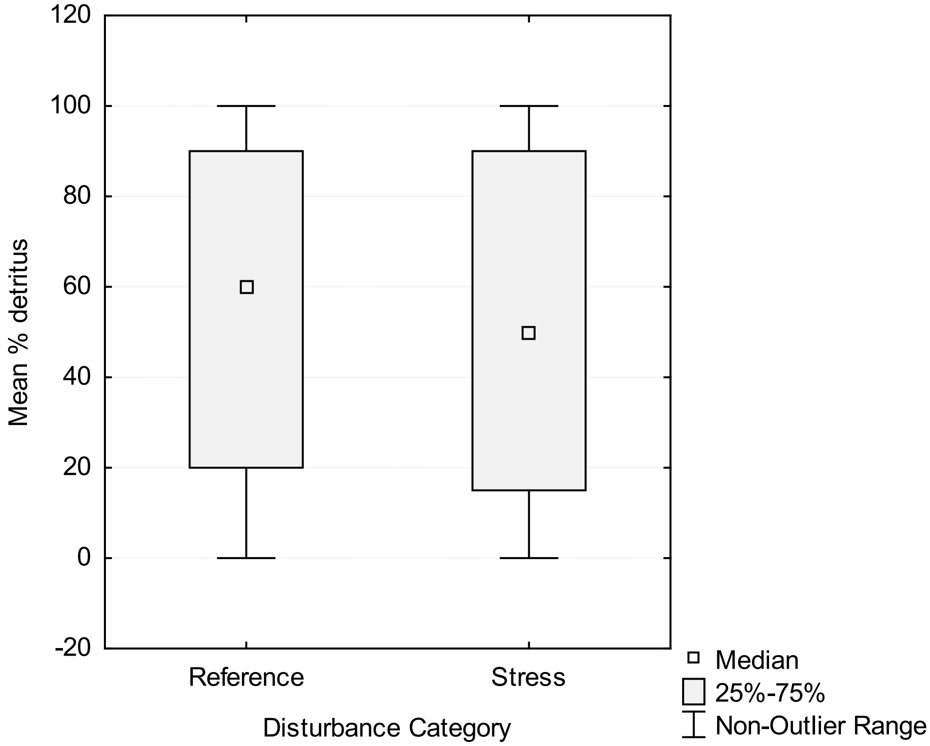
80
70
60
50
40
30
20
10
0
-10
Mean % gravel/pebble
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range

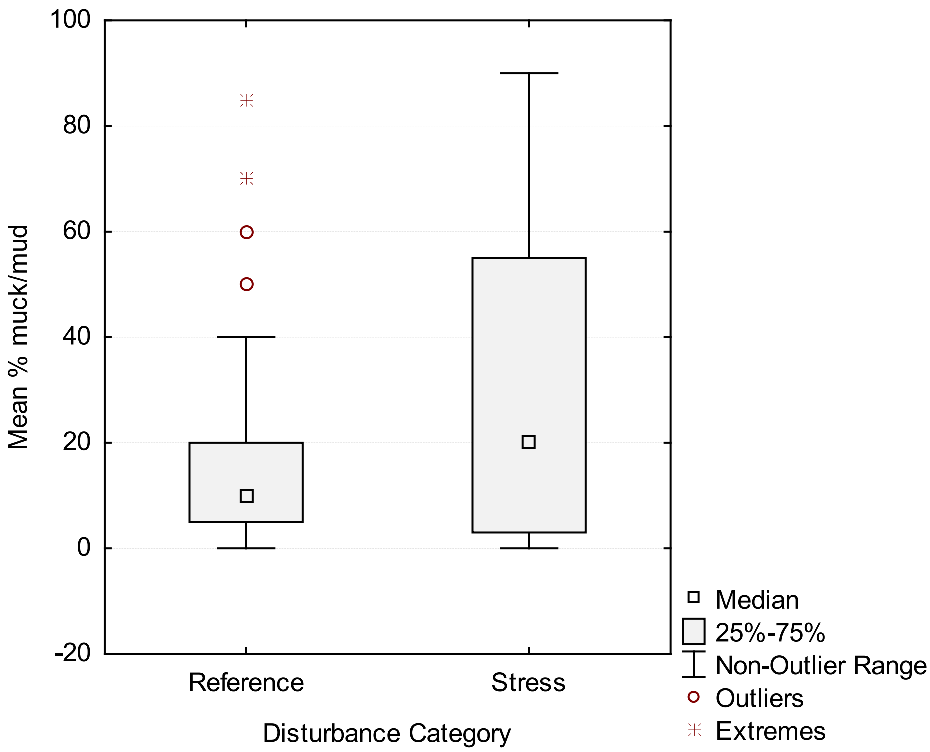


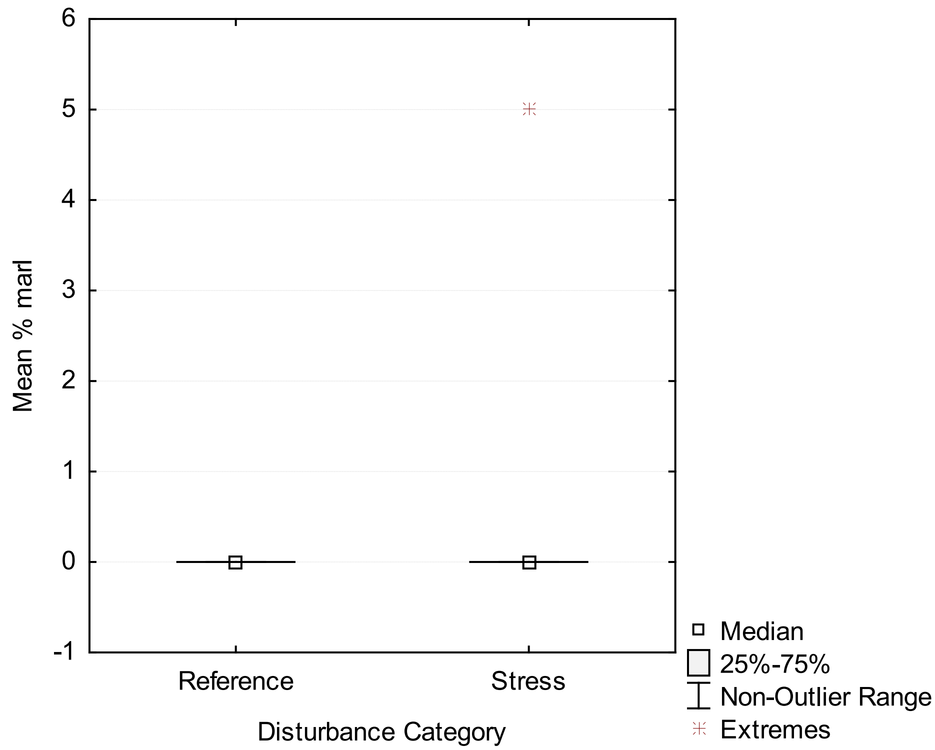
120
100
80
60
40
20
0
-20
Mean % sand
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range


50
40
30
20
10
0
-10
Mean % clay
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Extremes


70
60
50
40
30
20
10
0
-10
Mean % silt
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

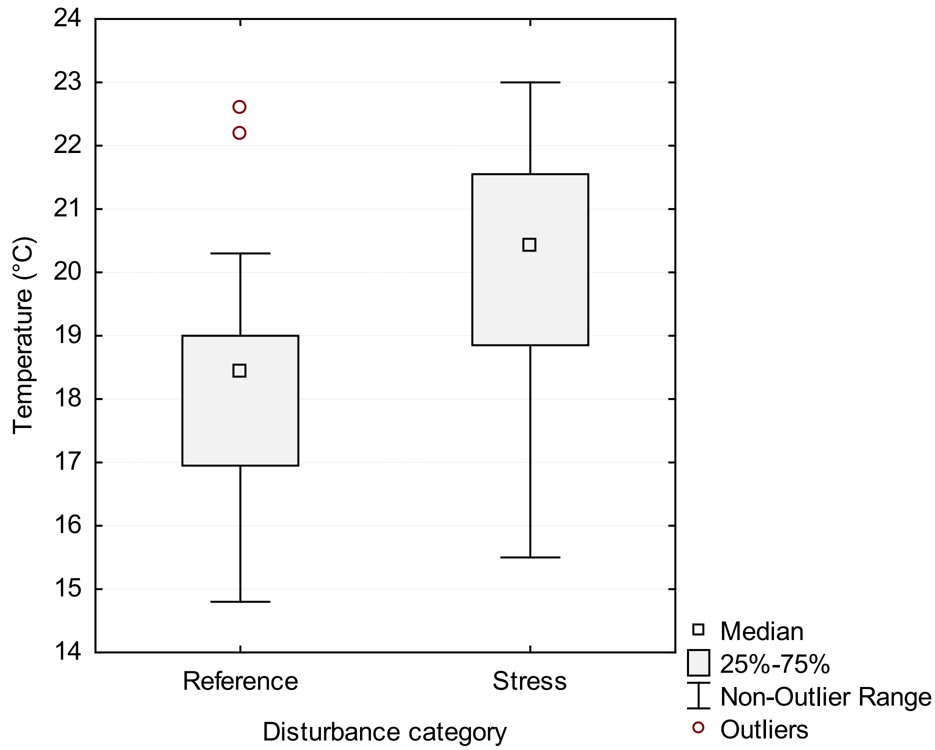

120
100
80
60
40
20
0
-20
Mean % detritus
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range

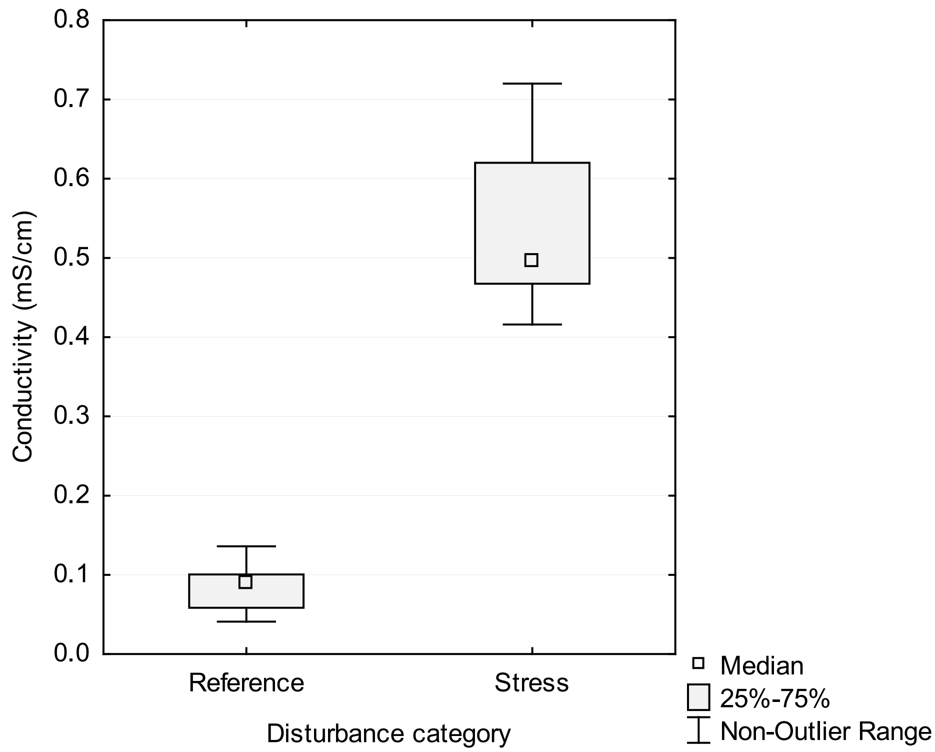

100
80
60
40
20
0
-20
Mean % muck/mud
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

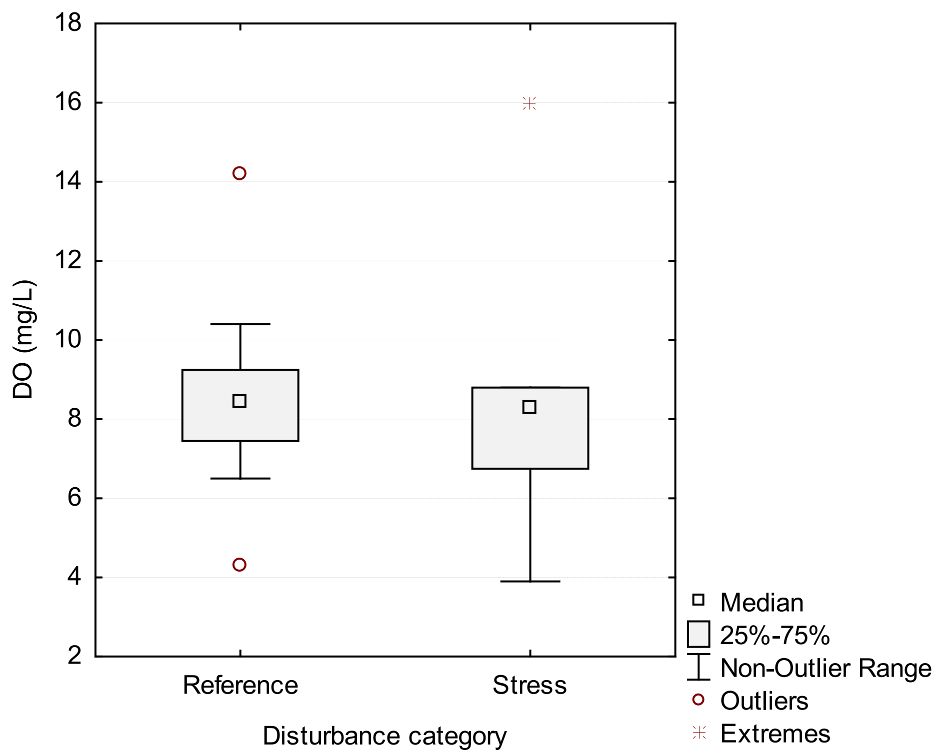

6
5
4
3
2
1
0
-1
Mean % marl
Reference
Stress
Disturbance Category
Median
25%-75%
Non-Outlier Range
Extremes

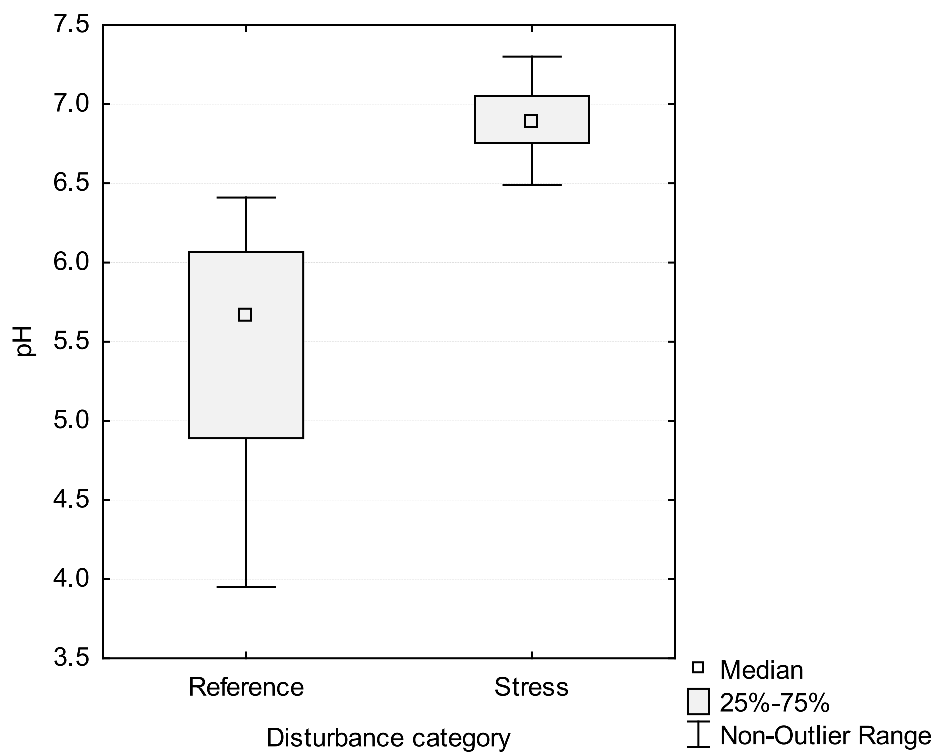
In situ water quality

Tetra Tech/SNEP sites only
20 reference sites, 8 stressed sites


24
23
22
21
20
19
18
17
16
15
14
Temperature (°C)
Reference
Stress
Disturbance category
Median
25%-75%
Non-Outlier Range
Outliers


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Conductivity (mS/cm)
Reference
Stress
Disturbance category
Median
25%-75%
Non-Outlier Range


18
16
14
12
10
8
6
4
2
DO (mg/L)
Reference
Stress
Disturbance category
Median
25%-75%
Non-Outlier Range
Outliers
Extremes


7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
pH
Reference
Stress
Disturbance category
Median
25%-75%
Non-Outlier Range

Appendix E

Site Classification Analysis

and

Additional maps –

- Baseflow
- Mean annual air temperature
- Mean summer stream temperature
- Elevation

Site Classification Analysis

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 by 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.

The classification investigation proceeded through the ordination of taxa and metrics in reference sites so that samples could be organized by similar biological characteristics. To increase the sample size for the classification analyses, two 'borderline reference' sites were included in the reference dataset (n=43; these sites are marked in Attachment C). 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. In each of these ordinations, the biological gradients were mapped in 2 dimensions, with each axis describing orthogonal composite aspects of the community. The axes were then associated with continuous natural variables through correlation. Categorical variables (i.e., level 4 ecoregions) were superimposed on the ordination diagrams to visually discern the separation of categories. Any strong associations of environmental factors with the axes prompted further investigation of the factors as possible classification variables.

Due to the small size of the region and data set, only a few discrete site classes could 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. For adjustment of expectations along a continuous gradient, the optimal metric values were defined relative to the strongly correlated environmental variables. Metric scoring was thereby specific to the natural factor in each site.

Non-metric multidimensional scaling (NMS)

The NMS ordinations were run on presence/absence data from the 43 reference sites, with the dataset limited to 115 common taxa. Taxa that occurred in less than four sites or more than 40 sites were removed to prevent a bias in the sample similarities. The ordination resulted in a 3-dimensional solution with a final stress of 16.2 (< 20 is acceptable). The first two axes explained 70% of the variance in the data, and these axes were explored with correlation analysis and visual inspection.

The first NMS axis (horizontal) was related to forested, western, hard-substrate samples collected in earlier years (in general). These samples had higher numbers of taxa, and notably, higher numbers of sensitive and Ephemeroptera, Plecoptera, and Trichoptera (EPT) taxa. On the left of the diagram were samples from eastern, soft-bottom, wetland influenced streams. These had greater percentages of non-insect individuals (in general) (Figure E-1, Table E-1). The second (vertical) axis was related to wide and large sites with higher base flow and diverse habitats at the top of the diagram. The opposite end of the axis had sites with higher sinuosity as measured in the NHDPlusV2 geospatial layer (Table E-1). The sinuosity measures traced in the GIS exercises did not confirm the relationship suggested by the NHD sinuosity. Biological metrics were not very responsive on the second axis.

On the first axis, longitude and substrate characteristics are the correlated natural variables that might be useful for site classification. Percent forest, percent water and wetland, and percent urban cover are inappropriate for classification because they could be directly influenced by human activities. Temporal variables (collection year and date) might help to explain differences in the current data set but are not reliable for extrapolation to future times. Substrate characteristics include the number of hard-bottom jabs (the habitat type sampled for macroinvertebrates) and percent muck-mud in the reach (estimated substrate areal percent).

Longitude is related to ecoregion and can be used as a continuous variable for classification whereas ecoregion could only be categorical. Eastern sites in the left of the diagram are in the Narragansett-Bristol Lowlands (NBL, L4 ecoregion 59e), Southern New England Coastal Plains and Hills (SNECPAH, L4 ecoregion 59c), and one from the Gulf of Maine Coastal Lowland (L4 ecoregion 59f) (Figure E-1). In Figure 9, it is evident that the level 4 ecoregions have some distinctive taxonomic characteristics, but because there is also considerable overlap in the diagram, ecoregions used as classes might lead to inappropriate biological expectations for some sites. The NBL sites are mostly in the lower left of the diagram, with some samples in the middle-right sections. The SNECPAH sites are mostly in the upper right of the diagram, though several also extend to mix with the NBL sites. The other ecoregions are sparsely represented by reference sites and fall mostly in the lower right of the diagram, between and mixed with the SNECPAH and NBL sites. Longitude and hard-bottom jabs show a relationship with the first axis, but on the right side of the diagram, some sites span the gradients, making identification of class thresholds untenable (Figure E-1).

Patterns related to baseflow were also evident in the second axis of the NMS ordination. The upper right of the diagram had sites with high base flow from the SNECPAH (L4 ecoregion 59c) of RI (Figure E-2). These might be distinctive, but they are mostly in southern RI (see baseflow map under Additional Maps) and might not be appropriate for establishing site classes in MA. The SNECPAH sites that are most distinctive in the upper right are also those with higher base flow (BFI>60) in southern RI. The NBL sites in the lower region are associated with small watersheds (< 5 km^2). Small watersheds might drive some of the taxa composition regardless of ecoregion, as has been acknowledged by the MassDEP biologists. If the small watersheds and high base flow sites are identified in the diagram, the groups are fairly distinctive (Figure E-3). The high BFI sites are biased on the right of the first axis, which is related to taxa richness. However, base flow and stream size are more strongly correlated with the second axis, which is not highly correlated with the biological metrics, so it is uncertain whether these groupings would be appropriate site classes. Plotting a few metrics in these categories shows that the high BFI sites are more distinctive than the small watersheds (for example, the median number of EPT taxa are slightly higher; Figure E-4).

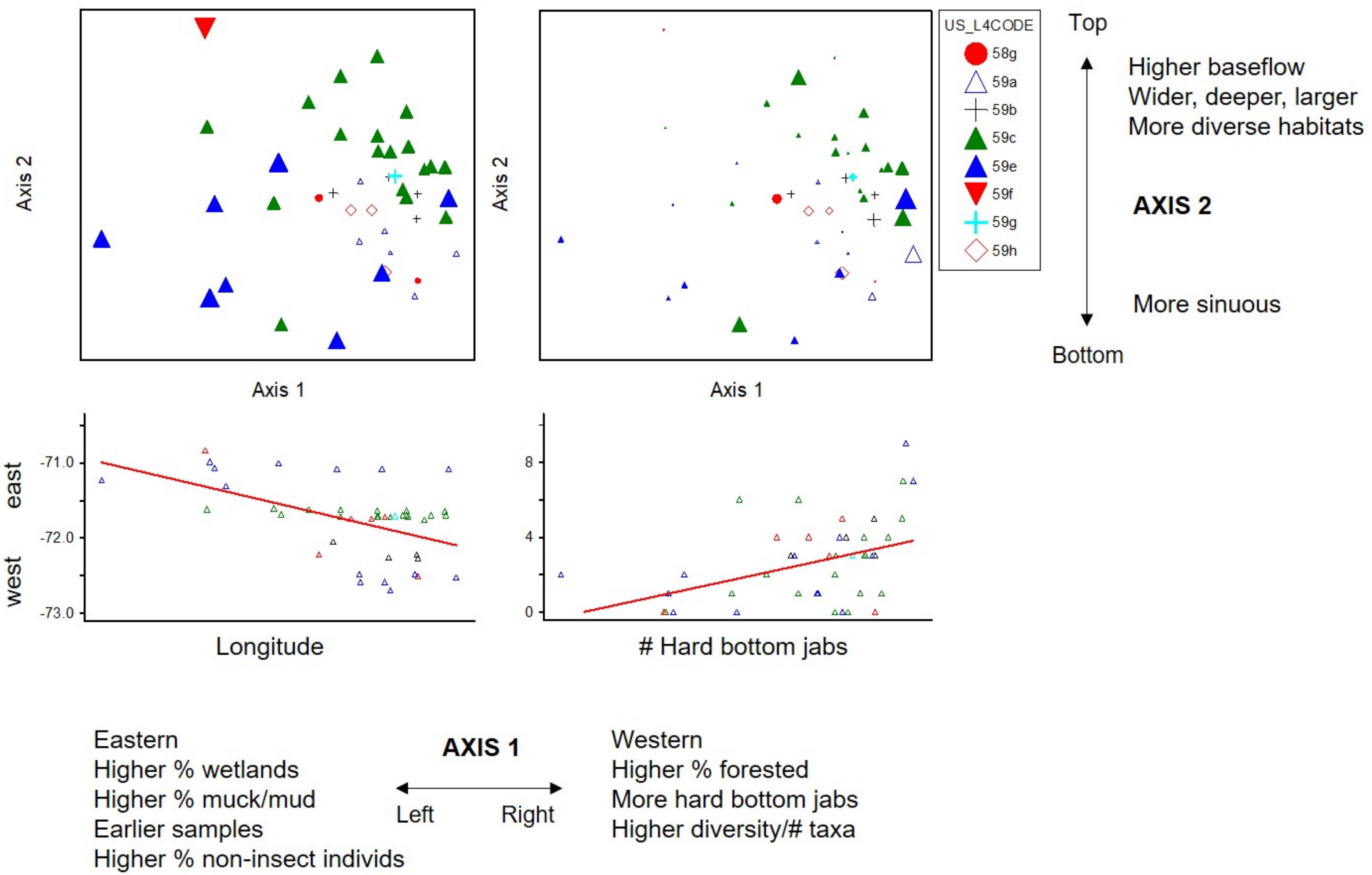


Figure E-1. NMS ordination diagrams based on presence/absence data with samples coded by Level 4 ecoregion. Samples with similar taxonomic composition are plotted in close proximity. In the main plots (upper), marker size shows patterns related to longitude (left) and number of hard bottom jabs (right). The lower plots show how these two metrics relate to axis 1. Axis 1 explains 50% of the variance and axis 2 explains 20%. See Table 1 for Level 4 ecoregion code names.

Table E-1. Environmental and biological metrics that were most strongly correlated with Axes 1 and 2 in the NMS. Negatively correlated metrics (in red) are associated with the left side of Axis 1 and bottom of Axis 2. Positively correlated metrics (in blue) are associated with the right side of Axis 1 and the top of Axis 2.

	Metrics	**r**
Axis 1 (50% variance)	Environmental or temporal	
	% Forest - watershed	0.49
	% Forest - local	0.49
	# Jabs - hard bottom	0.42
	Collection year	-0.41
	Collection date	-0.41
	% Muck/mud	-0.43
	% Wetland/open water - watershed	-0.47
	% Urban - watershed	-0.49
	Longitude	-0.50
	Biological	
	Shannon diversity	0.67
	# EPT taxa	0.63
	# Total taxa	0.43
	% Non-insect individuals	-0.67

	Metrics	**r**
Axis 2 (20% variance)	Environmental	
	Baseflow index	0.48
	Wetted width	0.42
	Drainage area	0.42
	Maximum depth	0.43
	# Habitats sampled	0.44
	# Jabs - vegetated margins/ undercut banks	0.46
	Sinuosity (NHDPlusV2)	-0.42

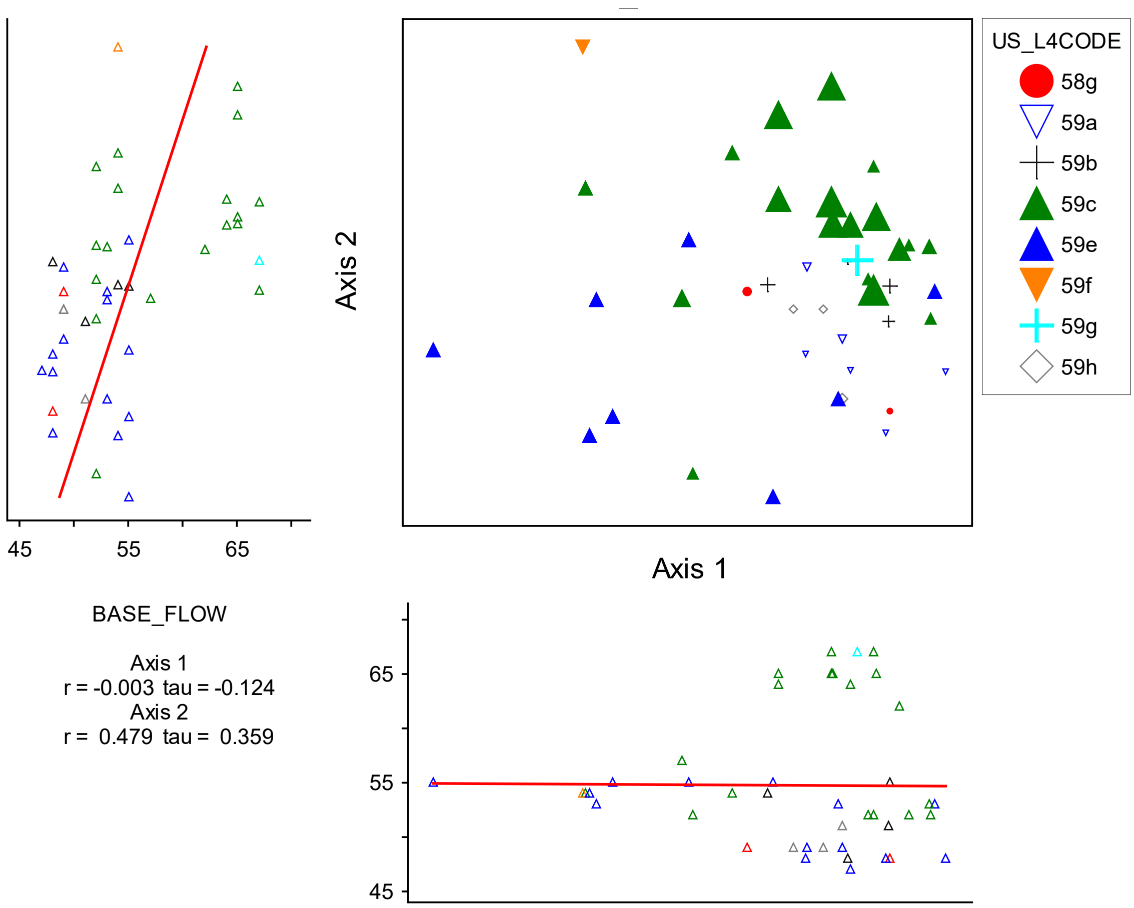


Figure 1. NMS ordination diagrams based on presence/absence data with samples coded by Level 4 ecoregion. Samples with similar taxonomic composition are plotted in close proximity. In the main plot (upper right), marker size indicates degree of baseflow influence. The plots on the bottom and left show how baseflow index values relate to the first and second axes, respectively. Axis 1 explains 50% of the variance and Axis 2 explains 20%. See report Table 1 for Level 4 ecoregion code names.

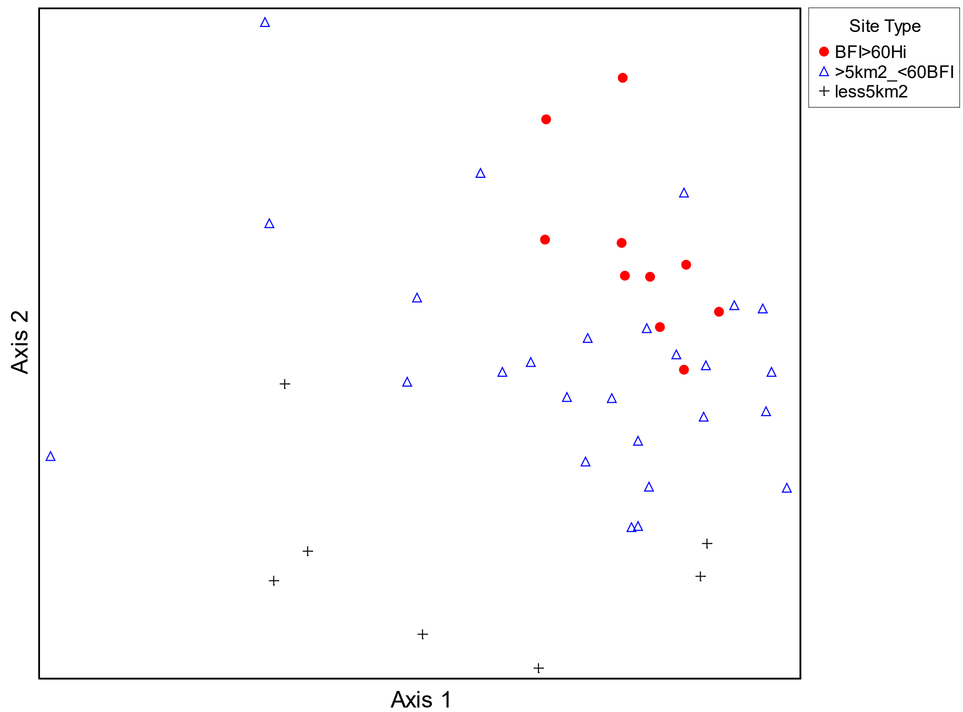


Figure 2. NMS ordination diagram with presence-absence data showing sites with relatively high base flow index (>60) and small watersheds (<5km2).

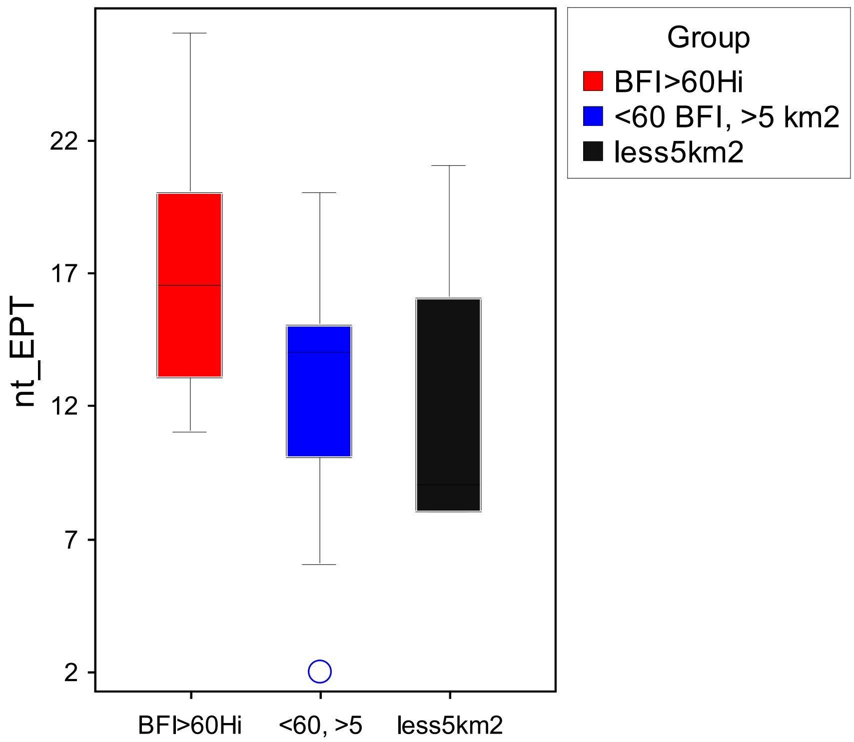


Figure E-4. Distribution of the number of EPT taxa (nt_EPT) in experimental groupings related to baseflow index (BFI) and watershed size.

Principle components analysis (PCA)

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 first five components were considered useful based on results from a broken-stick comparison. The first two components were plotted to illustrate relationships between environmental variables and samples with similar metric values. The first PCA axis (horizontal) was related to forest cover in steeper and larger watersheds on the left of the diagram and eastern watersheds with more wetland cover on the right (Figure E-5, Table E-2). The metrics associated with the steeper, forested watersheds were related to taxa richness overall and in specific groups (clingers, EPT, and intolerant taxa). Tolerant and non-insect individuals were associated with the eastern wetland sites (Table E-2).

On the second (vertical) axis, warmer eastern streams were at the top of the diagram and northern, higher elevation streams were at the bottom (Figure E-5, Table E-2). The northern streams also had more organic material (detritus). The warmer eastern streams had more sensitive insect individuals in contrast to the northern streams with more Diptera and short-lived, multivoltine individuals. The high percentage of midges are apparently in sites with more detritus. Swimmers are also correlated on the second axis, but the relationship appears to be driven by a few high outliers and might not be important in many sites. The relationships between the macroinvertebrate metrics and environmental variables were similar to those observed in the NMS presence/absence analysis. However, stream substrate was less important in the metric PCA than it was in the NMS analysis. Watershed land slope and annual air temperature were more important in the metric PCA than they were in the presence/absence ordination (Tables E-1 and E-2; air temperature map under Additional Maps below).

The total taxa and percent EPT individuals metrics are shown in relation to the Level 4 ecoregions. As with the presence/absence ordination, ecoregions were only somewhat distinctive in the metric PCA. There was considerable overlap between the two heavily sampled ecoregions, the NBL and the SNECPAH. The remaining ecoregions were somewhat distinct and mostly in the lower left of the diagram (Figure E-5). From this pattern and the correlation tables, it appears that the NBL and SNECPAH samples might be different from the other ecoregions in having more EPT and fewer midges. In addition, the NBL samples and sites with more wetland influence might have more tolerant organisms, in general (Table E-2).

The percent of detritus estimated in the sampling reach was correlated with the second PCA axis (Table E-2). This is illustrated in Figure E-6, which shows samples with high detritus coverage in the lower left of the main diagram. This figure also shows the groups that were suggested by the NMS presence/absence analysis. These groups, based on base flow and watershed size, did not appear to separate in the metrics PCA diagram (Figure E-6) and are therefore not dependable for site classification.

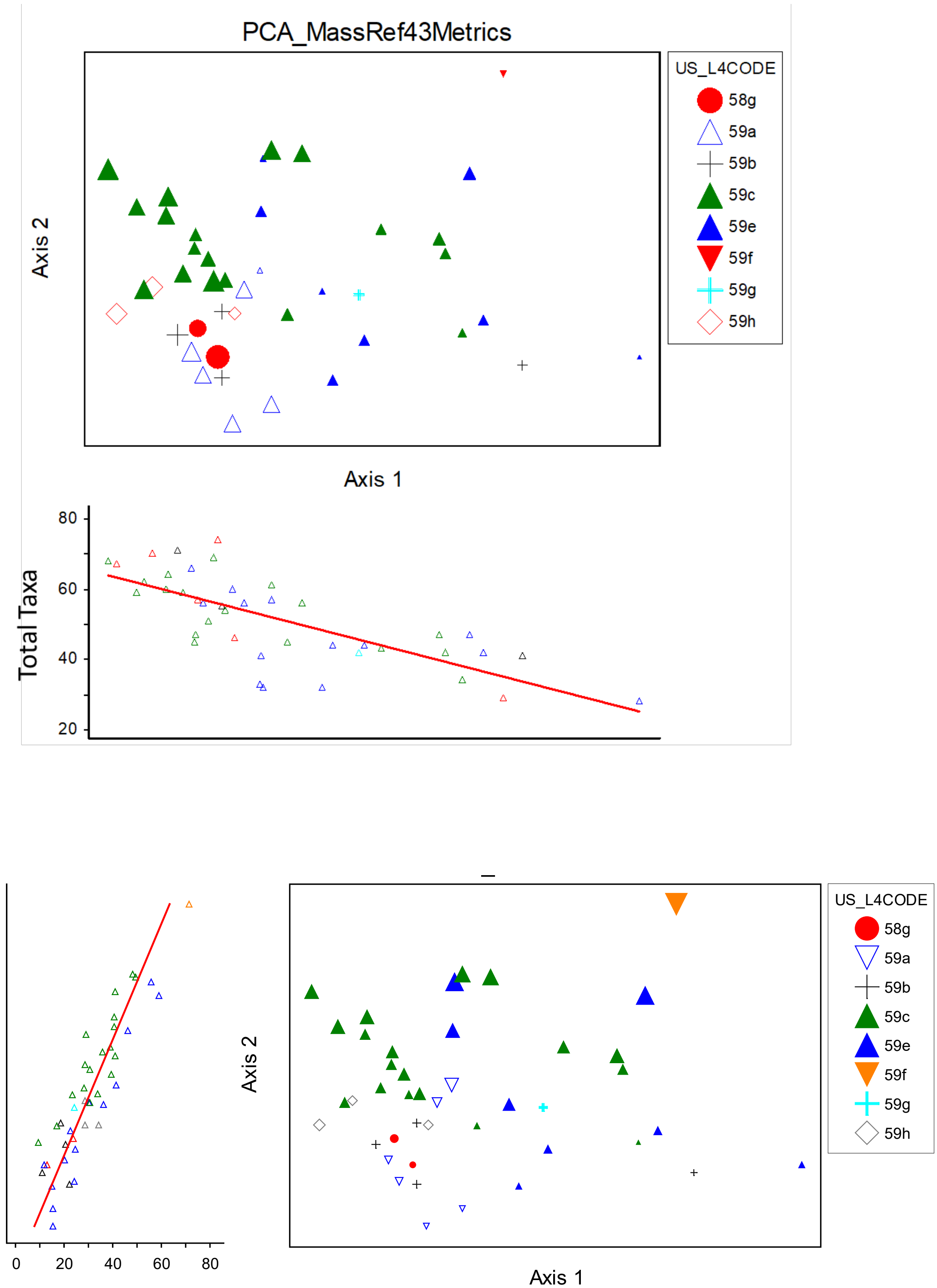


Figure E-5. PCA diagram showing total taxa (top) and percent EPT individuals (bottom), with samples coded by Level 4 ecoregion codes. See Table 1 for Level 4 ecoregion code names.

Table E-2. Environmental and biological metrics that were most strongly correlated with Axes 1 and 2 in the PCA. Negatively correlated metrics (in red) are associated with the left side of Axis 1 and bottom of Axis 2. Positively correlated metrics (in blue) are associated with the right side of Axis 1 and the top of Axis 2.

	Metrics	r		Metrics	r
Axis 1 (34% variance)	Environmental		**Axis 2 (13% variance)**	Environmental	
	% Forest - watershed	-0.61		Mean annual air temperature - local	0.53
	Watershed slope	-0.38		Baseflow index	0.50
	Drainage area	-0.31		Longitude	0.48
	Longitude	0.43		Latitude	-0.44
	% Wetland/open water - watershed	0.57		Elevation - local	-0.44
	Biological			% Detritus	-0.48
	# Clinger taxa	-0.90		Biological	
	Shannon diversity	-0.87		% COET individuals	0.90
	# EPT taxa	-0.83		% Ephemeroptera no Caenidae individuals	0.73
	# Intolerant taxa	-0.77		% Swimmer individuals	0.54
	# Total taxa	-0.73		% Chironomidae individuals	-0.54
	HBI	0.71		% Multivoltine individuals	-0.56
	% Non-insect individuals	0.82		% Diptera taxa	-0.70

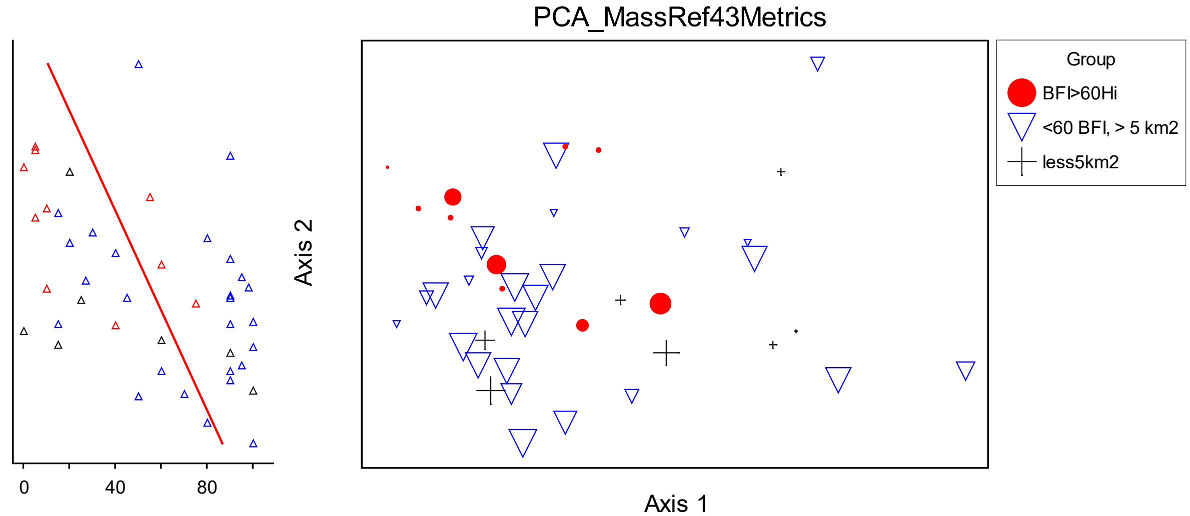


Figure E-6. PCA diagram showing percent detritus estimated in the sampling reach, marked by preliminary site groups of high base flow and small watershed size.

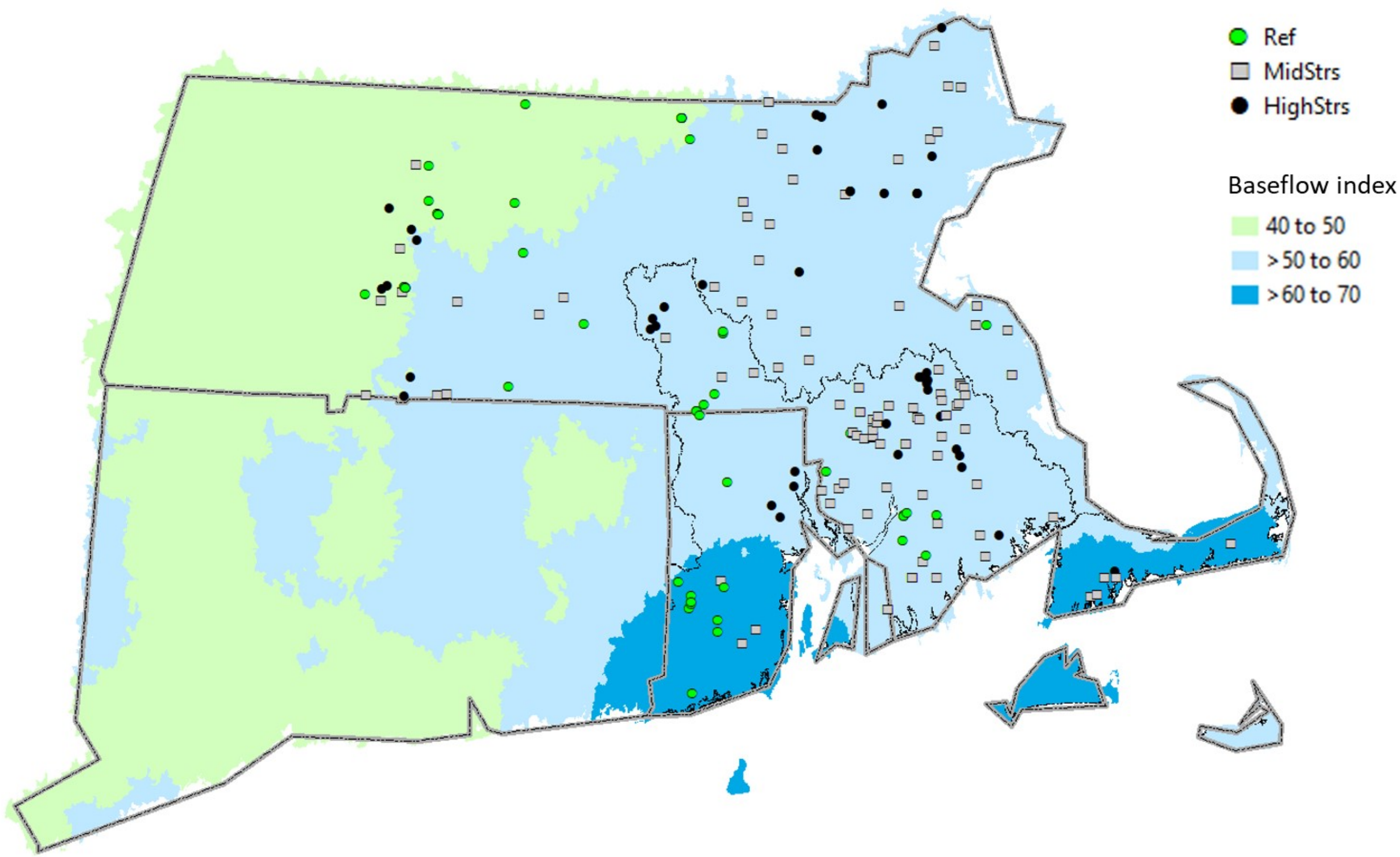
Classification Summary

Classification schemes related to Level 4 ecoregion, baseflow, and drainage area were considered but ruled out based on results from the NMS and PCA analyses. Level 4 ecoregions did not cluster distinctly in the ordinations. Moreover, defining site classes based on Level 4 ecoregions might be untenable because it would result in small sample sizes for index calibration. Patterns related to baseflow (BFI > 60) and watershed size (< 5 km^2) were evident in the NMS but were not strongly correlated with the biological metrics and did not show the same pattern in the PCA, so groupings based on these two metrics were not considered appropriate for site classification.

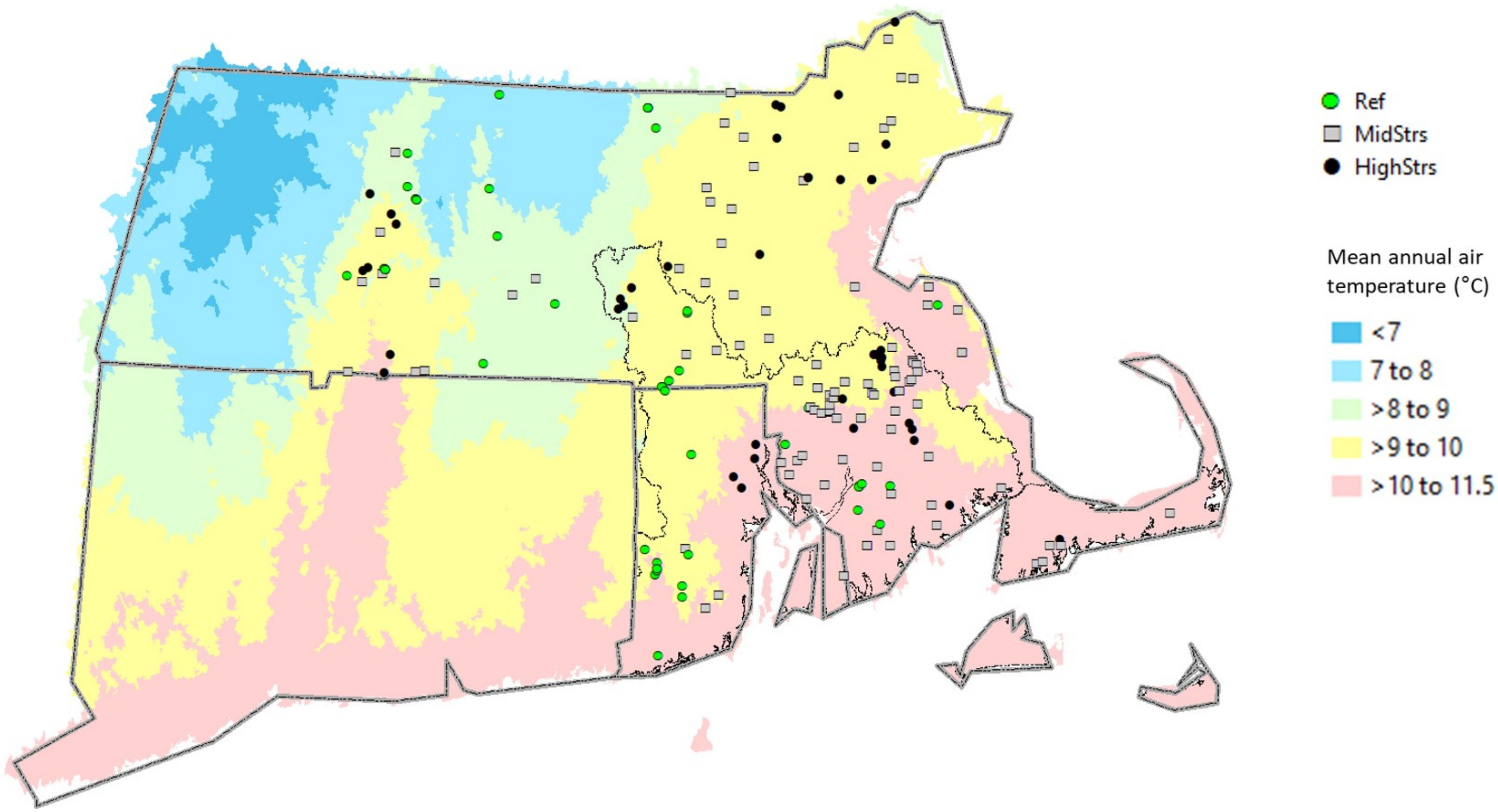
Of the continuous variables, the strongest associations were with percent forest, percent wetlands, and percent detritus. However, these are marginally-natural variables that are inappropriate for classification. Continuous variables that showed potential for classification included: mean annual air temperature (PRISM 1981-2010), latitude, longitude, elevation, watershed slope, 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).

Additional Maps

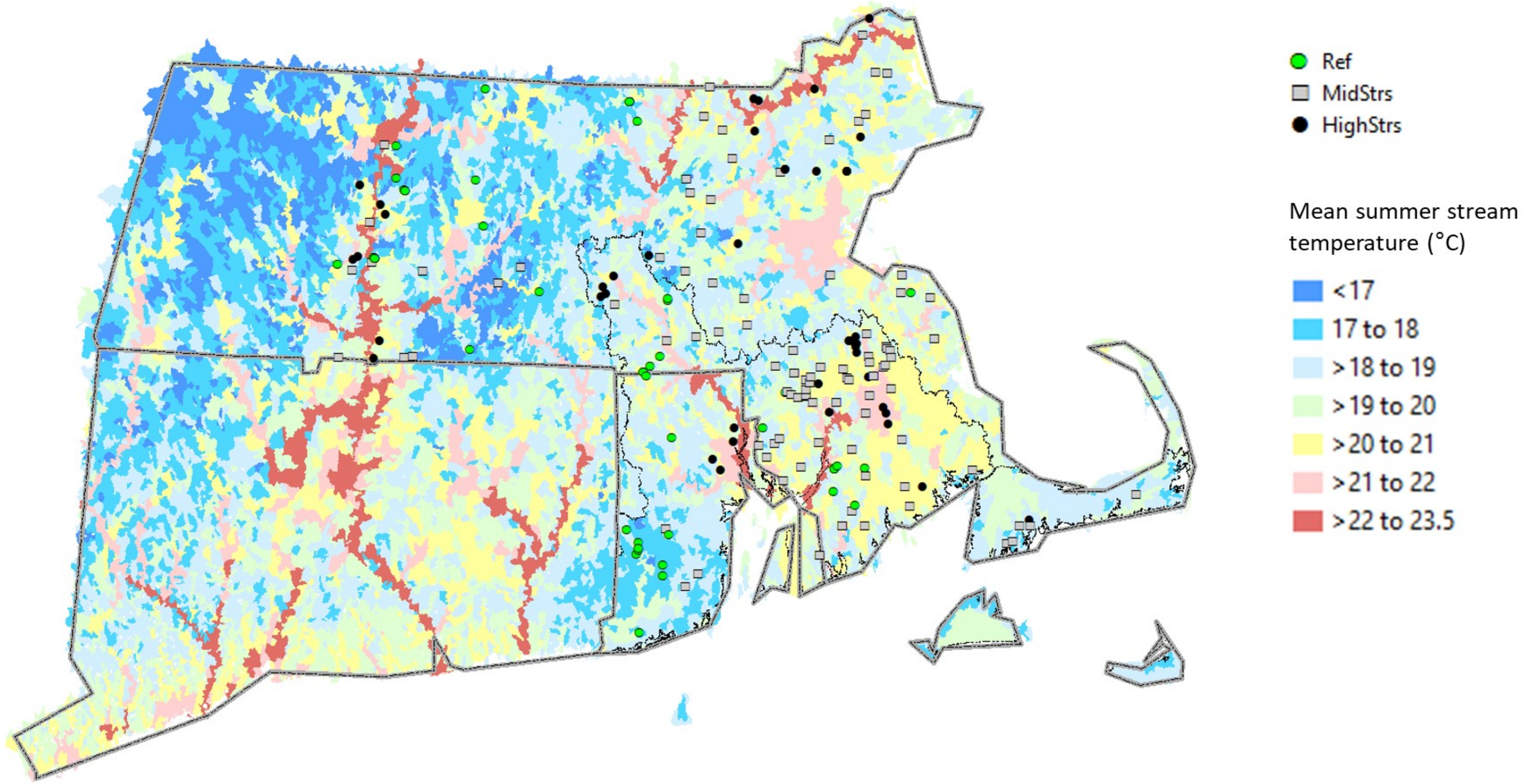
Baseflow index - https://water.usgs.gov/GIS/metadata/usgswrd/XML/bfi48grd.xml



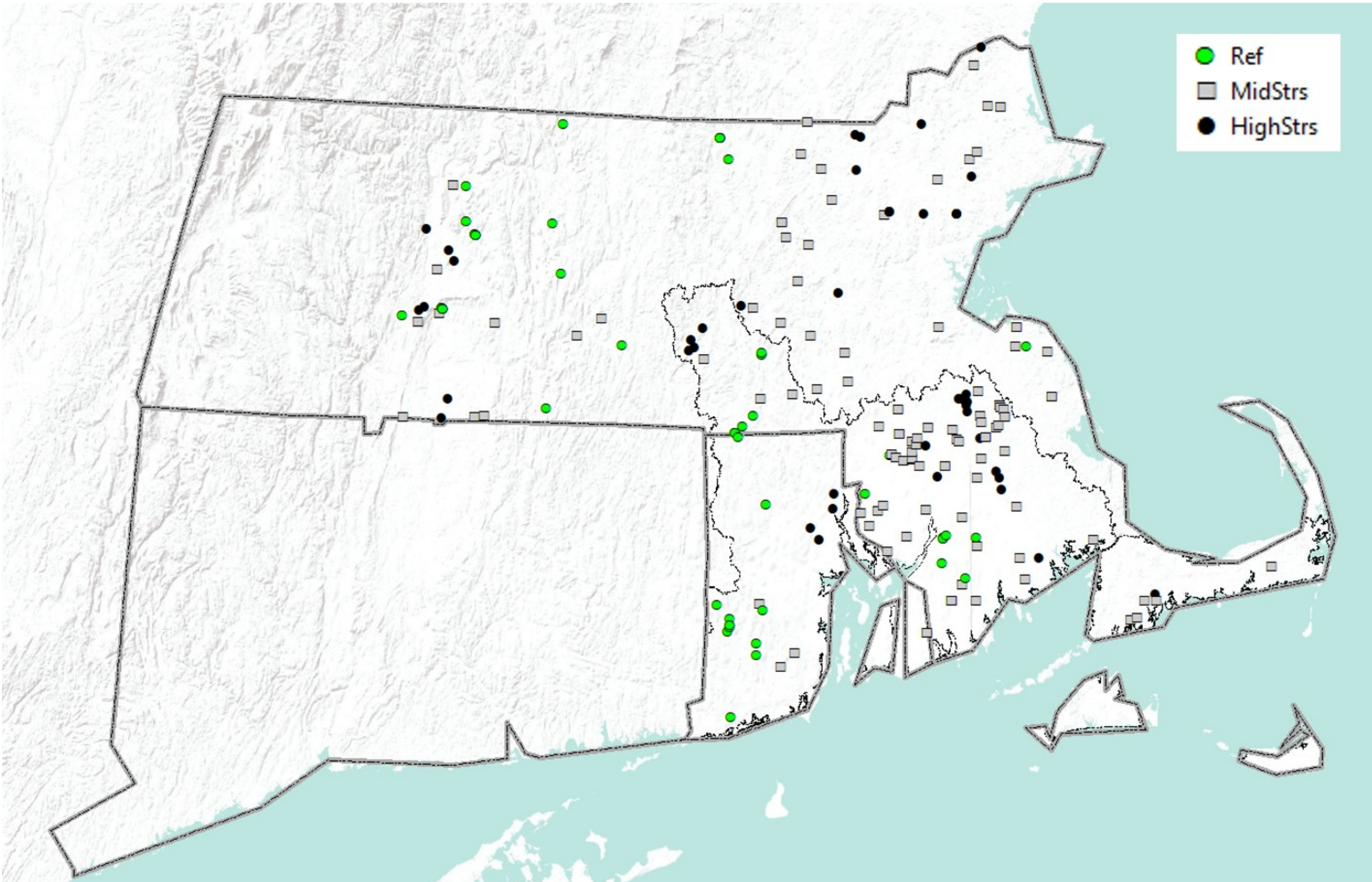
Mean annual air temperature (PRISM 1981-2010) - https://prism.oregonstate.edu/normals/



Mean summer stream temperature (July-August) –
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.



Elevation



Appendix F

MassDEP Low-Gradient IBI Metric Response Mechanisms

Metrics in the MassDEP low gradient IBI were selected for inclusion in the index based on performance statistics (DE and Z-score), response mechanisms, and metric diversity (metrics representative of many metric categories). The recommended IBI consists of metrics representative of relative taxonomic richness, community composition, pollution tolerance, functional feeding groups, and voltinism. The IBI input metrics (Table F1) have comprehensible mechanisms of response to increasing environmental stress, as described below. Interpretable metrics provide easier interpretation of assemblage structure in relation to index scores. Taxa attributes related to the metrics are in Attachment B.

Table F1. Metrics included in the low gradient IBI.

Metric (abbrev)
% Plecoptera, Odonata, Ephemeroptera, and Trichoptera (POET) taxa (pt_POET)
% Predator taxa (pt_ffg_pred)
% Non-insect taxa (pt_NonIns)
% Odonata, Ephemeroptera, and Trichoptera (OET) individuals (pi_OET)
% Tolerant taxa (pt_tv_toler)
% Semivoltine taxa (pt_volt_semi)

% Non-insect taxa (pt_nonIns)

Description: Of all taxa, the percentage of taxa that are non-insects

Taxa richness generally decreases with increasing stress, as the sensitive and specialist taxa emigrate or perish when exposed to intolerable conditions such as pollution, greater sedimentation, or reduced food quality. Non-insects (primarily gastropods, bivalves, crustaceans, and worms) can be tolerant or take advantage of stresses, and therefore, an increase in relative richness indicates the presence of disturbance. Relative richness of non-insects can increase either when non-insect taxa increase or when insect taxa decrease.

Metric Category: Relative Richness

Trend: Expected to increase with stress and increases in the SNEP dataset.

References: Barbour et al. 1999; Yuan and Norton 2003

% POET taxa (Plecoptera, Odonata, Ephemeroptera, and Trichoptera) (pt_POET)

Description: Of all taxa, the percentage of taxa that are in the insect orders Plecoptera (stoneflies), Odonata (dragonflies and damselflies), Ephemeroptera (mayflies), and Trichoptera (caddisflies)

In riffle dominated streams, EPT taxa are generally sensitive to environmental degradation such as reduced dissolved oxygen, unstable substrates, reduced food quality, and contamination due to heavy metals and other pollutants. EPT are also sensitive in low gradient streams and Odonata (dragonflies) can be a fourth sensitive insect order. As environmental conditions become worse, the sensitive and specialist taxa of these insect orders will emigrate or perish.

Metric Category: Relative Richness

Trend: Expected to decrease with stress and decreases in the SNEP dataset.

References: Angradi 1999; Barbour et al. 1999; Yuan and Norton 2003; Hutchens et al. 2009; Steele 2013; Onana et al. 2019; Gomez-Tolosa et al. 2020

% OET individuals (Percent of Odonata, Ephemeroptera, and Trichoptera individuals) (pi_OET)

Description: Of all individuals, the percentage of individuals that are in the insect orders Odonata (dragonflies and damselflies), Ephemeroptera (mayflies), and Trichoptera (caddisflies)

The stressor mechanisms described for % POET taxa also affect the relative abundance of sensitive insect individuals in a stream. Plecoptera (stoneflies) are more meaningful as a presence/absence signal than they are as a relative abundance signal because they are usually not abundant in low gradient streams. Therefore, this metric does not include stoneflies. The sensitive and specialist individuals of the dragonfly, mayfly, and caddisfly insect orders emigrate or perish with increasing stress.

Metric Category: Composition

Trend: Expected to decrease with stress and decreases in the Michigan dataset.

References: Angradi 1999; Barbour et al. 1999; Yuan and Norton 2003; Hutchens et al. 2009; Steele 2013; Onana et al. 2019; Gomez-Tolosa et al. 2020

% Predator taxa (Percent taxa of the predator (PR) Functional Feeding Group) (pt_ffg_pred)

Description: Of all taxa, the percentage of taxa that consume other organisms using different strategies to capture them

Predators employ a diversity of strategies for capturing prey, including modified mouth parts and behavior. Some species of invertebrates are predators in both the larval and adult stages of their life.

Metric Category: Functional Feeding Groups

Trend: Expected to decrease with stress and decreases in the SNEP dataset.

References: Kerans and Karr 1994; Merritt et al. 2008; Hutchens et al. 2009; Xu et al. 2014; Lan Fu et al. 2016;

% Tolerant taxa (Percent tolerant taxa with tolerance value ≥ 7) (pt_tv_toler)

Description: Of all taxa, the percentage of taxa that are relatively tolerant to stressors

Taxa respond differently to environmental stressors, therefore, can be arranged on a continuum from intolerant to tolerant. Intolerant taxa will emigrate or perish as environmental conditions worsen. Conversely, tolerant taxa may not respond negatively to environmental conditions and may actually increase as niches open from extirpated intolerant taxa.

Metric Category: Tolerance

Trend: Expected to increase with stress and increases in the SNEP dataset.

References: Hilsenhoff 1987; Yuan 2006; Megan et al. 2007; USGS 2013

% Semivoltine taxa (Percent Semivoltine taxa) (pt_volt_semi)

Description: Of all taxa, the percentage of taxa that require more than one year in a reproduction cycle

Taxa respond differently to environmental stressors, therefore, can be arranged on a continuum from intolerant to tolerant. Intolerant taxa will emigrate or perish as environmental conditions worsen. Conversely, tolerant taxa may not respond negatively to environmental conditions and may actually increase as niches open from extirpated intolerant taxa.

Metric Category: Voltinism

Trend: Expected to increase with stress and increases in the SNEP dataset.

References: Barbour et al. 1994; Dole´dec et al. 2006; Statzner and Beˆche 2010

References:

Angradi, T.R., 1999. Fine sediment and macroinvertebrate assemblages in Appalachian streams: a field experiment with biomonitoring applications. *Journal of the North American Benthological Society*, *18*(1):49-66.

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.

Barbour, M.T., M.L. Bowman, and J.S. White. 1994. Evaluation of the biological condition of streams in the Middle Rockies - Central ecoregion. Prepared for Wyoming Department of Environmental Quality.

Dole´dec S., N. Phillips, M. Scarsbrook, R.H. Riley, C.R. Townsend. 2006. Comparison of structural and functional approaches to determining landuse effects on grassland stream invertebrate communities. *Journal of the North American Benthological Society* 25: 44–60.

Gómez-Tolosa, M., Rivera-Velázquez, G., Rioja-Paradela, T.M., Mendoza-Cuenca, L.F., Tejeda-Cruz, C. and López, S., 2020. The use of Odonata species for environmental assessment: a meta-analysis for the Neotropical region. *Environmental Science and Pollution Research*, pp.1-16.

Hilsenhoff, W. L. 1987. An improved biotic index of organic stream pollution. *The Great Lakes Entomologist*, 20(1):7.

Hutchens Jr, J.J., J.A. Schuldt, C. Richards, L.B. Johnson, G.E. Host, and D.H. Breneman. 2009. Multi-scale mechanistic indicators of Midwestern USA stream macroinvertebrates. *Ecological Indicators*, 9(6):1138-1150.

Kerans, B.L. and J.R. Karr. 1994. A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. *Ecological Applications* 4:768 - 785

Lan Fu, Yuan Jiang, Jiao Ding, Qi Liu, Qiu-Zhi Peng, and Mu-Yi Kang. 2016. Impacts of land use and environmental factors on macroinvertebrate functional feeding groups in the Dongjiang River basin, southeast China, *Journal of Freshwater Ecology*, 31(1):21-35. DOI: 10.1080/02705060.2015.1017847

Megan, M.H., M.S. Nash, C. Neale, A.M. Pitchford. 2007. Biological integrity in mid-atlantic coastal plains headwater streams. *Environmental Monitoring and Assessment* 124:141–156.

Merritt, R. W., K. W. Cummins, and M. B. Berg. 2008. An introduction to the aquatic insects of North America. Fourth edition. Kendall, Dubuque, Iowa, USA.

Onana, F.M., S.H.Z. Togouet, A.A. Tamsa, N.L.N. Tchatcho, S. Tchakonte, E. Koji, A.W.Y. Yemeli, and A.N.S.M. Makong. 2019. Comparing freshwater benthic macroinvertebrate communities in forest and urban streams of the coastal ecological region of Cameroon. *Open Journal of Ecology*, 9(12):521.

Statzner B. and L. Beˆche. 2010. Can biological invertebrate traits resolve effects of multiple stressors on running water ecosystems? *Freshwater Biology* 55: 80–119.

Steele, A., 2013. *An analysis of the North Carolina Wetland Assessment Method (NCWAM) using quantitative metrics*. North Carolina State University. College of Natural Resources.

USGS (U.S. Geological Survey), 2013. 2012 National Anthropogenic Barrier Dataset Available. From: https://www.sciencebase.gov/catalog/item/56a7f9dce4b0b28f1184dabd (Accessed April 2019)

Xu, M., Z. Wang, X. Duan, and B. Pan. 2014. Effects of pollution on macroinvertebrates and water quality bio-assessment. *Hydrobiologia*, *729*(1), 247-259.

Yuan, L.L., and S.B. Norton. 2003. Comparing responses of macroinvertebrate metrics to increasing stress. *Journal of North American Benthological Society*, 22(2):308-322.

Yuan, L.L., 2006. Estimation and application of macroinvertebrate tolerance values. US EPA, ORD, National Center for Environmental Assessment.