Development of a Macroinvertebrate Multi-Metric Index for Wadeable Streams in Kansas

DRAFT REPORT

***Prepared for*:**

Kansas Department of Health and Environment
Tony Stahl, Work Assignment Manager

***Prepared by*:**

Tetra Tech

73 Main Street, Room 38, Montpelier, VT 05602

August 10, 2023

Executive Summary

The Kansas Department of Health and Environment (KDHE) is responsible for sampling and assessing Kansas'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. For both 305b and 303d screening assessments, aquatic life support is demonstrated through both biology and water chemistry data. Combining biological and chemical sampling provides a more complete picture of ecological status than either type of information can provide alone. Whereas water chemistry measurements provide information about chemical conditions at the moment they are collected, benthic macroinvertebrates (BMI) have life spans ranging from weeks to years. Therefore, the assemblage and structure of the BMI community can provide a time-integrated measure of environmental conditions ranging from months to years.

KDHE has two programs that assess the biological integrity in Kansas streams - the Stream Probabilistic (SP) program and the Stream Biological (SB) Monitoring Program. Together, the programs have collected over 3000 BMI samples from over 700 sites. BMI data from the SP program are used to estimate aquatic life support in streams statewide, for 305b "screening level" reporting. Data from the SB program data are used to document long-term trends in surface water quality and for TMDL follow-up monitoring and special projects.

For this project, we developed a Macroinvertebrate Multi-Metric Index (MMI) for wadeable streams in Kansas that can be calculated with a free R-based tool (referred to as a Shiny app) that can be accessed via this weblink: [placeholder]. Users do not need R on their computers to run the app, nor do they need to have any familiarity with R. They just need an internet connection. For those who prefer to work with R, the R code can be downloaded from this link: [placeholder].

The MMI is a numeric representation of biological conditions based on the combined signals of several different assemblage measurements. It was 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 MMI indicates the biological condition of the stream on a relative scale. The MMI scores in the reference sites are reasonable expectations for any stream in the region, and MMI scores that do not resemble the reference scores indicate that there might be stressors influencing the biological condition.

To account for natural biological variability, we used random forest (RF) models to predict metric expectations for each site based on multiple natural environmental variables at least-disturbed reference sites (sensu Carlisle et al. 2022). In the random forest analysis, the classification is continuous, not in discrete site classes, and observed metric values are compared to the model-predicted values to evaluate whether each metric is as predicted or underperforming. Steps for MMI development included data compilation and preparation, development of a disturbance index to identify reference and stressed sites, running random forest models for each metric, evaluating the performance of different combinations of metrics and selecting the MMI. The top candidate MMIs had the highest discrimination efficiencies (DEs) (lowest error when discriminating between reference and stressed sites) and metrics that were familiar to the workgroup members, ecologically meaningful, and diverse in response mechanisms.

The input metrics for the final MMI are listed in Table ES-1. They represented four different metric categories (habit, tolerance, composition, and voltinism). Each KS MMI input metric has its own random forest model and set of predictor variables, with 12-16 predictor variables per metric. The

KS MMI had a DE of 78% (Figure ES-1). As an alternate measure of performance, we used scatterplots and Spearman Rank correlation analyses to assess the strength and direction of MMI response to disturbance variables. We also explored relationships between MMI scores and program (SB vs. SP), time (Julian date, month, year), stream size, habitat, Level 3 ecoregion, Hydrologic Unit Code (HUC) and specific conductance.

Table ES-1. The KS MMI includes five input metrics from four different metric categories. Each metric was random forest (RF) adjusted based on natural predictor variables.

#	Metric	Response to Stress	Metric Category	Top 3 RF Predictors
1	Number of EPT taxa	Decrease	Composition	Precipitation, Flow (CFS), Elevation
2	Percent sensitive taxa (BCG attribute III+IV better)	Decrease	Tolerance	Longitude, Level 3 Ecoregion, Precipitation
3	Hilsenhoff Biotic Index (HBI)	Increase	Tolerance	Lithologic Potassium, Precipitation, Longitude
4	Number of climber + clinger taxa	Decrease	Habit	Elevation, Precipitation, Lithologic Magnesium
5	Number of semivoltine taxa	Decrease	Life cycle	Watershed Area, Elevation, Flow (CFS)

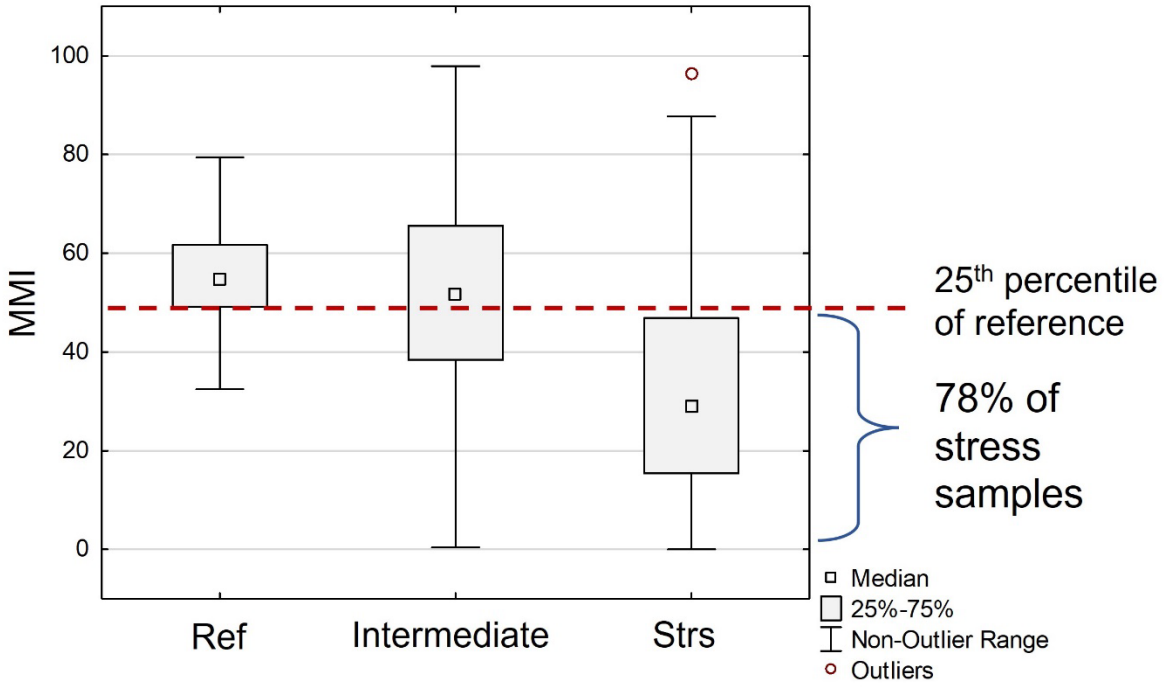


Figure ES-1. Distribution of MMI scores in the reference, intermediate and stress samples in the MMI calibration dataset. The overall DE of the MMI is 78%, which means 78% percent of stressed sites scored below the 25th percentile of reference (MMI=49).

In addition, we performed precision analyses to quantify the variability of the KS MMI and its input metrics based on sampling and temporal effects. Precision was best in the same site/same day replicates and decreased with the addition of temporal variability (seasonal and interannual) and programmatic differences (SB vs. SP). We also performed exploratory analyses to inform potential numeric thresholds for multiple biological condition categories (e.g., Very Good, Satisfactory,

Moderately Degraded, and Severely Degraded). We used three approaches: distribution statistics, balancing Type I and Type II errors, and crosswalking MMI scores with preliminary results from the Great Plains Biological Condition Gradient (BCG) model (Stamp et al., in progress). Results suggest potential MMI condition breakpoints of 30/50/70. These thresholds are preliminary and are subject to further review, refinement, and approval by KDHE before they are applicable in biological assessment programs.

Acknowledgements

The MMI development process was supported by the Kansas Department of Health and Environment (KDHE) through a contract with Tetra Tech, Inc. The contract was facilitated by Tony Stahl and William Carr of KDHE. The index development team, which consisted primarily of Tony Stahl, Elizabeth Smith, Clint Goodrich, Katlynn Decker and Jack Lapin, participated in regularly-scheduled calls and provided feedback throughout the process. Tony Stahl, Elizabeth Smith, and Clint Goodrich reviewed site disturbance assignments and index compilations. Elizabeth Smith and Clint Goodrich also reviewed taxa attribute assignments and provided guidance on resolving taxonomic nomenclature changes. Project authors and analysts included Ben Jessup, Diane Allen, Jen Stamp, Ben Block and Erik Leppo of Tetra Tech.

We are very grateful for the hard work and enthusiasm of all the project participants.

An appropriate citation for this report is as follows:

Tetra Tech. 2023. Development of a Macroinvertebrate Multi-Metric Index for Wadeable Streams in Kansas. Prepared for the Kansas Department of Health and Environment, Topeka, KS. Prepared by Tetra Tech, Montpelier, VT.

Cover photo: Smoky Hill River (SB268), provided by KDHE.

Acronyms

Acronym	Description
ALU	Aquatic Life Use
BCG	Biological Condition Gradient
BMI	Benthic Macroinvertebrates
BMP	Best Management Practice
BPJ	Best Professional Judgement
CAFO	Confined Animal Feeding Operation
CART	Classification and Regression Tree
CFS	Cubic feet per second; units for median flow statistic
CI90	90% Confidence Interval
CV	Coefficient of Variation
CWA	Clean Water Act
DE	Discrimination Efficiency
EPT	Ephemeroptera, Plecoptera, & Trichoptera
FFG	Functional Feeding Group
GIS	Geographic Information System
GP	Glide-Pool
HBI	Hilsenhoff Biotic Index
ITIS	Integrated Taxonomic Information System
IWI	Index of Watershed Integrity
KDHE	Kansas Department of Health and Environment
MMI	Multi-Metric Index
MSE	Mean-squared-error
MSST	Mean Summer Stream Temperature
NHD	National Hydrography Dataset
NLCD	National Land Cover Database
NPDES	National Pollutant Discharge Elimination System
NRSA	National Rivers and Streams Assessment
PCA	Principle Components Analysis
QAPP	Quality Assurance Project Plan
RHA	Rapid Habitat Assessment
RMSE	Root Mean Squared Error
RR	Riffle-Run
TRI	Toxics Release Inventory
NPDES	National Pollutant Discharge Elimination System
NPL	Superfund National Priority List
SB	Stream Biological Monitoring Program
SMASH	Stream Macroinvertebrate Assessment Sampled Habitat Index
SP	Stream Probabilistic Program
SWQS	Surface Water Quality Standards
TRI	Toxic Release Inventory
TMDL	Total Maximum Daily Load
USEPA	United States Environmental Protection Agency

WQS	Water Quality Standards

Table of Contents

List of Tables

List of Figures

1 Introduction

The Kansas Department of Health and Environment (KDHE) is responsible for sampling and assessing Kansas'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. For both 305b and 303d screening assessments, Aquatic life support is demonstrated through both biology and water chemistry data. Combining biological and chemical sampling provides a more complete picture of ecological status than either type of information can provide alone. Whereas water chemistry measurements provide information about chemical conditions at the moment they are collected, benthic macroinvertebrates (BMI) have life spans ranging from weeks to years. Therefore, the assemblage and structure of the BMI community can provide a time-integrated measure of environmental conditions ranging from months to decades.

There are two programs that assess biological integrity in Kansas streams - the Stream Probabilistic (SP) program and the Stream Biological (SB) Monitoring Program. Both programs collect BMI data as one of the CWA measures of biological integrity. The SP program began in 2006 while the SB program dates back to 1972. Together, the programs have collected over 3000 BMI samples from over 700 sites. BMI data from the SP program are used to estimate aquatic life support in streams statewide, for 305b "screening level" reporting. Data from the SB program data are used to document long-term trends in surface water quality and for TMDL follow-up monitoring and special projects.

In the past, the SB and SP programs have used a Biotic Index[1] to help assess full, partial and non-support of aquatic life. For this project, we developed a statewide multi-metric index (MMI) that builds upon their existing index. A MMI is a numeric representation of biological conditions based on the combined signals of several different assemblage measurements (Karr 1986). The raw measurements are recalculated or standardized as biological metrics, or numerical expressions of attributes of the biological assemblage (based on sample data) that respond to human disturbance in a predictable fashion. The process by which the metrics are selected and combined in a MMI follows established and innovative analytical methods of the reference condition approach (Hughes et al. 1986, Bailey et al. 2004). In this approach, biological conditions from relatively undisturbed sites and that account for natural variability, are set as a standard (or reference) to which other samples are compared (Stoddard et al. 2006). Using metrics to establish the numeric index scale, MMI values that are comparable to those found in reference sites meet the expectations for a well-balanced aquatic community. MMI values that are unlike the reference values indicate departure from the expected biological conditions and are assumed the result of anthropogenic disturbance.

In this report, we describe the development of a statewide MMI for Kansas wadeable stream BMI data. We utilized an approach similar to Carlisle et al. (2022) in which metric expectations for each site were predicted from multiple natural environmental variables at reference sites using a random forest model. In the random forest analysis, the classification is continuous, not in discrete site classes, and observed metric values are compared to the model-predicted values to evaluate whether each metric is as predicted or underperforming. Steps for MMI development included data compilation and preparation, development of a disturbance index to identify reference and stressed sites, running random forest models for each metric, evaluating the performance of different

[1]The KS Biotic Index is similar to the Hilsenhoff Biotic Index (Hilsenhoff 1987). Taxa are assigned tolerance ratings based on their sensitivity to nutrient enrichment and other stressors; tolerance ratings are multiplied by total number of individuals, then summed; this number is then divided by the total number of individuals in the sample to determine the Biotic Index value.

combinations of metrics and selecting the MMI. In addition, the report includes a section on precision analyses to quantify the variability of the KS MMI and its input metrics based on sampling and temporal effects, and a section on exploratory analyses to inform potential numeric thresholds for multiple biological condition categories.

2 Data Compilation and Preparation

MMI 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

We obtained BMI samples from the KDHE SB and SP monitoring programs. The macroinvertebrate dataset spanned forty years (1980-2020) and included a total of 2995 samples from 709 unique sites. 175 of the samples were replicates (same site, same day, different time). 2289 of the samples were collected by the SB program and 706 from the SP program, which began in 2006.

The SB program samples targeted sites that are part of a fixed long-term monitoring network or were selected based on agency informational needs (e.g., TMDL, special studies/investigations). SB sites are on perennial streams that tend to be in larger systems than SP sites. SP sites are mostly randomly selected and include more headwater sites, as well as some intermittent streams. The SP program also samples targeted reference sites, including four long-term Regional Monitoring Network sites (USEPA 2016). Figure 1 shows the distribution of sites across Omernik Level 3 ecoregions.

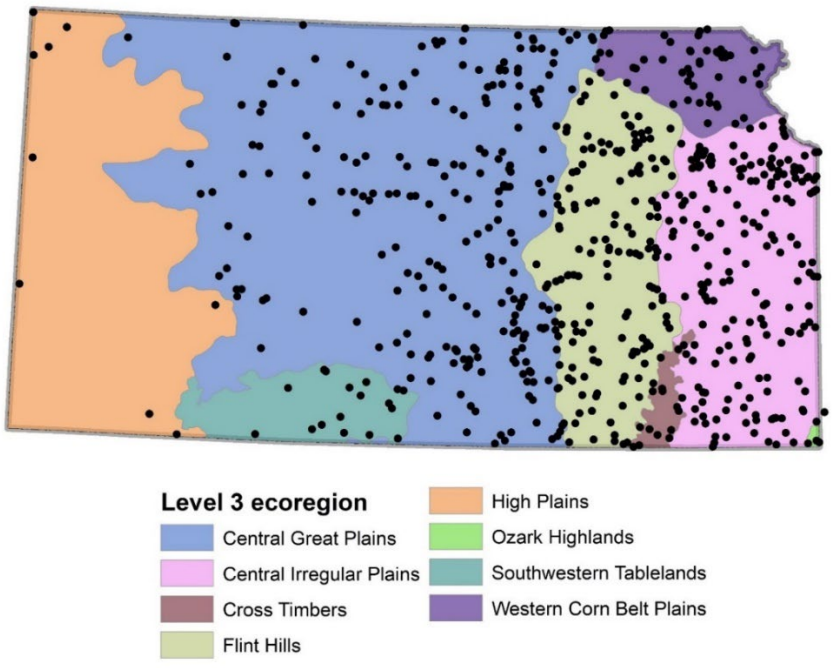


Figure 1. Locations of KDHE BMI sites, with Omernik Level 3 ecoregion as the backdrop.

2.1.2 Collection method

BMI samples were collected in accordance with KDHE SP and SB program standard operating procedures and Quality Assurance Project Plans (QAPP) for wadeable streams (KDHE 2021, KDHE 2022). Table 1 summarizes the componens of the SP and SB protocols. Methods between the two programs are largely similar. Both programs sample from late spring through fall and composite organisms from multiple habitats. They try to maximize the diversity of organisms being captured. Targeted macrohabitats include riffles, leafpacks, undercut banks and rootmats, fine substrate, aquatic vegetation, and large woody debris (LWD). The one or two-person field crews collect organisms with 500-µm mesh D-frame nets, using sweeps or kicks, or by hand picking organisms off large hard substrates. Organisms are picked in the field, roughly proportional to their relative abundance. Samples are preserved in the field with denatured 95% ethanol. KDHE staff identify the samples in the lab, going to genus or species-level as allowed by available keys, specimen condition, and specimen maturity. Differences between the SP vs. SB programs include: length of sampling reach (150-m for the SP program versus SB reaches are established at the sampler's discretion (undefined length) to include areas with the most biologically rich habitat); and target number of total organisms (200 for the SP program vs. 'all that can be captured in one-person hour of effort' for the SB program).

Table 1. Components of the SP and SB BMI collection and processing methods.

Method Component	Probabilistic (SP)	Targeted (SB)
Index period for sampling	April 15-September 30	May to leaf fall (October)
Sampling reach length	150 m	Undefined, flexible
How sampling locations are selected	Targeted macrohabitats within reach	All available macrohabitats that can be sampled in one person-hour. Emphasis on biologically rich microhabitats within representative macrohabitats
Sampling device/s	500-µm mesh D-frame net and fine point forceps	
Sampling duration	60 minutes of active sampling (1 person x 60 minutes or 2 people x 30 minutes). Time is stopped when moving from location to location. SP sampling time may be extended to achieve sample size of 200 total organisms	
Other constraints or components to sampling	No more than 50 individuals per D-net sweep/kick or hand-picked substrate; field pick in proportion to apparent abundance of productive habitats (riffles, leafpacks, undercut banks and rootmats, fine substrate, aquatic vegetation, LWD, etc.) while being conscious to capture all available diversity, including large/rare	
Field or lab pick	Field	
Target sample size	200	All that can be captured within one-person hour of effort, given the constraints listed above
Level of taxonomic resolution	Lowest practicable, usually genus/species	

2.1.3 Taxa attributes

We compiled the BMI data into a MS Access relational database. There were 729 unique taxa names in the dataset. KDHE checked taxa names against the Integrated Taxonomic Information System (ITIS; https://itis.gov/) for validity and also provided the phylogeny (Phylum/Class/Order/Family/Genus/Genus Species). Over the 40-year span of the dataset, some TaxaIDs had changed. Per input from KDHE, we changed TaxaIDs where needed to account for these updates (Table 2).

Table 2. Some TaxaIDs were updated to account for changes in nomenclature during certain time periods.

Original TaxaID	Update
CAENIS	Prior to 2005, changed to AMERCAENIS/CAENIS
ARIGOMPHUS LENTULUS	Prior to 2009, changed to ARIGOMPHUS LENTULUS/GOMPHUS MILITARIS
BAETIS	Prior to 2005, changed to BAETIS/FALLCEON/HETEROCLOEON/PLAUDITUS
CHOROTERPES	Prior to 2000, changed to CHOROTERPES/NEOCHOROTERPES
HYDROPORUS	All records were collapsed to tribe (HYDROPORINI)
LARSIA	Prior to 2010, changed to LARSIA/TELOPELOPIA
PELTODYTES LENGI	Prior to 2012, changed to PELTODYTES DUODECIMPUNCTATUS/LENGI
PELTODYTES SEXMACULATUS	Prior to 2012, changed to PELTODYTES SEXMACULATUS/DUNAVANI/SHERMANI
TRIBELOS	Prior to 2010, changed to PHAENOPSECTRA/TRIBELOS
CRICOTOPUS	Prior to 2010, changed to CRICOTOPUS/ORTHOCLADIUS
ORTHOCLADIUS	Prior to 2010, changed to CRICOTOPUS/ORTHOCLADIUS

With input from KDHE staff, we designated 'non-target' taxa, which were excluded from MMI calculations. Non-target taxa included Arachnida, Argulus, Branchiobdellidae, Collembola, Gordiidae, Gordius, Isotomurus, Metacopina, Nemata, Spongillidae and Staphylinidae.

We compiled five types of traits: functional feeding group (FFG), habit, tolerance to general disturbance, life cycle/voltinism, and thermal preference (Table 3). Trait assignments and sources of the assignments are provided in Attachment 1. We made use of the traits table that had been compiled for the Great Plains Biological Condition Gradient (GP BCG) project (Stamp et al., in progress), which included traits from biomonitoring programs in six states (Kansas, Iowa, Nebraska, Missouri, Oklahoma, Minnesota). FFG and habit assignments were based on level of agreement across data sources, using the most commonly-used trait assignment ('majority rules'). If there was no clear majority, we assigned multiple traits. The life cycle/voltinism assignments were based on the EPA National Rivers and Streams Assessment (NRSA) traits table where available, supplemented by Twardochleb et al. (2021).

There were two types of tolerance traits: numeric (scaled 0-10) and categorical (based on BCG taxa attributes (USEPA 2016) from the GP BCG project). Assignments were based on multiple lines of evidence, including KDHE's existing tolerance ratings (Huggins and Moffett 1988, KBTOLR), which were based on literature reviews and best professional judgment. In addition, we considered results from taxa tolerance analyses that were performed on several variables from the United States Environmental Protection Agency (USEPA) Stream-Catchment (StreamCat) dataset (Hill et al. 2016) that capture general disturbance in Midwestern streams: the chemistry component of the Index of

Watershed Integrity (IWI-WCHEM version 2.1) (Thornbrugh et al. 2018, Johnson et al. 2018), percent row crop and percent urban land use. 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).

We characterized the relationship between BMI taxa and the disturbance variables with three measures: 1) central tendency, based on weighted average optima calculations and relative abundance data; 2) lower and upper limits, based on the 10th and 90th percentiles of taxon occurrence; and 3) response shape, based on Generalized Additive Model (GAM) plots (Hastie and Tibshirani 1999) (see examples in Figure 2). We performed the taxa tolerance analyses on two datasets: KDHE (alone) and the six-state regional dataset compiled for the GP BCG project.
A group of regional biologists (including several from KDHE) reviewed results and assigned taxa to the following BCG attribute categories: highly sensitive (Attrib II), sensitive (Attrib III), intermediate tolerant (Attrib IV)[2] and tolerant (Attrib V). Results and documentation of the KDHE taxa tolerance analyses can be found in Tetra Tech (2023)[3].

Thermal preference assignments were exploratory. Thermal stream conditions can be natural. The metrics were tested to determine whether thermal indicators were also related to disturbance. We assigned taxa to two categories (cool or warm) based on results from a taxa tolerance analysis on the KDHE dataset, using the modeled mean summer (July/August) stream temperature (MSST) metric (Hill et al. 2013) from USEPA's StreamCat dataset (Hill et al. 2016).

[2] Attribute IV was broken into three subgroups: 'better', which are mostly habitat specialists that are not particularly sensitive to water quality, but appear to be sensitive to siltation/high% fines; 'middle', which occur everywhere; 'worse', which are ubiquitous, but typically have higher abundance at poorer sites, such as those affected by nutrient enrichment or sedimentation.

[3] At a later time, we also ran taxa tolerance analyses on selected water chemistry variables from the KDHE dataset: specific conductance, total phosphorus (TP), nitrate and total suspended solids (TSS). Results are included in Tetra Tech 2023

Table 3. Traits that were used in biological metric calculations.

Attribute	Description	Categories	Sources*	Percent of total
Functional feeding group (FFG)	Refers to the primary process for acquiring food resources	collector-gatherer (CG), collector-filterer (CF), predator (PR), scraper (SC), shredder (SH), macrophyte herbivore (MH), omnivore (OM), parasite (PA), piercer-herbivore (PH), xylophage (XY).	KDHE, IA DNR, MO DNR, NE DEQ, MPCA, OCC/OWRB, EPA NRSA, Twardochleb et al. 2020	99.2%
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	KDHE, IA DNR, MO DNR, NE DEQ, MPCA, OCC/OWRB, EPA NRSA, Twardochleb et al. 2020	98.9%
Tolerance value (numeric)	Tolerance to general disturbance	scores range from 0 (most intolerant) to 10 (most tolerant)	KS tolerance analysis (Tetra Tech 2023), Huggins and Moffett (1988), KBTOLR, Regional BCG	86.9%
BCG taxa attribute (categorical) (USEPA 2016)	Tolerance to general disturbance	highly sensitive (II), sensitive (III), intermediate tolerant (IV) - better, middle, worse, tolerant (V)	Regional BCG, KS tolerance analysis (Tetra Tech 2023)	86.8%
Life Cycle/ Voltinism	Number of broods or generations a species typically produces in a year	uni (one), semi, multi (multiple)	EPA NRSA, Twardochleb et al. 2020	82.3%
Thermal preference (exploratory)	Thermal optima based on the modeled mean summer (July/August) stream temperature (MSST) metric from USEPA's StreamCat dataset (Hill et al. 2013, 2016)	cool or warm	KS tolerance analysis (Tetra Tech 2023)	43.0%**

*Source abbreviations: KDHE, Iowa Department of Natural Resources (IA DNR), Missouri Department of Natural Resources (MO DNR), Nebraska Department of Environmental Quality (NE DEQ), Minnesota Pollution Control Agency (MPCA), Oklahoma Conservation Commission (OCC), Oklahoma Water Resources Board (OWRB), and EPA National Rivers and Streams Assessment (NRSA).

**We only used results for taxa that occurred in 30 or more samples in the KDHE dataset (n=297 taxa). A more detailed description of the tolerance analysis can be found in Tetra Tech (2023).

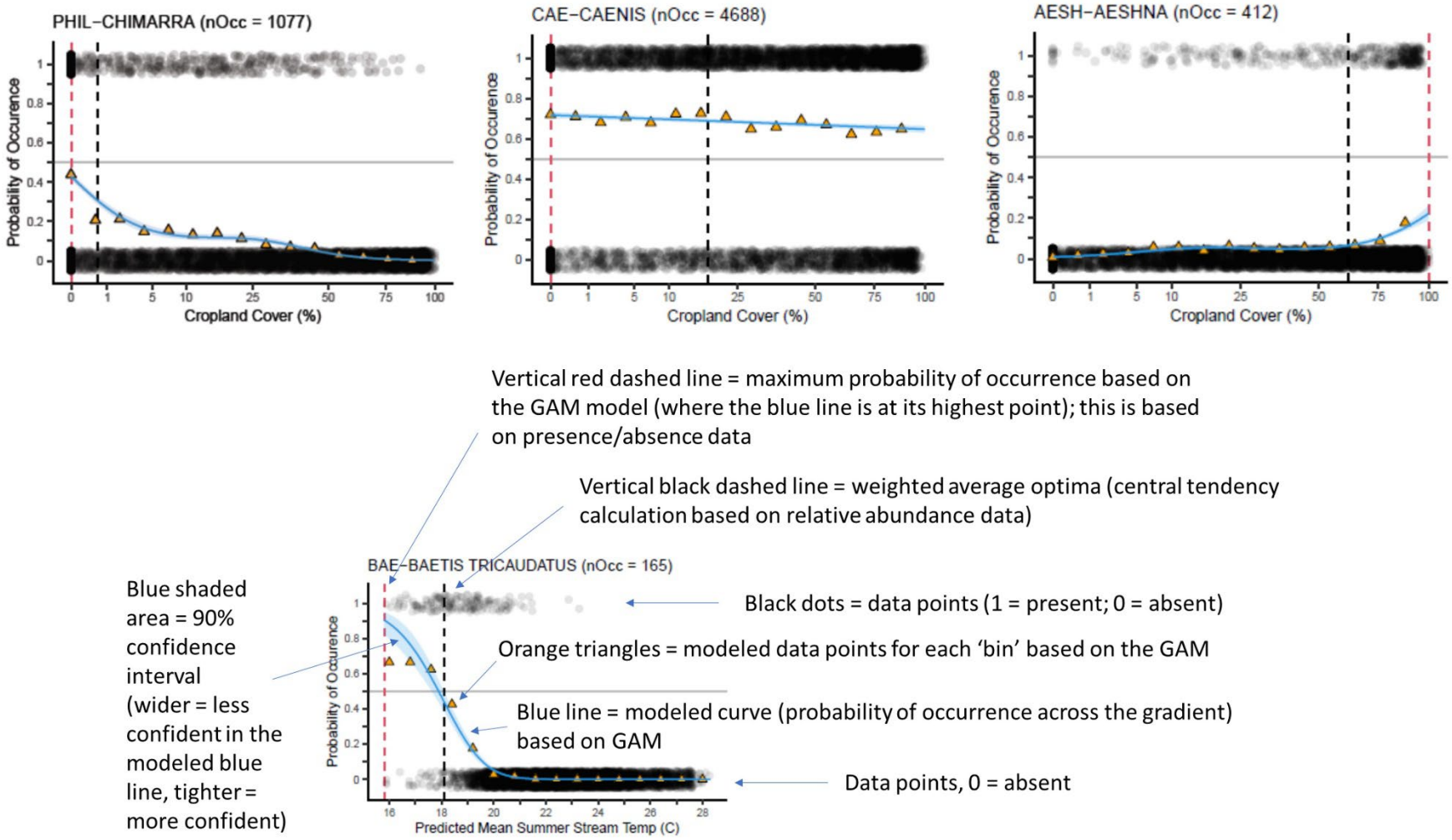


Figure 2. General Additive Model (GAM) plots showing taxon probability of occurrence vs. disturbance variables (in this case, % cropland cover). We included examples of response curves associated with sensitive (*Chimarra* - top left), intermediate tolerant (*Caenis* – top middle) and tolerant taxa (*Aeshna* – top right). The number of samples in which the taxon was detected is indicated by nOcc.

2.1.4 Metric calculations

Metrics were calculated with the BioMonTools R package (Leppo et al. 2021). Appendix A contains the list of metrics that were considered as candidates for inclusion in the MMI. When making metric calculations, non-target taxa were excluded from all metrics, and redundant/non-distinct taxa were excluded from the richness metrics (for more information, see Appendix A).

2.2 Habitat and water chemistry

Habitat and water quality data were utilized where available. Water quality data included over 50 different parameters collected by the Stream Chemistry and SP programs. For the site disturbance characterizations, we used a subset of parameters (specific conductance, pH, turbidity) as secondary screening criteria. During the reference site review process, KDHE staff also noted violations for additional parameters, including lead, copper, selenium and atrazine. We also used median and maximum values for specific conductance, total phosphorus, nitrate and total suspended solids (calculated based on all available data for each site) for the MMI performance evaluation and taxa tolerance analyses.

Habitat data differed across the SP and SB programs. SP surveys included scores from Rapid Habitat Assessment (RHA) surveys (Barbour et al. 1999), which have 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. RHA survey forms differ slightly depending on whether the stream reach is predominantly riffle/run (RR) or glide-pool (GP). SP habitat surveys also included visual estimates of substrate composition, percent flow habitat (riffle, run/glide, pool), and flags for deep (non-wadeable) areas, pooling (due to stream drying) and beaver activity. The SB program used the Stream Macroinvertebrate Assessment Sampled Habitat Index (SMASH), which included evaluations of riffle, run and pool habitat. Each type of habitat was scored based on flow, depth, substrate, edge habitat, algae, embeddedness, and organic debris and then composited into a total overall score.

More detailed information on SP and SB sampling protocols can be found in each program's Quality Assurance Management Plan (KDHE 2021 and KDHE 2020, respectively).

2.3 Landscape-scale information (GIS-based)

Landscape-scale metrics were used for site disturbance characterizations and for predictor variables in the random forest analyses. A primary data source was the USEPA 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). Disturbance variables include measures of overall watershed condition (Index of Watershed Integrity (IWI); Thornbrugh et al. 2018, Johnson et al. 2019), land cover statistics (source: National Land Cover Database (NLCD)), road density, and specific discharges or activities (National Pollutant Discharge Elimination System discharges, mining activity, etc.). Natural variables include geology, soils and climate (air temperature and precipitation). To associate the BMI sampling sites with the StreamCat dataset, we used Geographic Information System software (ArcGIS Pro 2.8) to spatially join the sites with NHDPlusV2 stream segments, and then linked

the sites to the StreamCat data tables via the unique identifiers COMIDs and FEATUREIDs from the NHDPlusV2 dataset.

3 Disturbance index

3.1 Purpose

When developing new biological indices, one of the first steps is to identify reference and stressed sites. Biotic indices are calibrated and validated based on a disturbance gradient, therefore, accurately capturing the full gradient, from best to worst, is necessary and important. Reference sites serve several purposes including to identify metric expectations with the least levels of disturbance, calibrate and validate the index, identify site classes, and set biocriteria thresholds. Stressed sites (i.e., most disturbed sites), identified using criteria from the opposite end of the disturbance scale than reference, can be used in conjunction with reference sites to evaluate the response of metrics along the stressor gradient from worst to best. The direction and strength of responses can be used for selecting candidate metrics for inclusion in a biotic index as well as properly scoring them using a common scale.

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, MMIs 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 biology (USEPA 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 a single index or multiple measures) may be based on partial catchments or the entire watershed. Land use categories that are commonly summarized and used as criteria include imperviousness, agriculture, and urban (USEPA 2013). Similarly, other landscape-scale factors such as point-source pollution sources and infrastructure are evaluated at multiple scales. Once developed, the disturbance gradient is applied to all sites within a dataset to separate them into disturbance categories.

To identify reference and stressed sites for MMI development, we worked collaboratively with the KDHE to develop a statewide disturbance index for wadeable streams in Kansas. The disturbance index was calibrated using 709 sites that were spatially distributed throughout the state and represented a wide range of stream sizes and ecoregions (Figure 1). Disturbance index development was an iterative process that followed these general steps:

1. Compile a list of sites for index calibration
2. Compile geospatial data

3. Select disturbance variables
4. Select threshold values
5. Score sites
6. Assign calibration sites to disturbance categories
7. Provide KDHE staff the opportunity to review designations
8. Conduct subsequent revisions as necessary and finalize designations

The full list of candidate disturbance variables that were considered during index development can be found in Appendix B. Candidate variables captured a mosaic of potential stressors to BMI communities in Kansas streams and were selected based on input from KDHE, data availability and the Kansas Reference Stream Report (Angelo et al. 2010)[4]. We delineated the exact watersheds for each site using the USGS StreamStats Batch Processor, and then used ArcGIS Pro 2.8 to create polygons for three additional spatial scales: 1) 100-m riparian buffer extending 5 km upstream of each site; 2) 1-km upstream radius; and 3) 5-km upstream radius (Figure 3). We generated statistics for each spatial scale where appropriate (not all data lended themselves to geospatial analyses), and then averaged the statistics from the three spatial scales. Next, we ran principal component analyses (PCAs) to help determine which variables most effectively separated reference and stressed sites. We selected variables that had the strongest correlations with the PCA axes, were not conceptually redundant with each other, and are known to be associated with degraded key watershed functions. The results of the PCAs and correlation analyses are provided in Appendix B.

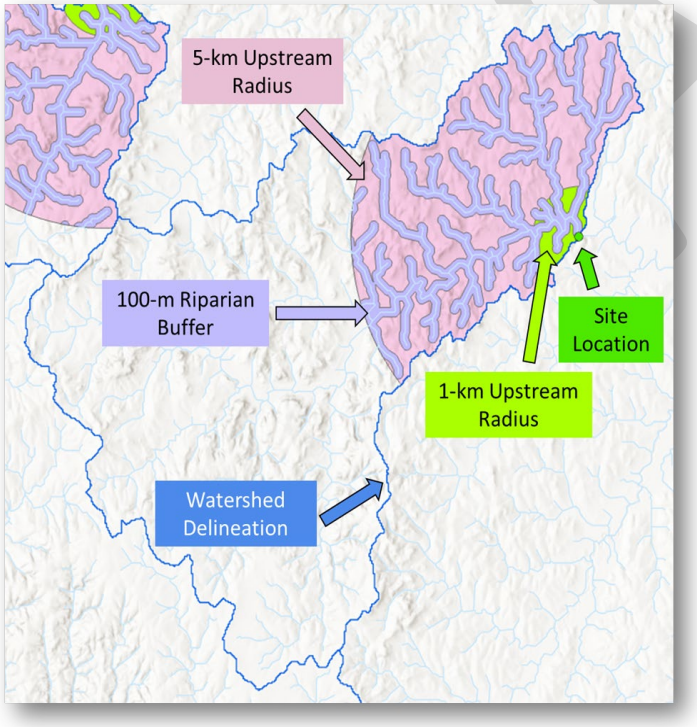


Figure 3. Illustration of the four spatial scales considered during the site disturbance screening process. The WCHEM was based on total watershed scale. The other variables were based on the average of the 100-m riparian buffer, the 1-km radius, and the 5-km radius.

[4] we tried to use the similar variables as Angelo et al. (2010) where possible.

Ultimately we selected five primary screening variables: percent impervious, percent row crop, percent hay/pasture, number of road crossings and the IWI, an overall measure of watershed condition (version 2.1; Thornbrugh et al. 2018, Johnson et al. 2019) (Table 4). We scored each variable on a scale of +3 (least amount of disturbance) to -3 (most amount of disturbance) based on their value in relation to the thresholds in Table 5. The scoring thresholds were not ecologically important; rather they were based entirely on statistics (10th, 25th, 50th, 75th percentile, etc.). After scoring each disturbance variable, we used quantitative 'combination rules' to assign sites to one of six preliminary disturbance categories: Best Reference (lowest level of disturbance), Reference, Other, Somewhat Stressed, Stressed, and Highly Stressed (highest level of disturbance). This multi-level approach is similar to approaches used by several other Midwestern states in recent years (e.g., Illinois (Tetra Tech 2015), Indiana (Jessup and Stamp 2017), Michigan (Tetra Tech 2020), Wisconsin (Tetra Tech 2021). Having six levels of disturbance allows for more transparency and easier communication about differing levels of disturbance in reference and stressed datasets (e.g., is reference in a given ecoregion 'truly natural' vs. 'as natural as possible in the context of ubiquitous disturbance'; this is similar in concept to the minimally disturbed and least disturbed categories proposed in Stoddard et al. (2006)).

In addition to the five primary variables, KDHE staff also considered 'secondary' screening variables that are known to affect key watershed functions but are better suited to be examined on a site-by-site basis (for example, dam attributes vary widely, ranging from very small inactive dams to very large hydropower dams). Secondary variables included dams (presence/absence and storage volume), point source pollutants (National Pollutant Discharge Elimination System (NPDES) major discharges, Superfund sites, Toxic Release Inventory (TRI) locations, mines, confined animal feeding operations (CAFOs)), water withdrawals from irrigation or public water supply water wells, active oil/gas wells, water chemistry (turbidity, pH, conductivity) and 303d listings (Table 6).

We compiled the preliminary disturbance category assignments, primary variable values and scores, secondary screening variables, and site information into a MS Excel worksheet. KDHE staff reviewed the worksheet and either confirmed or recommended changes to the preliminary disturbance categories based on local knowledge and information in the worksheets. The final version of the worksheet can be found in Attachment 2.

Table 4. Five primary disturbance variables were used in the KS disturbance index.

Primary stressor variable	Spatial extent	Source
% Impervious	average of 100-m, 1-km and 5-km (based on exact watershed delineation)	NLCD 2016
% Row crop		
% Hay/pasture		
# Road crossings		Roads - www.kansasgis.org; streams - NHDPlusV2
Index of Watershed Integrity (IWI) (Thornbrugh et al. 2018)	Total watershed*	StreamCat (Hill et al. 2016)

*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.

Table 5. Percentile-based thresholds were used to score each primary variable on a scale of +3 (least amount of disturbance) to -3 (most amount of disturbance). To meet reference level, all variables had to score 0 or higher and the majority of metrics needed to have a positive score.

Metric Scoring	% Row crop	% Hay/pasture	% Impervious	IWI*	# Road crossings
3 (least disturbed)	< 2.7	0	< 0.3	> 0.58	< 2.7
2	2.7 to 13.3	> 0 to 0.1	0.3 to 0.4	0.42 to 0.58	2.7 to 4.3
1	13.4 to 29.2	0.2 to 4.1	0.5 to 0.7	0.33 to 0.41	4.4 to 7.6
0	29.3 to 46.7	4.2 to 23.3	0.8 to 1.6	0.26 to 0.32	7.7 to 11.3
-1	46.8 to 55.4	23.4 to 33.5	1.7 to 4.2	0.23 to 0.25	11.4 to 13.6
-2	55.5 to 68.1	33.6 to 47.8	4.3 to 17.8	0.19 to 0.22	13.7 to 17.3
-3 (most disturbed)	> 68.1	> 47.8	> 17.8	< 0.19	> 17.3

*IWI scores range from 0 (worst) to 1 (best)

Table 6. When reviewing disturbance ratings, KDHE considered primary variables (Table 4) and 'secondary' screening variables.

Secondary screening variable	Spatial extent	Source
303d listings	Site-specific	Known 303d listings
# Confined animal feeding operations (CAFO)	5-km, 100-m	KDHE
# Active oil/gas wells	5-km, 100-m	https://hub.kansasgis.org/
# Irrigation or public water supply water wells	5-km, 100-m	Kansas WWC5 Database - http://www.kansasgis.org/catalog/index.cfm
% Irrigated land cover	5-km	Moderate Resolution Imaging Spectroradiometer (MODIS) Irrigated Agriculture Dataset for the US - https://www.usgs.gov/media/images/mirad
Road length (km)	100-m	www.kansasgis.org
% Canal/Pipeline	100-m	https://nhdplus.com/NHDPlus/
# Mines	100-m	USGS Mineral Resources Data System (MRDS) - https://mrdata.usgs.gov/usmin/
# Dams	100-m	USACE Dams - https://nid.sec.usace.army.mil/
Average normal storage of dams (acre ft)	local catchment, 100-m	
# NPDES major discharge permits	100-m	USEPA Facility Registry Service - https://www.epa.gov/frs/geospatial-data-download-service
# Toxic Release Inventory (TRI) locations	100-m	
# Superfund sites	100-m	
Turbidity	Site-specific	KDHE
pH	Site-specific	
Nitrate	Site-specific	
Conductivity*	Site-specific	KDHE, Olson and Cormier 2019

*We initially considered modeled conductivity data (Olson and Cormier 2019) but ultimately decided not to use it after finding discrepancies with observed data from KDHE. Kansas streams have localized areas with naturally high conductivity, which makes it difficult to model and use as a universal stressor screening variable.

3.3 Results

Table 7 shows the distribution of sites across disturbance categories and Figure 4 shows their spatial distribution across the landscape. Of the 709 sites, the greatest number (n=219) were in the 'Stress' group and the fewest (n=39) were in the Best Reference group. The sites with the least amount of disturbance were concentrated in the Flint Hills and Southwestern Tablelands Level 3 ecoregions, while sites with the highest levels of disturbance occurred mostly in the Great Plains ecoregions (Central Great Plains, Central Irregular Plains and Western Corn Belt Plains), which have intensive agriculture as well as urban areas like Kansas City and Witchita.

Table 7. Distribution of sites across disturbance categories, grouped by Omernik Level 3 ecoregion. For purposes of MMI development, Best Reference + Reference were combined to form the 'reference' dataset and sites in the High Stress group comprised the 'stress' dataset.

Level 3 ecoregion	Best Reference	Reference	Other	Some Stress	Stress	High Stress	*Totals*
Central Great Plains	6	36	38	36	79	52	***247***
Central Irregular Plains	0	9	17	37	85	59	***207***
Flint Hills	22	48	32	30	20	3	***155***
Western Corn Belt Plains	0	5	1	8	33	12	***59***
Southwestern Tablelands	8	5	1	0	1	0	***15***
Cross Timbers	0	6	4	3	0	0	***13***
High Plains	3	1	4	0	0	1	***9***
Ozark Highlands	0	1	1	0	1	1	***4***
Totals	***39***	***111***	***98***	***114***	***219***	***128***	***709***

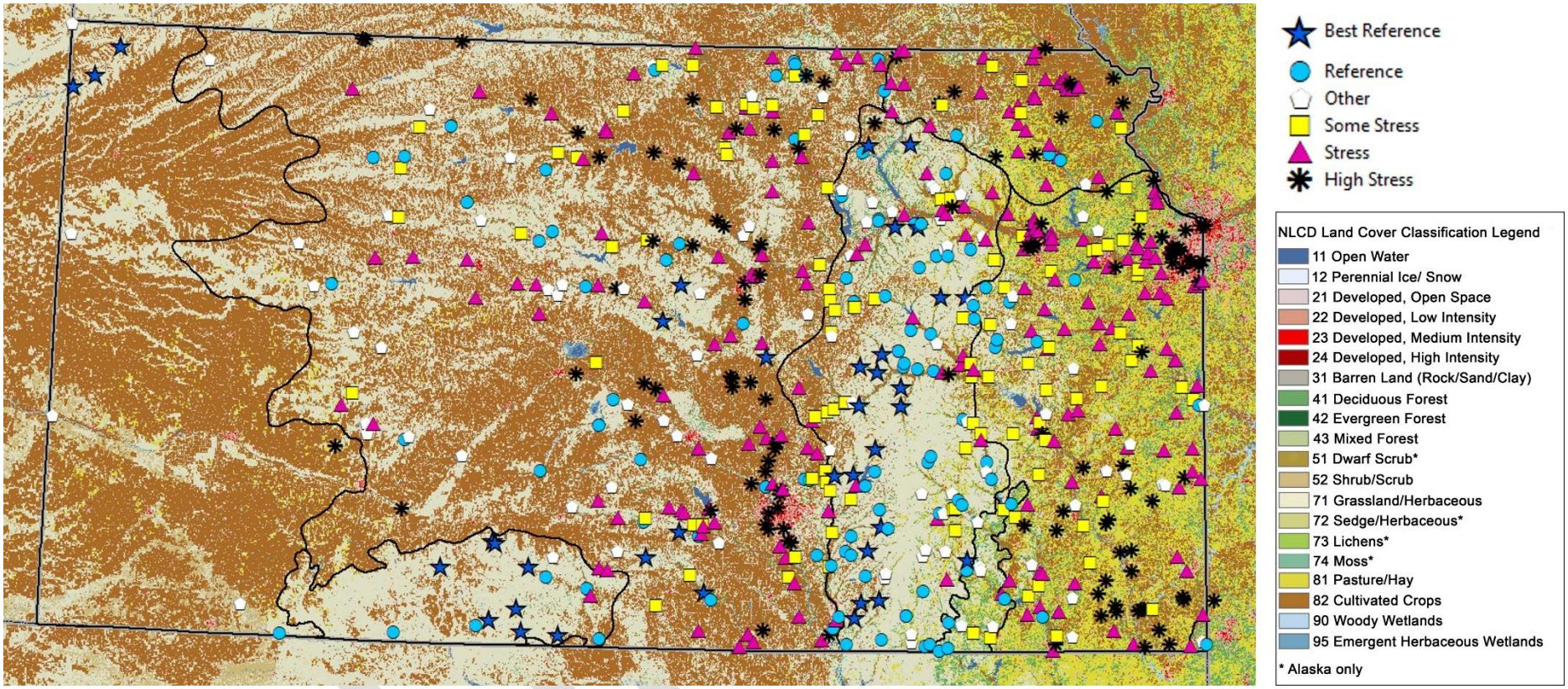


Figure 4. Sites color-coded by disturbance category, against a land cover (NLCD 2016) and Omernik Level 3 ecoregion backdrop.

4 Method for MMI development

4.1 Overview

Initially we headed down the path of doing a 'traditional' MMI, which involved performing a site classification analysis to develop discrete bio-geographical site classes. While the results (provided in Appendix C) were informative, they were ultimately inappropriate because they did not result in categories of site types that could be reliably used in metric and index sensitivity analyses. Because the performance of preliminary MMIs was mediocre, we decided to try the random forest approach (Tang et al. 2016, Carlisle et al. 2022). In the random forest analysis, all samples across the state are incorporated into one model that adjusts metric expectations for each site. The classification is continuous, not in discrete site classes. Random forest accounts for all important variables at one time and can interpret step, sigmoidal, bimodal, asymptotic, linear, or no response relationships. This results in specific metric predictions for each site based on natural environmental variables observed at reference sites. Observed metric values are compared to model-predicted values to evaluate whether each metric is as predicted or underperforming. By modelling BMI metrics at reference sites to account for natural gradients, it increases metric precision (decreases variation).

Steps in the random forest MMI model development are shown in Figure 5 and are described in greater detail in the ensuing subsections.

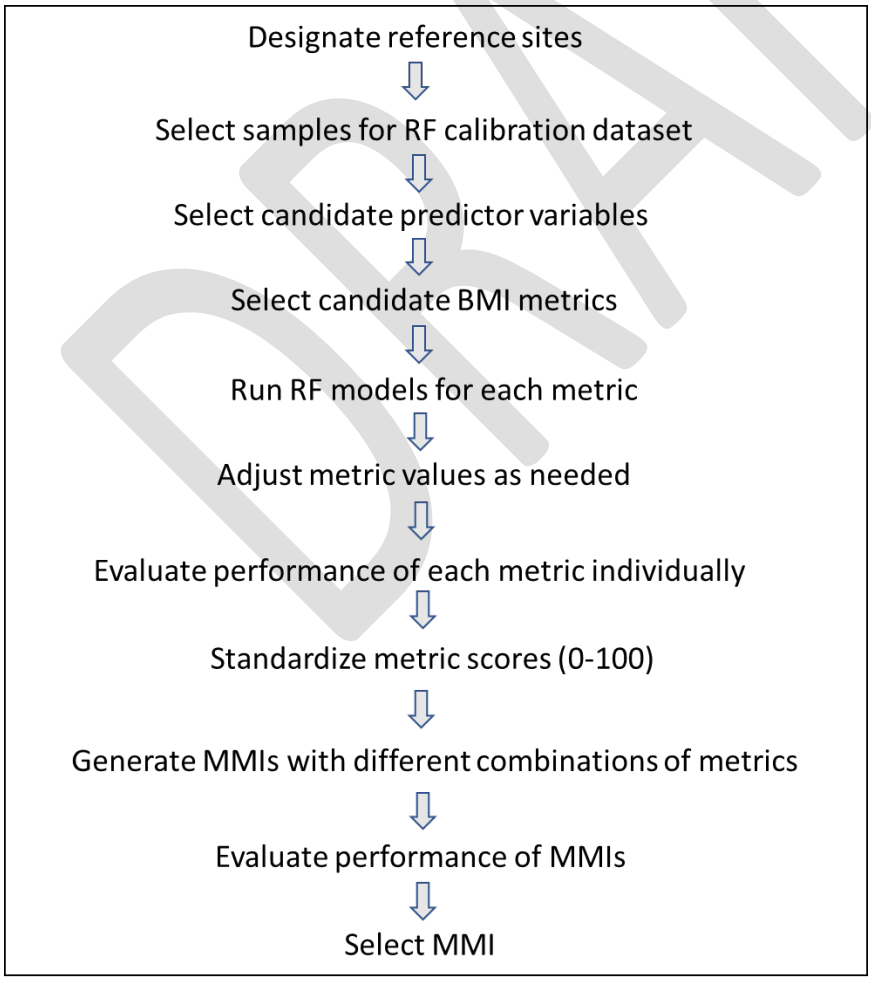


Figure 5. Process for random forest MMI development.

4.2 Prepare calibration dataset

To prepare the MMI calibration dataset, we selected a subset of samples from the full BMI dataset, and then compiled candidate natural predictor variables and BMI metrics.

4.2.1 Sample selection

First we selected reference samples, which are used to build the random forest models. We combined sites from the Best Reference and Reference disturbance categories (n=150; Section 3). We ultimately excluded five of the reference sites because they were missing modeled median flow (cubic feet per second (CFS); Perry et al. 2004). Median flow is known to be an important classification variable for KS BMI communities. KDHE recognizes three CFS classes: < 10 (headwaters), 10-100 and > 100. Each Level 3 ecoregion and CFS class was represented in the reference dataset (Table 8), some to a greater extent than others. The Flint Hills ecoregion had the most number of reference sites (n=66) and many were in the CFS < 10 class (n=92).

Next we selected the 'stress' samples, which were used to evaluate MMI performance. The 'stress' dataset was comprised of the High Stress sites (n=122; Section 3). Some reference and stressed sites had multiple samples. In these situations we selected one sample per site for the analysis. We tried to avoid samples collected before 2000, outside the normal May-September collection period and during years with extreme drought or high flow events. These limitations were disregarded when few samples were available at a site and none met all the preferred conditions. In addition, we included 974 samples that represented intermediate levels of disturbance (Other, Some Stress, Stress; Section 3). These were used to help evaluate the response of the MMI to stressors and to score the MMI metrics (metrics were scaled based on the 5th and 95th percentiles of the dataset). In total, 1241 samples (out of 2995) were included in the MMI calibration dataset. Figure 6 shows the spatial distribution of reference, stressed and intermediate sites across the landscape. The box plots in Figure 7 show the levels of disturbance represented in the reference and stress datasets, as measured by the five primary variables used in the disturbance index (Section 3).

Table 8. Distribution of reference sites across Omernik Level 3 ecoregions and modeled median flow class (CFS).

Level 3 ecoregion	Total	CFS < 10	CFS 10 - 100	CFS > 100
Flint Hills	***66***	48	17	1
Central Great Plains	***42***	28	5	9
Southwestern Tablelands	***13***	7	5	1
Central Irregular Plains	***8***	4	2	2
Cross Timbers	***6***	2	4	
Western Corn Belt Plains	***5***	2	3	
High Plains	***4***		4	
Ozark Highlands	***1***	1		
Total	***145***	***92***	***40***	***13***

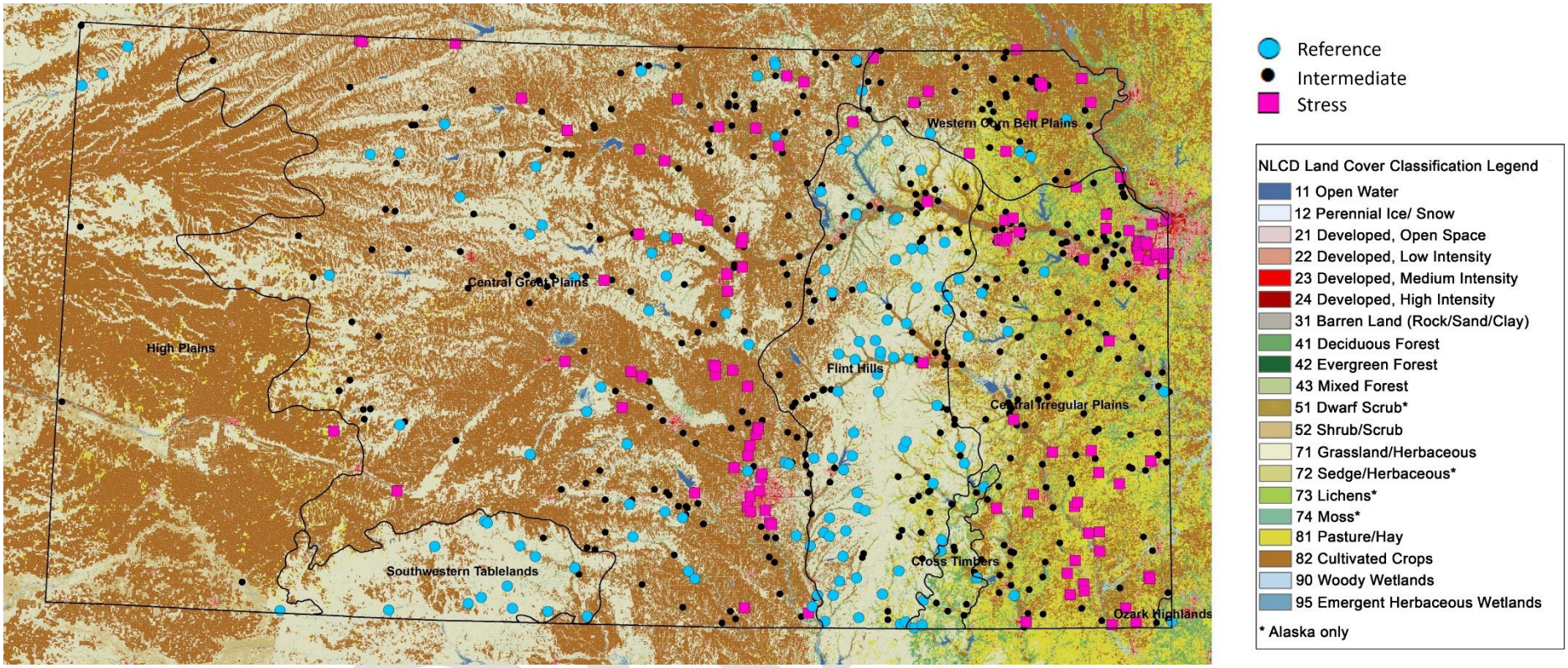


Figure 6. Spatial distribution of reference (Best Reference + Reference), intermediate (Other, Some Stress, Stress) and stressed (High Stress) sites overlaid on land cover (NLCD 2016). Level 3 ecoregion boundaries are also shown.

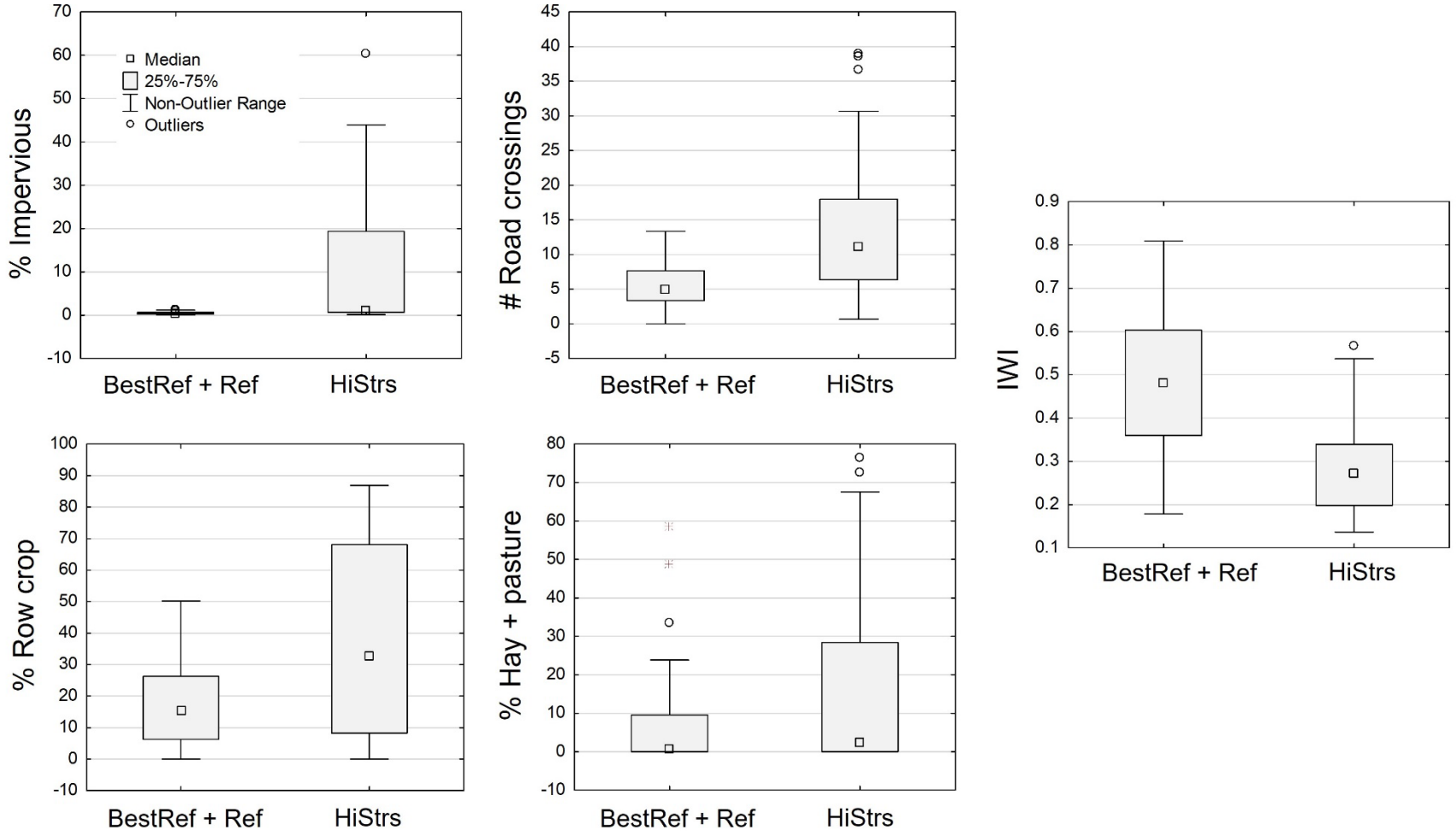


Figure 7. Box plots showing the range of disturbance represented in the reference (Best Reference + Reference) and stressed (High Stress) datasets, as measured by the five primary variables used in the disturbance index (Table 4).

4.2.2 Candidate predictor variables

Predictor variables are used in the random forest analysis to account for natural gradient effects on BMI metrics. For each site, the random forest model looks at site characteristics through decision 'trees' and arrives at a predicted BMI metric value for that site based on the site characteristics, as captured by the predictor values. We started with over 80 candidate predictor variables. Most were continuous, but a few were categorical (e.g., Level 3 ecoregion, HUC). Most came from the USEPA StreamCat dataset[5] (Hill et al. 2016), which includes measures of geology, soils, hydrology and climate. Other candidate variables included longitude, elevation, flowline slope, watershed area, sinuosity and modeled median flow (CFS; Perry et al. 2004). CFS values of 0 were changed to 0.01, which was the dataset minimum value above 0. Predictor selection was an iterative process. After running several iterations of random forest models for the candidate BMI metrics, we narrowed the list down to ~20 predictor variables based on which variables were occurring most frequently in the models (and thus, were consistently explaining the most amount of variability). Attachment 3 contains the initial and final lists of candidate predictor variables that were considered during MMI development.

4.2.3 Candidate BMI metrics

BMI metrics were calculated for the traits listed in Table 3[6], using the BioMonTools R package (Leppo et al. 2021). The metrics represented different aspects of the BMI community (richness, composition, evenness, tolerance, functional feeding group, habit and life cycle). Selection of BMI metrics was an iterative process. We started with over 100 candidate BMI metrics. Metrics were evaluated based on 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
- Precision

We used discrimination efficiency (DE) (Flotemersch et al. 2006, Maxted et al. 2000) and z-scores to assess metric sensitivity. DE was calculated as the percentage of metric scores in stressed sites that were more extreme than each quartile of those in the reference sites. For metrics with a pattern of decreasing value with increasing environmental stress, DE is the percentage of degraded values below the 25th percentile of reference site values. For metrics that increase with increasing stress, DE is the percentage of degraded sites that have values higher than the 75th percentile of reference values. DE can be visualized on box plots (Barbour et al. 1999) of reference and degraded metric or index values with the inter-quartile range plotted as the box (Figure 8). Higher DE denotes more frequent correct association of metric values with site disturbance conditions. DE values ≤ 25% show no discriminatory ability in one direction. Metrics with DE values ≥ 40% were generally considered for inclusion in the KS MMI. However, metric selection was usually dependent on relative DE values within a metric category.

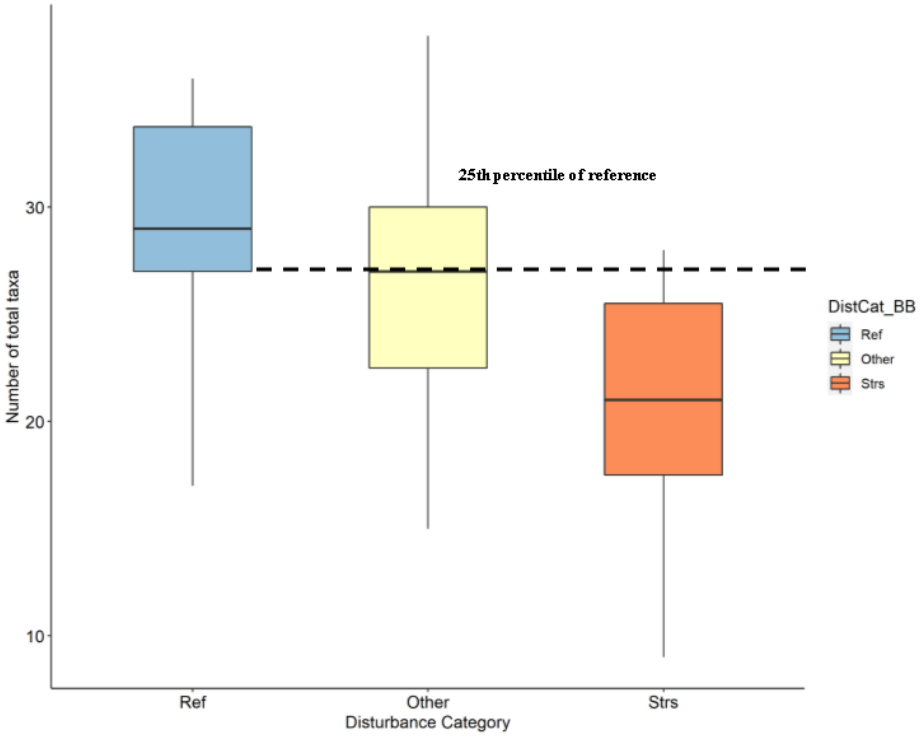
[5] https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset

[6] An exception was the thermal preference metrics, which were removed because they were exploratory and difficult to interpret.

The z-score was a second measure of metric sensitivity to stress. It was calculated as the difference between mean reference and degraded 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 degraded values. Cohen proposed that z values ≥ 0.80 indicated a "large" and preferred effect.

To assess strength and direction of response of metrics to stressors, we used scatterplots and Spearman Rank Correlation Analysis. We also used Spearman Rank Correlation Analysis to assess redundancy between metric pairs and excluded pairs correlated at rho ≥ |0.8|. Precision was analyzed as the coefficient of variation (CV) of sample sets that were collected at the same site on the same day (Cao et al. 2003). Low CVs (e.g., < 30%) are more desirable, as they indicate high precision for a metric. Conversely, high CVs (e.g., > 75%) could contribute to more variability in an index and are less desirable. CVs were calculated from an ANOVA using a replicate set identifier as the grouping variable and metrics as dependent variables. The Root Mean Squared Error (RMSE) was derived as an estimate of the standard deviation of each metric or index. Then the RMSE was standardized to the replicate sample mean to get the CV, which is comparable among metrics.

When we made the decision to switch to the random forest analysis, we had already started screening the BMI metrics (using the techniques described above) and had a list of ~70 candidate metrics. Prior to exploring index compilations, we further reduced the list to < 20 BMI candidate metrics[7]. Selection criteria included representation across metric categories (composition, tolerance, FFG, habit, and life cycle/voltinism), sensitivity as measured by DE and z-score, relative strength of responses to stressors, and metric transparency (e.g., rational response mechanism and uncomplicated calculation). Appendix A includes the lists of BMI metrics that were considered for the random forest MMI analysis.



[7]The 'all subsets' analysis that we later used to explore index alternatives (with different metric combinations) is limited to 20 or fewer metrics due to limited computer capacity to iteratively test all possible subsets of the metrics.

Figure 8. 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).

4.3 MMI development

We followed these steps when developing the MMI:

- Build random forest models for each metric and adjust metrics (if needed)
- Evaluate metric performance
- Score metrics
- Generate MMIs with different metric combinations & evaluate performance

4.3.1 Build Random Forest Models

Using the predictor variables, we created random forest models for each of the 10 top candidate BMI metrics based on the reference dataset. Accounting for natural gradients with the random forest models increases metric precision and decreases variation (Cao et al. 2007, Tang et al. 2016). We used the randomForest package (Liaw and Wiener 2002) in R (R Core Team 2020) to build the random forest models to predict reference site BMI metrics, using methods adapted from Carlisle et al. (2022; National Diatom Index) and Tang et al. (2016). A random forest is a set of 500 separate classification and regression tree models (CART) models, each built with subsets of the training data (data used to build a model, as opposed to testing the result). Each CART model (tree) is built with bootstrap samples from available metric data (about 2/3 of total number of samples) with the BMI metric value as the dependent variable. Each split in each tree considers only a subset of the available predictors selected randomly. The final model consists of all 500 trees. Predictions for training or new data are calculated by using the sample's predictors to identify the terminal node of each tree to which it belongs and averaging the BMI metric values associated with those nodes from each of the 500 trees[8].

Model building was an iterative process. We started with over 80 predictor variables before narrowing it down to < 20 based on which variables were occurring most frequently in the models and were consistently explaining the greatest amount of variability. Metrics were adjusted on a statewide basis depending on the pseudo-Rsq values. The pseudo-Rsq is a relative estimate of model performance, with higher values indicating that the model predictors have greater influence on metric values.

Pseudo r-squared:

$$1 - \frac{\sum(y - \hat{y})2}{\sum(y - mean(y))2}$$

[8] This website offers a good explanation of how random forest models work: https://mlu-explain.github.io/random-forest/

Each BMI metric for which pseudo r-squared values were greater than 15 (meaning 15% of the variation at reference sites was explained) were adjusted by the random forest model (Carlisle et al. 2022). Metrics for which less than 15% of the variation was explained were left in their original (unadjusted) forms. The 15% threshold is halfway between the 20% threshold used by Carlisle et al. (2022) and the 10% threshold used by Tang et al. (2016). Adjusted metrics were then calculated as:

Adjusted valued = Observed value – predicted value

The residual (adjusted) value can be thought of as the variation remaining after accounting for natural factors.

4.3.2 Evaluate metric performance

To assess metric performance, we followed the methods described in Section 4.2.3. We calculated DE and z-scores for each of the top candidate BMI metrics (which were random forest adjusted if needed), and evaluated how effectively they discriminated between reference and 'stress' samples (the higher the DE and z-scores, the better). We also used scatterplots and Spearman Rank Correlation Analysis to evaluate observed vs. expected direction of metric response to stress. During the selection process, the top performing metrics were mostly in the tolerance group (in particular, the BCG taxa attribute metrics). To ensure that the index would capture multiple aspects of the BMI community, we selected the best performing metrics *within* at least three metric categories 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/or that were correlated at rho ≥ |0.8|). Table 9 lists the top 10 BMI metrics that came out of that process, eight of which were random forest adjusted. The best performing metric had a DE of 81%.

Table 9. Metrics that passed selection criteria for consistent responsiveness and were retained for index analysis. RF adj = adjusted by random forest model.

#	Metric category	Metric	DE
1	Composition	# Ephemeroptera, Plecoptera, & Trichoptera taxa (EPT) (RF adj)	70%
2		% EPT individuals minus Hydropsychidae	49%
3		# Coleoptera, Odonata, Ephemeroptera & Trichoptera taxa (COET) (RF adj)	66%
4		% Chironomidae taxa	39%
5	Tolerance	% sensitive* taxa (RF adj)	77%
6		# sensitive* taxa (RF adj)	81%
7		Hilsenhoff Biotic Index (RF adj)	68%
8	Habit	# climber + clinger taxa (RF adj)	70%
9	Voltinism	# semivoltine taxa (RF adj)	66%
10		% semivoltine taxa (RF adj)	55%

*sensitive = BCG attribute III + IV_better

4.3.3 Score metrics

We used the full MMI calibration dataset (which included reference, stressed and intermediate disturbance samples) for metric scoring. Both adjusted metric residuals and unadjusted metric

values were standardized to a 100-point scale such that higher values represented better conditions. The metric standardization was an interpolation of metric values between the 5th and 95th percentiles of each metric distribution (Blocksom 2003). Values calculated as less than 0 or greater than 100 were truncated at 0 and 100. If metric values met the reference expectation (predicted = observed), they received scores in the 50-60 range.

For metrics that decreased with increasing stress (referred to as 'decreasers'; an example is the number of EPT 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 Hilsenhoff Biotic Index (HBI) 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:

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

4.3.4 Generate index compilations and evaluate performance

The final step involved generating MMI index compilations and evaluating their performance. MMI values were calculated as the average of the individual metric scores. Indices comprised of multiple metrics are more sensitive and robust than any single metric and capture multiple response mechanisms to multiple stressors. To find metric combinations that resulted in responsive index options, we used the "all subsets" routine in R software (R Core Team 2020) to combine and evaluate all possible sets of the the top 10 BMI metrics (Table 10). The all-subsets analysis allowed consideration of a plethora of diverse index compositions that simply could not be computed by hand. In successive trials, the code iteratively substituted or removed metrics until all possible subsets of metrics were tested. This resulted in approximately 103,000 different index combinations. Metric combinations were randomly determined, and enough combinations were tested to ensure that every possible combination was included. The combinations were run on four randomized training data sets comprised of 75% of the total datasets. The average DE and z-score of each index across the four data sets was calculated. Looking across the four randomized datasets allowed us to identify the best and most robust MMIs to sample selection variability.

Compiling and evaluating indices was an iterative process. During the first run, we did not put restrictions on combinations of the 10 BMI metrics and included up to six metrics. Having more than six metrics does not necessarily improve index discrimination, can overburden hardware (RAM) limitations and may increase the risk of model over-fitting (Carlisle et al. 2022). We found that the top performing metric combinations were consistently dominated by tolerance metrics (in particular, the BCG taxa attribute metrics). To reduce the redundancy of tolerance metrics, we decided to limit index compilations to five metrics. We also 'forced' the random forest model to include at least one metric from four or more different metric categories (composition, tolerance, FFG, habit, and life cycle/voltinism). We considered the top metric combinations with DEs > 75% and maximum correlations between pairs of metrics < |0.8|. We also evaluated z-scores (the higher, the

better), relative strength of responses to stressors, and metric transparency (e.g., could the metrics be readily communicated and were they responding in ways that could be explained relative to environmental stressors and conditions). We presented the top performing indices to KDHE staff in tables like the one shown in Table 10 and collaboratively decided which index to select.

Table 10. The top performing index compilations were presented to KDHE in tables that showed DE and z-scores (the higher, the better) and mean and maximum Spearman Rank Order Correlations among the metric pairs. The five MMIs shown here were the final ones that were considered. They included slightly different combinations of 10 different metrics. The sensitive taxa metrics were based on BCG attribute III + IV better taxa. All but the percent Chironomidae taxa metric were random forest adjusted (RF adj).

Index #	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5	DE*	z-score*	Mean corr	Max corr
MMI - 1	% semivoltine taxa (RF adj)	# sensitive taxa (RF adj)	# EPT taxa (RF adj)	HBI (RF adj)	# COET taxa (RF adj)	71.4%	1.79	0.36	0.67
MMI - 2	# climber + clinger taxa (RF adj)	# sensitive taxa (RF adj)	# EPT taxa (RF adj)	HBI (RF adj)	% Chironomidae taxa	71.5%	1.69	0.30	0.72
MMI - 3	# climber + clinger taxa (RF adj)	% sensitive taxa (RF adj)	# EPT taxa (RF adj)	HBI (RF adj)	# semivoltine taxa (RF adj)	72.5%	1.97	0.40	0.72
MMI - 4	# climber + clinger taxa (RF adj)	# sensitive taxa (RF adj)	# semivoltine taxa (RF adj)	HBI (RF adj)	% Chironomidae taxa	70.5%	1.68	0.27	0.60
MMI - 5	# climber + clinger taxa (RF adj)	# sensitive taxa (RF adj)	% semivoltine taxa (RF adj)	HBI (RF adj)	% Chironomidae taxa	70.5%	1.60	0.22	0.55

*DE and z-scores in this table were based on an earlier iteration where we used Stress+High Stress samples in the 'stress' dataset

5 Index Selection and Performance

5.1 MMI selection

Through consensus of KDHE staff, MMI #3 (from Table 10) was selected. Its input metrics represented four different metric categories (habit, tolerance, composition, voltinism) (Table 11). The mostly strongly correlated metric pairs in the KS MMI were number of EPT taxa and number of climber + clinger taxa, at rho = 0.71 (Table 12). Other pairs had rho values < 0.6. Each metric had a pseudo R-squared value > 15 and was random forest adjusted using the predictor variables listed in Table 13. Figure 9 shows an example of how the random forest adjustment improved the DE of the percent sensitive taxa metric.

Each KS MMI input metric has its own random forest model and set of predictor variables. There are 12-16 predictor variables per metric, with overlap across models (e.g., eight of the variables are included in all five random forest metric models; Table 14). The importance of the predictor variables varies across models. For example, mean annual precipitation (local catchment scale) and CFS are two of the most important predictor variables in the number of EPT taxa model (meaning they explain more variance than the other predictors in that model) (Table 15). Some of the predictor variables are highly correlated (Attachment 3). For example, elevation, longitude and precipitation are strongly correlated (rho > |0.8|), with a clear east-west gradient (elevation gradually decreases from west to east, while precipitation increases from west to east). Overall, watershed area and CFS are highly correlated (rho = 0.82), with some exceptions in western Kansas where rivers with large watershed areas have limited water due to low precipitation. Figure 10 shows spatial patterns for three of the most used predictor variables (precipitation, watershed area and sand). Maps for all the predictor variables are included in Appendix D.

Table 11. The KS MMI includes five input metrics from four different metric categories. Each metric was random forest adjusted.

#	Metric	Response to stress	Metric Category	Pseudo R-squared	RF Adjust
1	Number of EPT taxa	Decrease	Composition	24.24	Y
2	Percent sensitive taxa (BCG attribute III+IV better)	Decrease	Tolerance	41.26	Y
3	Hilsenhoff Biotic Index (HBI)	Increase	Tolerance	30.33	Y
4	Number of climber + clinger taxa	Decrease	Habit	23.85	Y
5	Number of semivoltine taxa	Decrease	Life cycle	24.48	Y

Table 12. Spearman rho values for the KS MMI input metrics. All correlations are significant at p<0.05.

Variable	HBI (RF adj)	# climber + clinger taxa (RF adj)	# semivoltine taxa (RF adj)	% sensitive taxa (RF adj)	# EPT taxa (RF adj)
HBI (RF adj)	1.00				
# climber + clinger taxa (RF adj)	-0.19	1.00			
# semivoltine taxa (RF adj)	-0.23	0.57	1.00		
% sensitive taxa (RF adj)	-0.45	0.27	0.20	1.00	
# EPT taxa (RF adj)	-0.33	0.71	0.38	0.39	1.00

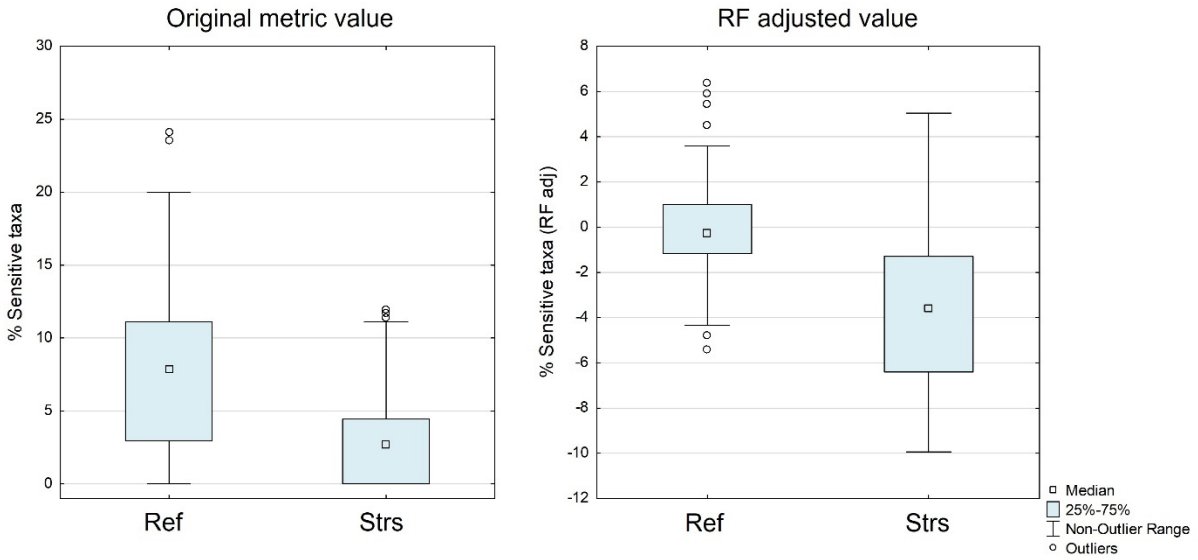


Figure 9. The random forest (RF) adjustment reduced variation in the percent sensitive taxa metric at reference sites. The box plot on the left shows distribution of the raw metric values compared to the adjusted values on the right.

Table 13. Predictor variables used in KS MMI random forest models to account for natural gradient effects on BMI metrics. Minimum and maximum (min/max) values are based on the reference samples in the MMI calibration dataset. MMI scores for sites with values outside of these ranges should be interpreted with caution.

Predictor (abbrev)	Min (Ref)	Max (Ref)	Description	Source
LONG	-102.01	-94.62	Longitude	Site information
PrecipCat	451.78	1150.04	PRISM climate data - 30-year normal mean precipitation (mm): Annual period: 1981-2010 within the catchment	EPA StreamCat
PrecipWs	433.51	1150.04	PRISM climate data - 30-year normal mean precipitation (mm): Annual period: 1981-2010 within the watershed	EPA StreamCat
ElevCat	228.73	1067.03	Mean catchment elevation (m)	EPA StreamCat
WetIndexWs	719.66	1006.10	Mean Composite Topographic Index (CTI) [Wetness Index] within watershed	EPA StreamCat
WtDepWs	87.93	182.92	Mean seasonal water table depth (cm) of soils (STATSGO) within watershed	EPA StreamCat
Fe2O3Cat	1.01	16.02	Mean % of lithological ferric oxide (Fe2O3) content in surface or near surface geology within catchment	EPA StreamCat
K2OWs	0.39	2.29	Mean % of lithological potassium oxide (K2O) content in surface or near surface geology within watershed	EPA StreamCat
L3Eco	--	--	Omernik Level 3 ecoregion - https://www.epa.gov/eco-research/level-iii-and-iv-ecoregions-continental-united-states	EPA ecoregion website*
SandWs	4.04	56.90	Mean % sand content of soils (STATSGO) within watershed	EPA StreamCat
WsAreaSqKm	4.04	97404.96	Watershed area (square km) at NHDPlus stream segment outlet, i.e., at the most downstream location of the vector line segment	EPA StreamCat
MgOCat	0.70	6.85	Mean % of lithological magnesium oxide (MgO) content in surface or near surface geology within catchment	EPA StreamCat
CFS	0.01	562.00	Median flow statistics (ten-year median)	Perry et al. 2004
NWs	0.02	0.35	Mean % of lithological nitrogen (N) content in surface or near surface geology within watershed	EPA StreamCat
Al2O3Ws	1.34	12.27	Mean % of lithological aluminum oxide (Al2O3) content in surface or near surface geology within watershed	EPA StreamCat
ClayWs	13.94	45.08	Mean % clay content of soils (STATSGO) within watershed	EPA StreamCat
PermWs	0.95	18.79	Mean permeability (cm/hour) of soils (STATSGO) within watershed	EPA StreamCat
SWs	0.02	6.87	Mean % of lithological sulfur (S) content in surface or near surface geology within watershed	EPA StreamCat
TmeanCat	10.73	14.78	PRISM climate data - 30-year normal mean temperature (C°): Annual period: 1981-2010 within the catchment	EPA StreamCat

Table 14. Matrix showing which predictor variables are used in which random forest model. Predictors are sorted from most to least commonly used across the five metric models.

Predictor	Number of EPT taxa	Percent sensitive taxa (BCG attribute III+IV better)	HBI	Number of climber + clinger taxa	Number of semivoltine taxa	# BMI metrics
LONG	x	x	x	x	x	5
PrecipCat	x	x	x	x	x	5
PrecipWs	x	x	x	x	x	5
ElevCat	x	x	x	x	x	5
Fe2O3Cat	x	x	x	x	x	5
WtDepWs	x	x	x	x	x	5
K2OWs	x	x	x	x	x	5
WetIndexWs	x	x	x	x	x	5
L3Eco	x	x	x	x		4
SandWs	x		x	x	x	4
WsAreaSqKm	x		x	x	x	4
MgOCat	x			x	x	3
CFS	x				x	2
NWs		x	x			2
Al2O3Ws		x				1
ClayWs		x				1
PermWs		x				1
TmeanCat			x			1
SWs			x			1

Table 15. The importance of the predictor variables varies across the five KS MMI random forest models (e.g., elevation is the most important predictor for # clinger + climber taxa). This table shows the top six predictor variables for each model. Appendix D shows predictor importance plots for each metric.

Metric	Pseudo R-squared	Adjust	Pred1	Pred2	Pred3	Pred4	Pred5	Pred6
Number of climber + clinger taxa	23.85	Y	ElevCat	PrecipCat	MgOCat	LONG	WtDepWs	PrecipWs
Number of semivoltine taxa	24.48	Y	WsAreaSqKm	ElevCat	CFS	Fe2O3Cat	PrecipWs	WetIndexWs
Percent sensitive taxa (BCG attribute III+IV better)	41.26	Y	LONG	L3Eco	PrecipWs	K2OWs	PrecipCat	Al2O3Ws
HBI	30.33	Y	K2OWs	PrecipCat	LONG	PrecipWs	L3Eco	SandWs
Number of EPT taxa	24.24	Y	PrecipCat	CFS	ElevCat	LONG	WtDepWs	WsAreaSqKm

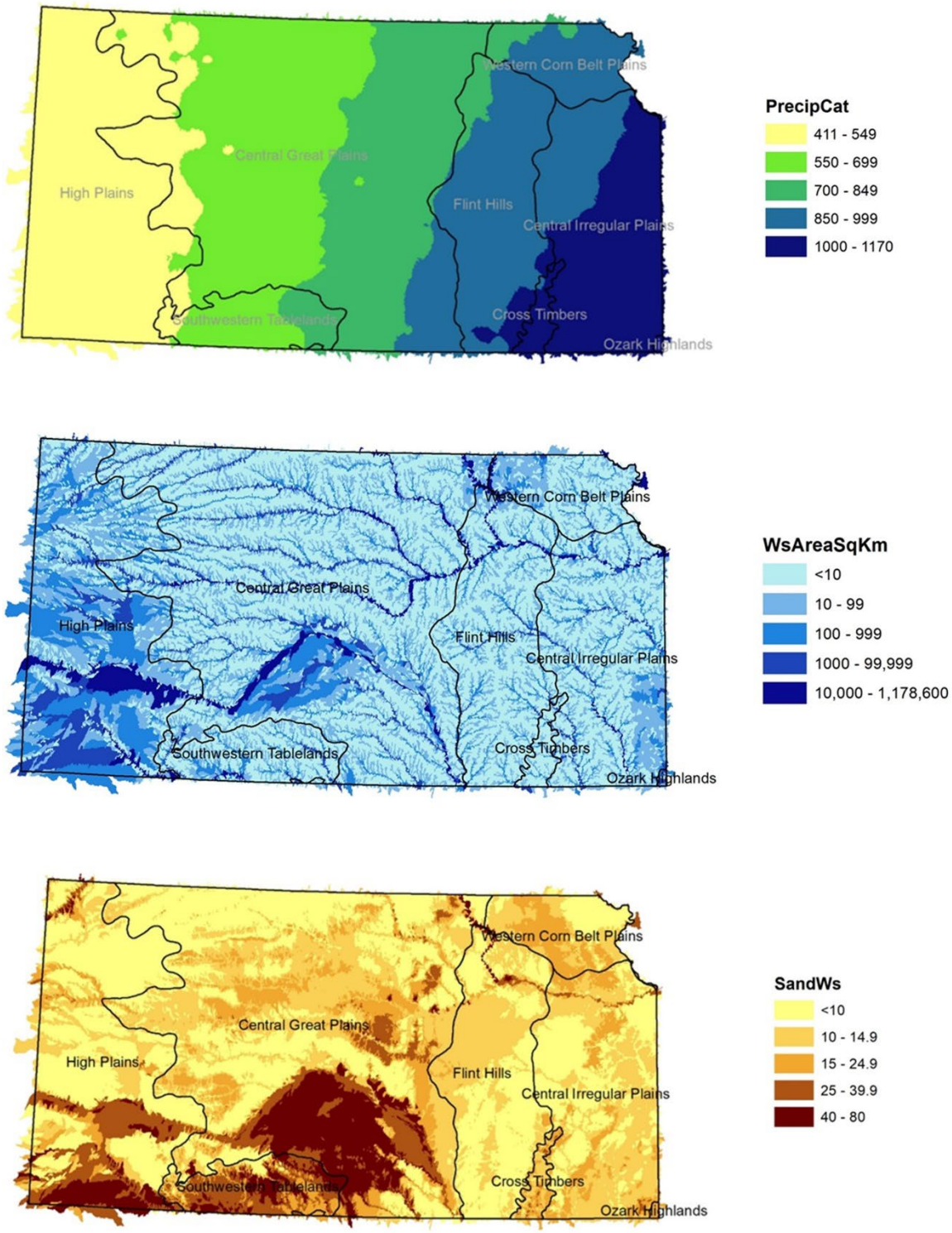


Figure 10. Maps showing spatial patterns for three of the predictor variables (precipitation, watershed area and sand). Maps for all the predictor variables are included in Appendix D.

5.2 MMI performance

We used DE as the primary measure of MMI performance. DE scores for the overall calibration dataset and the four random subsets ranged from 77-78% (Table 16). This meant that 77-78% percent of stressed sites scored below the 25th percentile of reference (Figure 11). DE scores for the individual metrics ranged from 66% to 77% (Table 16).

We also used scatterplots and Spearman Rank correlation analyses to assess the strength and direction of MMI response to the primary and secondary disturbance variables (Tables 4 & 6, respectively). Among the primary disturbance variables, the strongest relationships were with percent impervious, number of road crossings and the IWI, with MMI scores showing a noticeable decrease when percent impervious was ≥ 1%, road crossings were ≥ 10 and IWI scores were < 0.35 (Figure 12). MMI scores did not show a decipherable relationship with percent row crop, nor with percent hay/pasture until it reached ~30% (Figure 13).

Results for the correlation analyses with the secondary disturbance variables are shown in Tables 17-20. They were mostly weak (rho < |0.2|) and did not always occur in the expected direction. Table 17 shows the rho values for the KS MMI and MMI metrics vs water chemistry. Where significant ($p< 0.05$), the MMI was positively correlated with median and maximum values for total phosphorus (TP), nitrate and total suspended solids (TSS)[9]. The strongest relationships were with TP (rho = 0.23) and TSS (rho = 0.31). Of the land cover metrics, which are shown in Table 18, the KS MMI had the strongest relationship with percent natural land cover[10] at the total watershed scale (rho = 0.31). The others, which were based on the 100-m riparian buffer, were very weak (rho < |0.15|). The other two sets of variables (point source pollutants in Table 19 and hydrologic variables (e.g., water withdrawals and impoundments) in Table 20), show weak, mixed signals with rho values ≤ |0.15|. This was not entirely surprising, given how these types of variables tend to have skewed distributions across the landscape (e.g., point source pollutants, like NPDES major discharges and Superfund sites, mines, and CAFOs may occur sparsely, or in clumps) and also have attributes that vary widely (e.g., dams range from very small inactive dams to very large hydropower dams).

Table 16. Discrimination Efficiency (DE) scores for the MMI and MMI input metrics, based on the reference and stress samples in the MMI calibration dataset.

Metric	DE
MMI - overall	78%
Mean MMI - four random subsets	77%
RF adjusted - Percent sensitive taxa (BCG attribute III+IV better)	77%
RF adjusted - Number of climber + clinger taxa	70%
RF adjusted - Number of EPT taxa	70%
RF adjusted - Hilsenhoff Biotic Index (HBI)	68%
RF adjusted - Number of semivoltine taxa	66%

[9] Median and maximum values were calculated for all available data for each site.

[10] % natural land cover was defined as the sum of forest + shrub + grass + water + wetlands

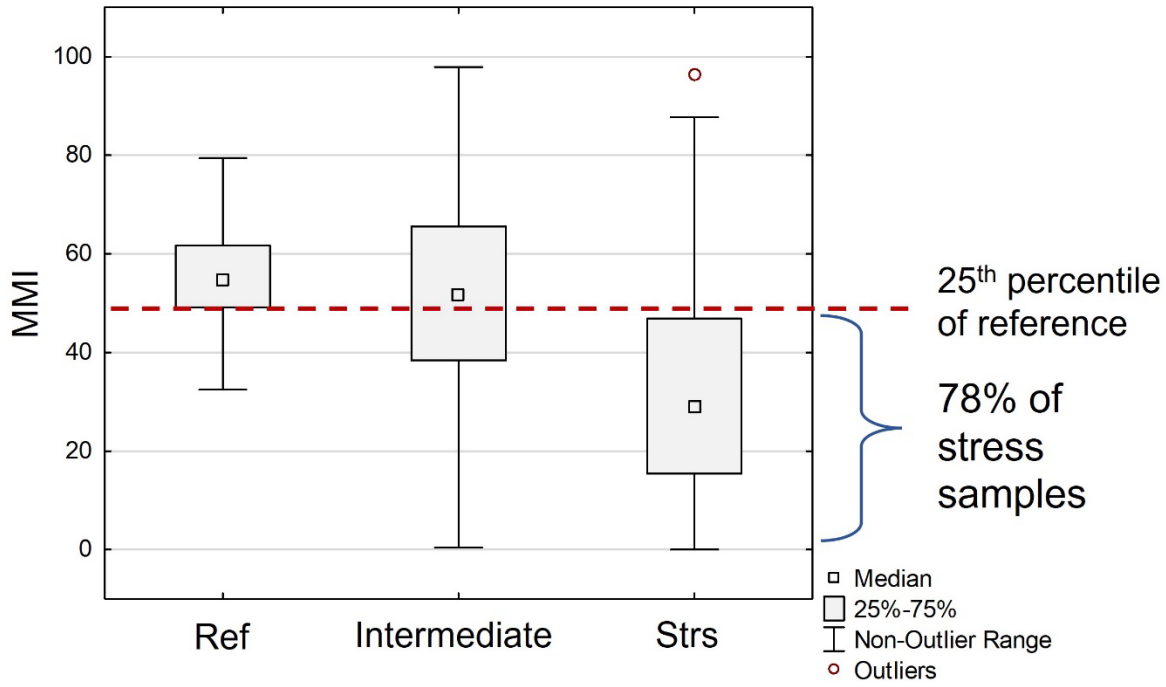


Figure 11. Distribution of MMI scores in the reference, intermediate and stress samples in the MMI calibration dataset. The overall DE of the MMI is 78%, which means 78% percent of stressed sites scored below the 25th percentile of reference (MMI=49).

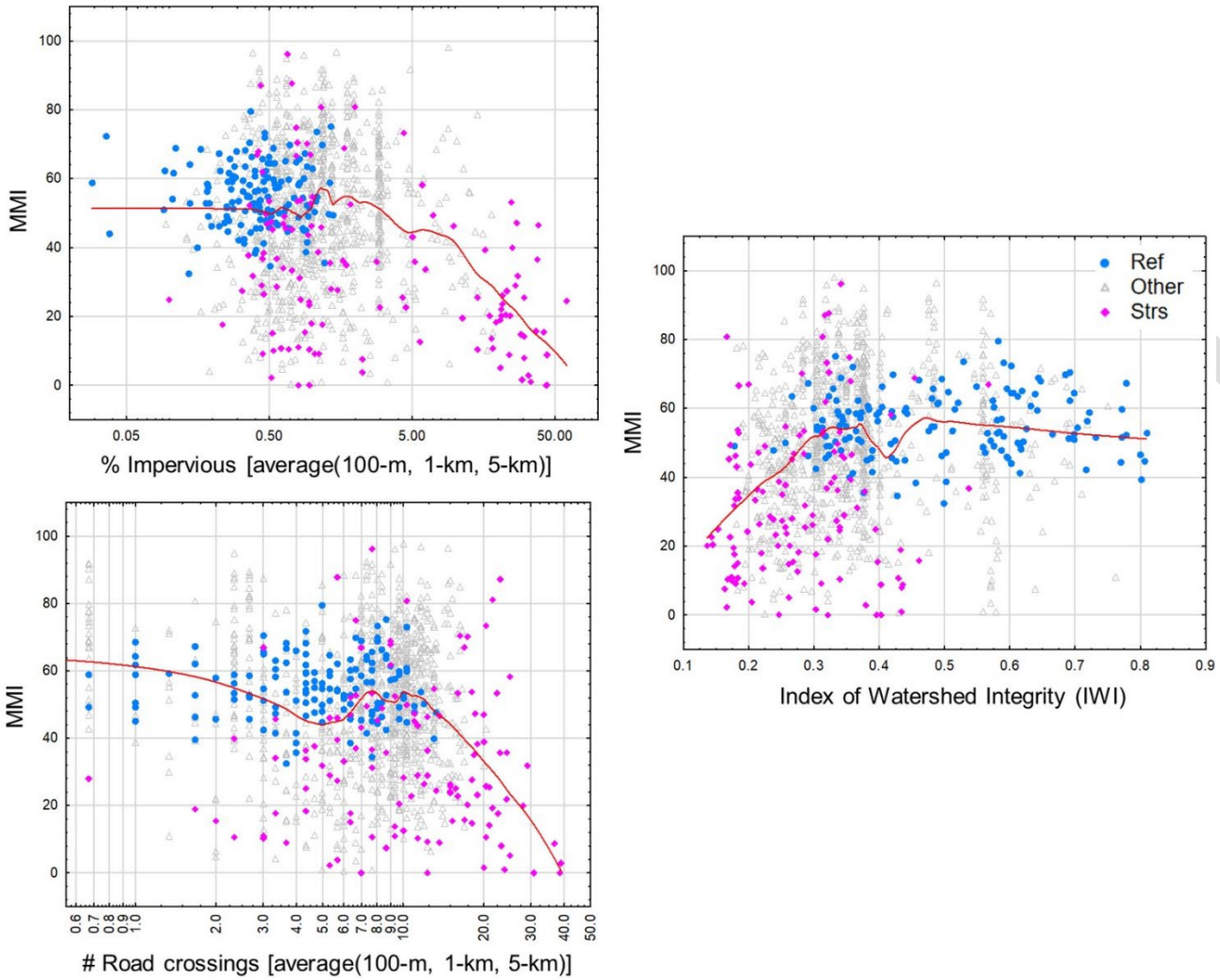


Figure 12. Scatterplots of KS MMI scores vs. three of the primary stressor variables (percent impervious, number of road crossings and IWI), fit with a Locally Weighted Scatterplot Smoothing (LOWESS) line. IWI scores range from 0 (worst) to 1 (best). The IWI is based on the total watershed scale, while the other two metrics are based on the average of 100-m, 1-km and 5-km polygons, using exact watershed delineations. This plot is based on the MMI calibration dataset (n=1241).

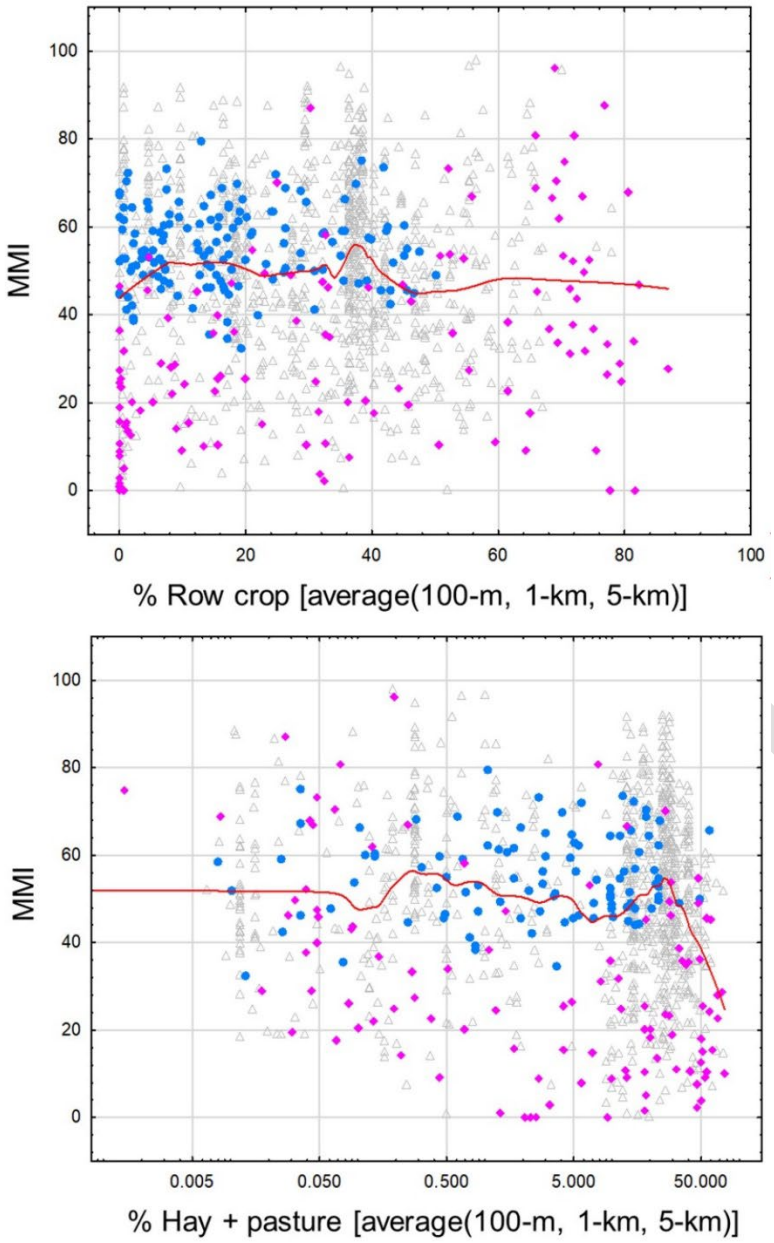


Figure 13. Scatterplots of KS MMI scores vs. percent row crop (top) and percent hay/pasture, fit with a LOWESS line. The land cover metrics are based on the average of 100-m, 1-km and 5-km polygons, using exact watershed delineations, and the NLCD 2016 dataset. This plot is based on the MMI calibration dataset (n=1241).

Table 17. Rho values from Spearman Rank correlation analyses on KS MMI and MMI metrics vs. water chemistry variables. Median and maximum water chemistry values were provided by KDHE and are based on all available grab samples for each site. Blank cells indicate non-significant (p>0.05) relationships. This analysis was performed on the MMI calibration dataset (n=1241).

MMI & input metrics	Phosphorus, total (as P) mg/L		Nitrate (as N) mg/L		Total suspended solids mg/L	
	Median	Max	Median	Max	Median	Max
MMI		0.23	0.12	0.21	0.10	0.31
HBI (RF adj)	0.12	-0.07				-0.13
# climber + clinger taxa (RF adj)		0.24	0.16	0.26	0.09	0.28
# semivoltine taxa (RF adj)		0.11	0.09	0.14		0.14
% sensitive* taxa (RF adj)	-0.10				-0.07	0.09
# EPT taxa (RF adj)	0.08	0.31	0.15	0.29	0.26	0.41

Table 18. Rho values from Spearman Rank correlation analyses on KS MMI and MMI metrics vs. land cover metrics. Blank cells indicate non-significant (p>0.05) relationships. This analysis was performed on the MMI calibration dataset (n=1241).

MMI & input metrics	Land cover within the 100-m riparian buffer (NLCD 2016)							Total watershed
	% Impervious	% Urban	% Row crop	% Hay+pasture	% Forest	% Shrub	% Grass	% Natural land cover*
MMI	-0.12	-0.11		-0.10			0.14	0.31
HBI (RF adj)	0.12	0.11	0.07		-0.09			-0.14
# climber + clinger taxa (RF adj)	-0.06	-0.06	0.07	-0.14	-0.11	-0.08	0.21	0.26
# semivoltine taxa (RF adj)								0.13
% sensitive* taxa (RF adj)	-0.20	-0.19			0.07	0.07	0.09	0.25
# EPT taxa (RF adj)	-0.06		0.18	-0.19	-0.11	-0.06	0.20	0.36

*natural land cover was defined as the sum of forest + shrub + grass + water + wetlands and was based on the NLCD 2019 dataset

Table 19. Rho values from Spearman Rank correlation analyses on KS MMI and MMI metrics vs. point source pollutants. Blank cells indicate non-significant (p>0.05) relationships. This analysis was performed on the MMI calibration dataset (n=1241). # Mines includes sand, gravel, and coal mines. NPDES = National Pollutant Discharge Elimination System. NPL = Superfund National Priorities List. TRI = Toxic Release Inventory. CAFO = Concentrated Animal Feeding Operation.

MMI & input metrics	within the 100-m riparian buffer						within 5-km upstream	
	# Mines	# Major NPDES permits	# NPL Superfund sites	# TRI sites	# Active oil/gas wells	CAFOs	# Active oil/gas wells	CAFOs
MMI	0.10	0.08		-0.10				
HBI (RF adj)	-0.06			0.07				
# climber + clinger taxa (RF adj)	0.09			-0.07		0.06		
# semivoltine taxa (RF adj)	0.13	0.10		-0.07	0.08		0.09	
% sensitive* taxa (RF adj)	0.06	0.10		-0.11	0.08		0.08	
# EPT taxa (RF adj)		-0.06	-0.06	-0.08		0.10		0.06

Table 20. Rho values from Spearman Rank correlation analyses on KS MMI and MMI metrics vs. hydrologic variables (water withdrawals, impoundments). Blank cells indicate non-significant (p>0.05) relationships. This analysis was performed on the MMI calibration dataset (n=1241). PWS = public water supply.

MMI & input metrics	within the 100-m riparian buffer					1-km upstream	5-km upstream		3 scales*
	# Surface water withdrawals	# Water wells	Mean normal storage of dams (Acre Ft)	# Irrigation or PWS water wells	# Dams	# Dams	% Irrigated land cover	# Irrigation or PWS water wells	Mean normal storage of dams (Acre Ft)
MMI	0.06	0.08		0.13		-0.11	0.13	0.10	-0.07
HBI (RF adj)	-0.09					0.07	-0.06		
# climber + clinger taxa (RF adj)		0.10	-0.08	0.09	-0.08	-0.10	0.10	0.06	-0.09
# semivoltine taxa (RF adj)				0.11		-0.07			
% sensitive* taxa (RF adj)				0.12		-0.06	0.10	0.08	
# EPT taxa (RF adj)	0.07	0.17	-0.09	0.15	-0.09	-0.09	0.20	0.16	-0.12

*based on the average of 100-m, 1-km and 5-km polygons, using exact watershed delineations

5.3 Patterns

In this section we examined relationships between MMI scores and:

- Program (SB vs. SP)
- Time (Julian date, month, year)
- Stream size (CFS, watershed area)
- Habitat type (riffle-run vs. glide-pool)
- Level 3 ecoregion
- Hydrologic Unit Code (HUC)
- Specific conductance

5.3.1 Program

The random forest model calibration dataset (reference samples only) included 27 samples from the SB program (1984-2019) and 118 from the SP program (2006-2020) (Figure 14). When we added in all the reference site data (some sites had multiple years of data) and compared distributions of MMI scores across the two programs, we found that SB samples had slightly higher median MMI scores (SB=59 vs. SP=54.7). Median scores for the HBI, climber/clinger, EPT and semivoltine metrics were slightly lower in SP samples and the median score for the percent sensitive taxa metric was slightly higher in SP samples (Figure 15). We also found that the MMI was better at discriminating between reference and stress sites with SP samples (meaning less overlap between the 25th percentile of reference and 75th percentile of stress). This pattern held true with the input metrics as well (Figure 16).

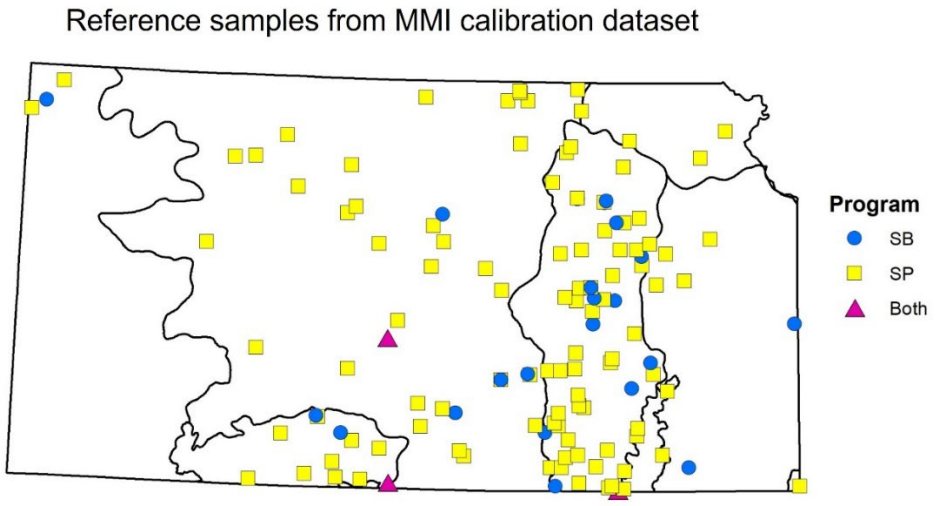


Figure 14. Sites that were used in the MMI calibration dataset, color-coded by program.

Some sites were sampled by both programs. At one site, both programs had collected a sample on the same day (Cottonwood River; April 20, 2010) and MMI scores were within 0.3 points of each other (SP = 63.1, SB = 62.8). Eleven sites were sampled by both programs during the same year but on different days. On average, SP MMI scores were 14 points lower, with differences ranging from 0.3 to 43 (Table 21). Appendix E has additional information on comparisons across the two programs.

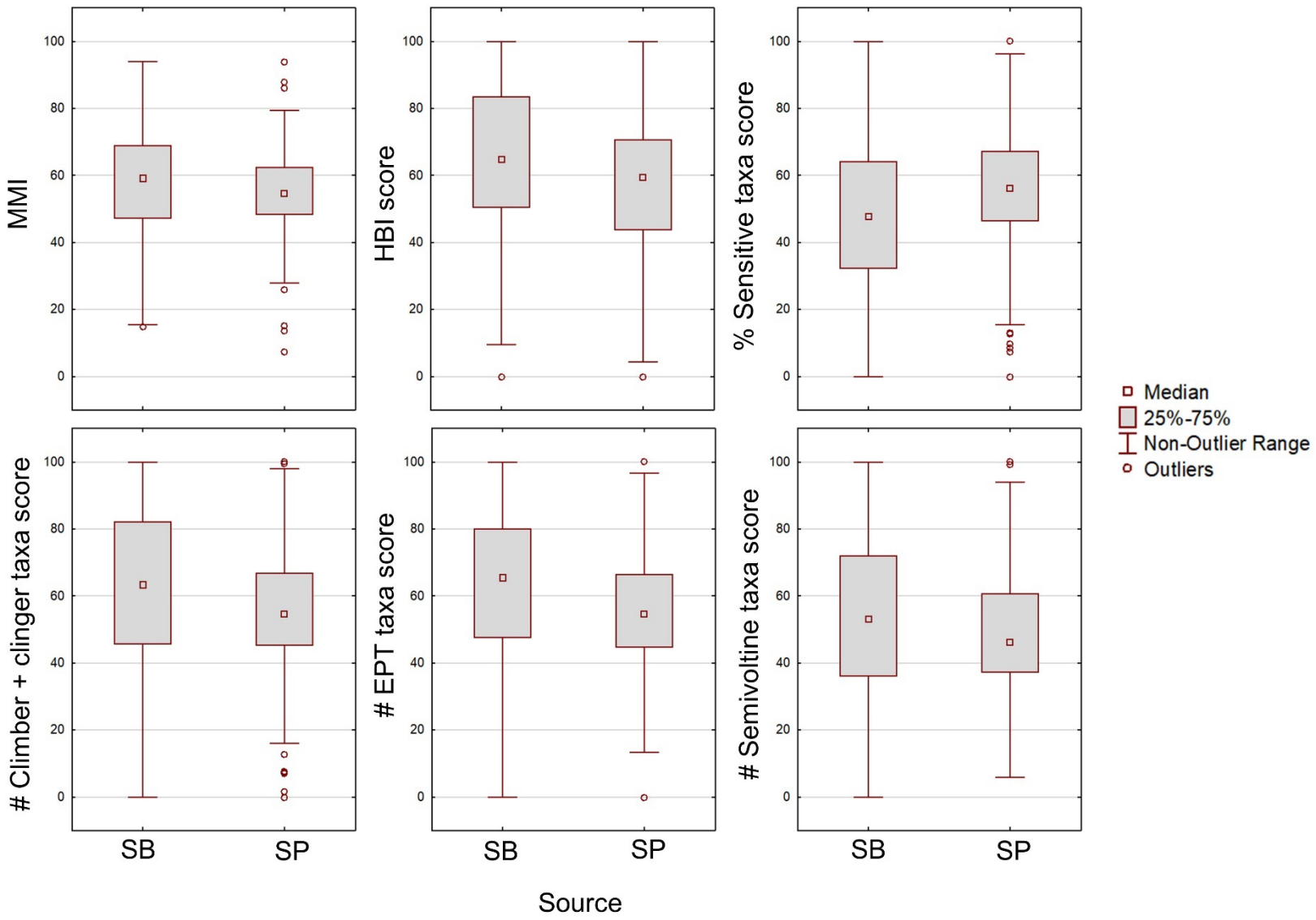


Figure 15. Box plots showing distributions of MMI and input metric scores for SB and SP program samples, based on all reference site samples (145 sites, 383 samples).

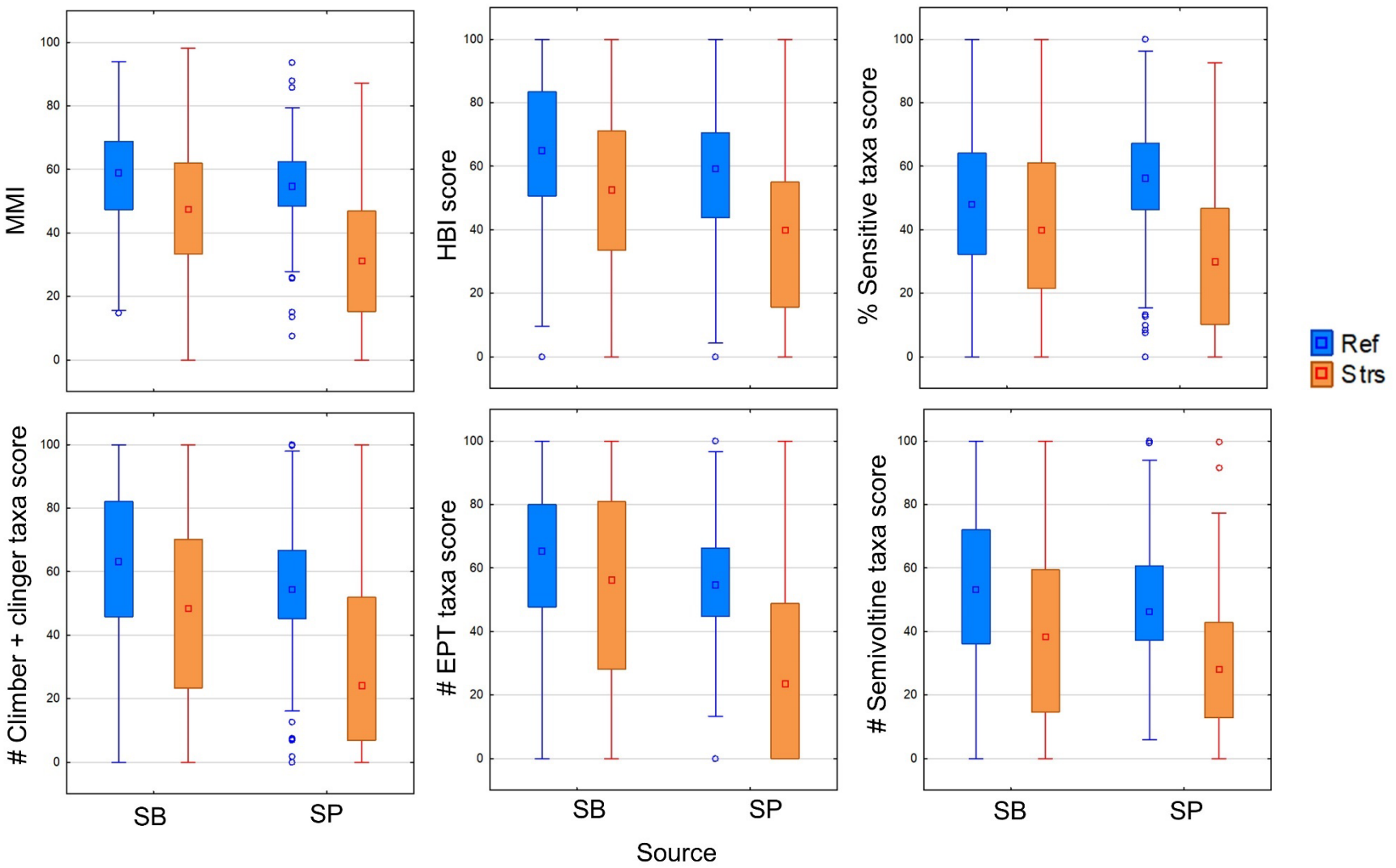


Figure 16. Box plots showing distributions of MMI and input metric scores for SB and SP program samples in reference and stressed samples (Ref dataset = 145 sites, 383 samples; Strs dataset =122 sites, 499 samples).

Table 21. Comparison of MMI scores at 11 sites that were sampled by both programs during the same year but on different days. Where available, same day replicates were also included (Rep1, Rep2).

Waterbody Name	Tt_StationID	Year	Difference (SP MMI – SB MMI)	SP				SB			
				KDHE SITE	Coll Date	MMI		KDHE SITE	Coll Date	MMI	
						Rep1	Rep2			Rep1	Rep2
Neosho River	KS_MMI_Site014	2014	-4.9	SPB260	21-May-14	33.8		SB214	17-Jul-14	38.7	
Neosho River	KS_MMI_Site316	2010	-6.8, -35.8	SPB036, SPB612	17-Aug-10	22.6		SB273	20-Apr-10	29.4	58.4
		2017	-24.1		19-Jun-17	48.4			13-Jul-17	72.4	
Spring River	KS_MMI_Site049	2012	-29.8	SP822	16-May-12	46.8		SB568	06-Jun-12	76.6	
		2015	-21.4		21-Sep-15	38.0	57.3		14-Oct-15	59.4	
Verdigris River	KS_MMI_Site080	2012	0.9	SP810	03-Jul-12	54.2		SB105	09-Aug-12	53.3	
		2015	-18.8		05-May-15	35.5			05-Aug-15	54.3	
		2017	-43.1, -37.6		18-Jul-17	16.6			19-Jul-17	59.6	54.1
Marais des Cygnes River	KS_MMI_Site290	2014	-23.2	SP809	15-Jul-14	47.4		SB745	31-Jul-14	70.6	
Arkansas River	KS_MMI_Site314	2017	5.6	SPA170, SPB643	08-Aug-17	68.9		SB284	06-Jul-17	63.3	
Cottonwood River	KS_MMI_Site315	2010	0.3	SPA387	20-Apr-10	63.1		SB274	20-Apr-10	62.8	
Kansas River	KS_MMI_Site551	2010	-20.6	SPB032	23-Sep-10	41.9		SB260	04-Oct-10	62.5	
Crooked Creek	KS_MMI_Site577	2007	-8.2	SPA150	05-Jul-07	56.7		SB683	14-Sep-07	64.9	
		2014	-7.7		07-May-14	49.6			23-Sep-14	57.3	
		2016	22.1, 4.4		27-Jun-16	74.3	56.6		16-Jun-16	52.2	
North Fork Solomon River	KS_MMI_Site609	2014	-30.2	SPB369	27-May-14	51.1		SB014	02-Jul-14	81.3	
Republican River	KS_MMI_Site710	2011	-7.2	SPB162	21-Jul-11	80.6		SB231	09-May-11	87.8	

5.3.2 Time

We examined relationships between MMI scores and Julian date, month and year in the reference dataset to see if any temporal patterns were evident. We looked at SB and SP program samples separately and limited the dataset to April 15-October 31, which spans the normal sampling periods for both programs (SP: April 15-September 30; SB: May 1-October 31[11]). Overall, MMI scores for the SB samples were lower at the beginning and end of the SB index period, with MMI scores decreasing noticeably in late October (Figure 17). Median MMI scores for SB samples were highest in August (64.6) and ranged from 55-60 during the other months (Figure 18). Temporal patterns were less evident in the SP samples.

We also looked for patterns in MMI scores at reference sites across all years (1980-2020). Most of the reference samples were collected by the SP program from 2006 onward (Figure 19). There were no clear trends over time. Rather, median MMI scores increased and decreased in cyclical patterns. Programmatic differences may account for some of the patterns (overall, SP reference samples had lower MMI scores; Section 5.3.1), along with weather events. For example, some of the lower MMI scores from 2011-2013 corresponded with drought years (Figure 19).

[11]The SB program stops sampling at leaf fall; since leaf fall dates vary from year to year and site to site, we used October 31 as the end date.

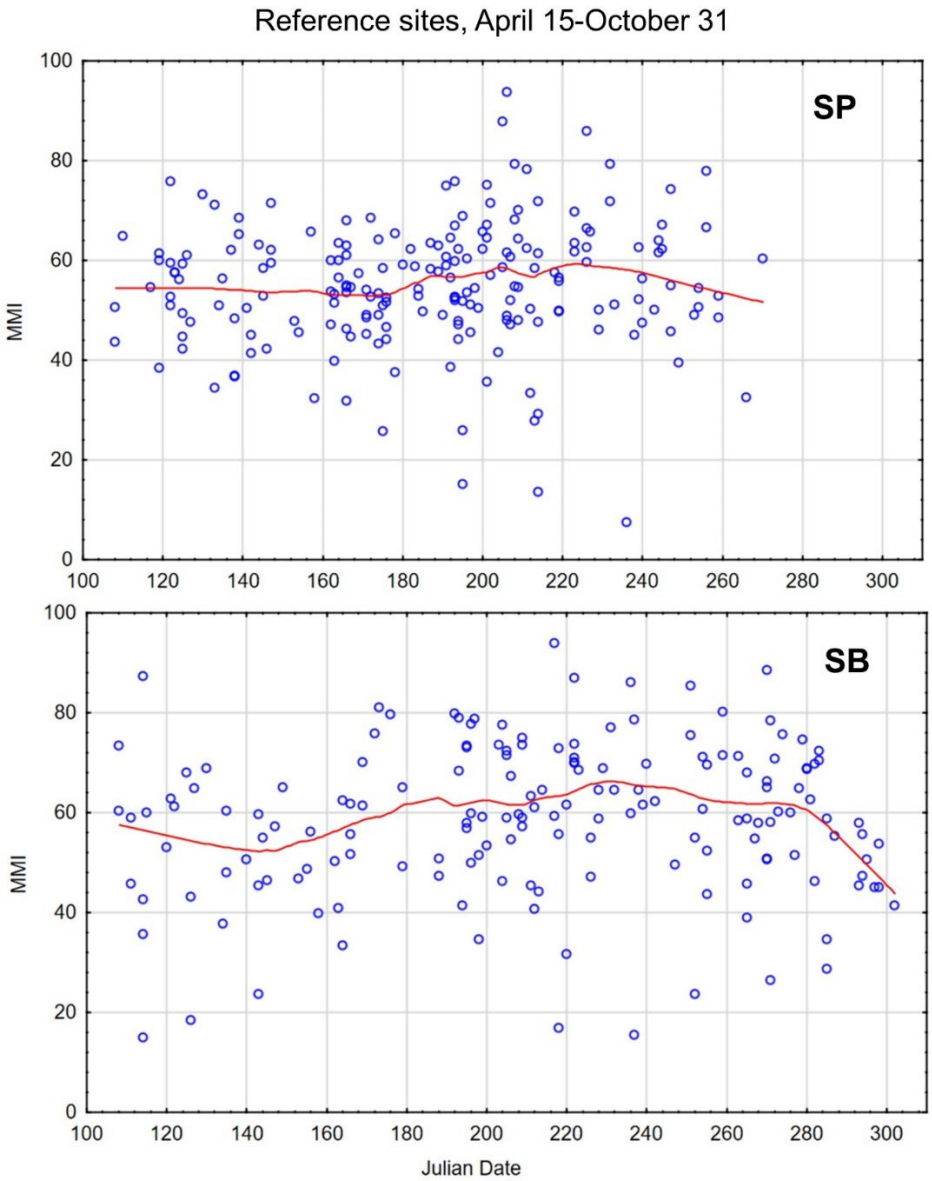


Figure 17. Scatterplots of MMI scores vs. Julian date for SP (top) and SB (bottom) reference samples, fit with a Lowess line. The date range was limited to April 15 through October 31. Sample sizes: SP = 200 samples from 118 sites; SB = 166 samples from 27 sites.

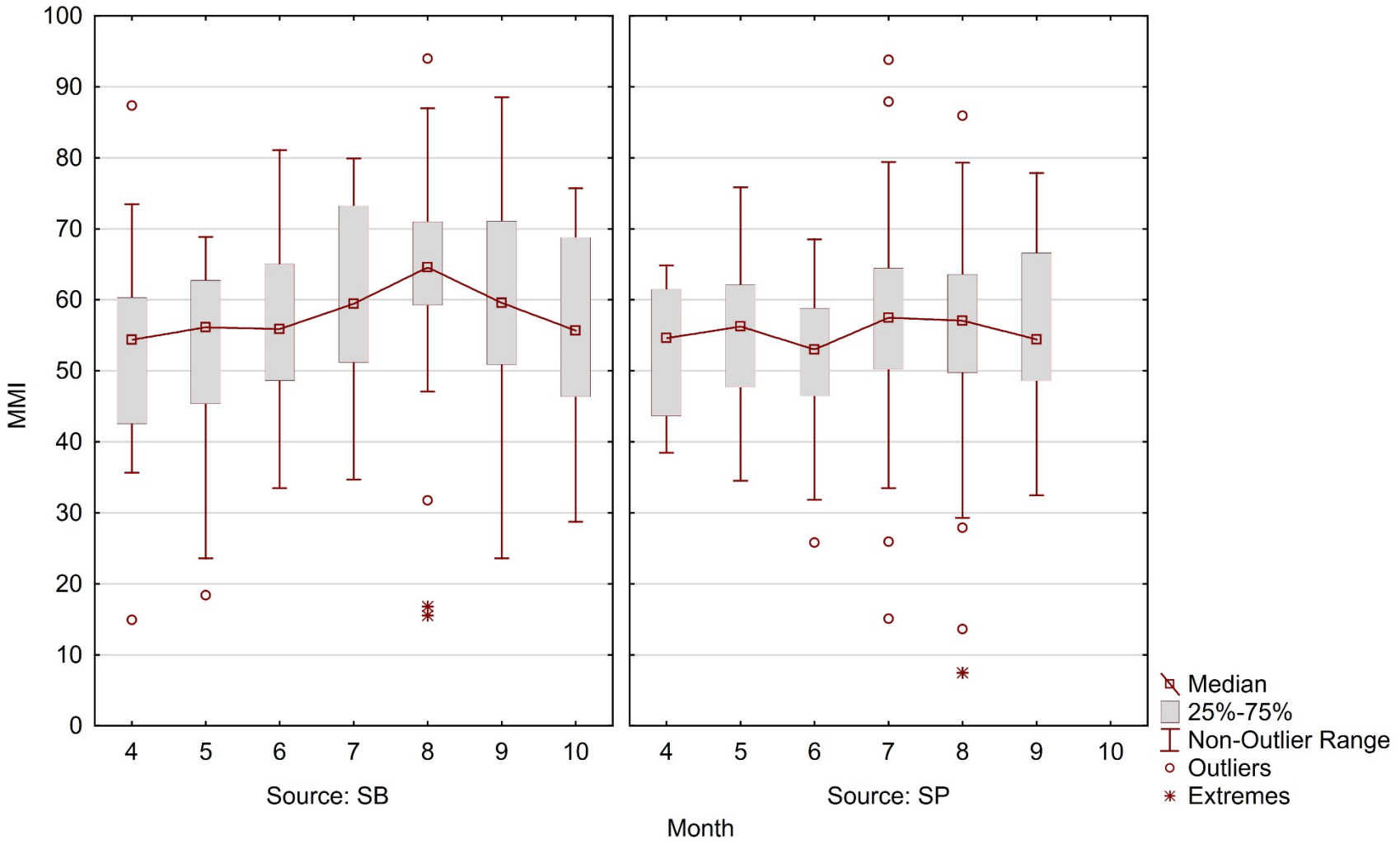


Figure 18. Box plots showing distributions of MMI scores vs. month for SB (left) and SP (right) reference samples. Sample sizes: SP = 200 samples from 118 sites; SB = 166 samples from 27 sites.

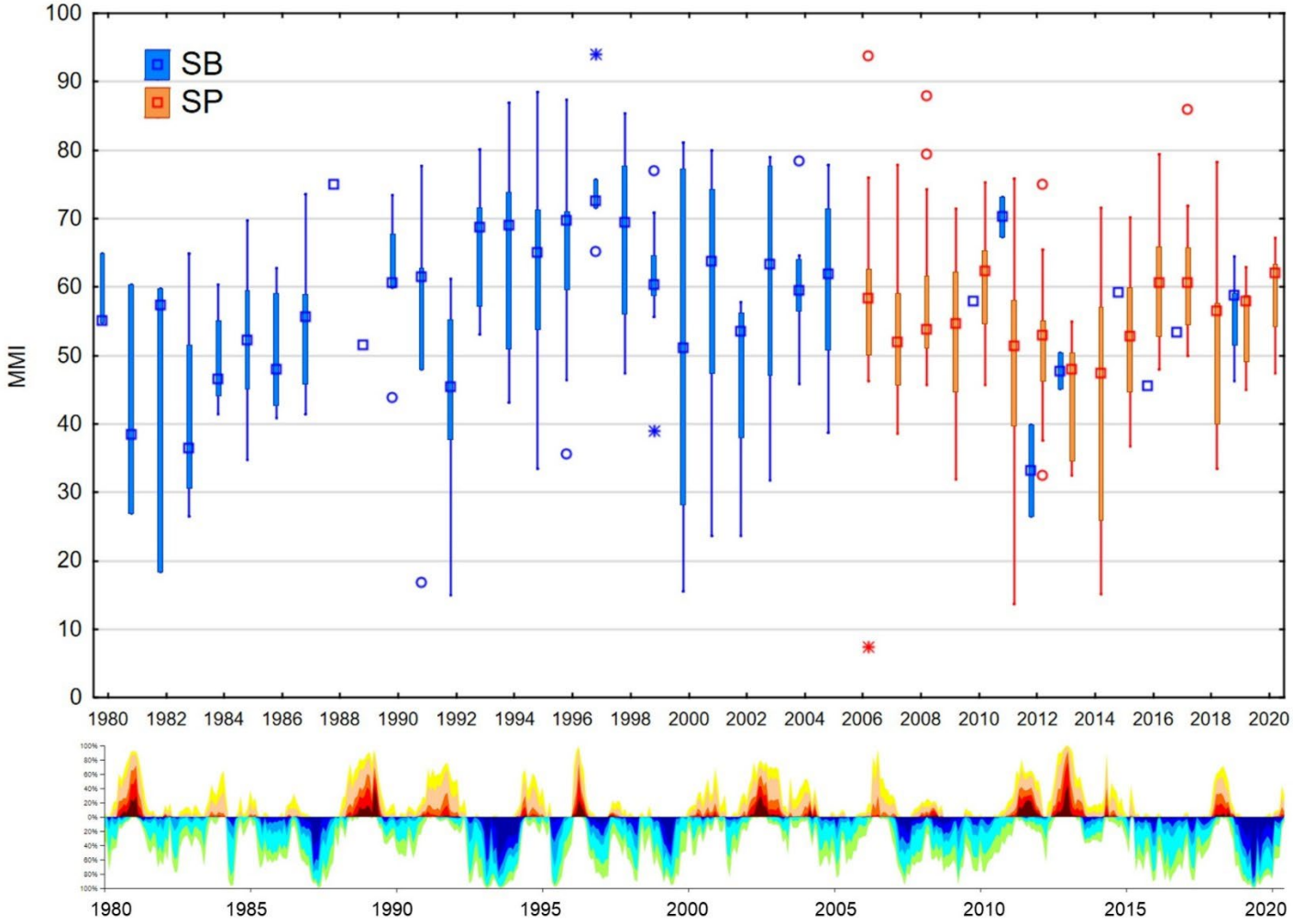


Figure 19. Box plots showing distributions of MMI scores vs. year for SB (blue) and SP (orange) reference samples. Sample sizes: SP = 200 samples from 118 sites; SB = 166 samples from 27 sites. Also shown is the Standardized Precipitation Index (SPI), which captures how observed precipitation (rain, hail, snow) deviates from the climatological average over the 9 months leading up each month. Red hues indicate drier conditions, while blue hues indicate wetter conditions[12]

[12] SPI: https://www.drought.gov/historical-information?dataset=1&selectedDateUSDM=20101221&selectedDateSpi=19580601

KDHE has a rich dataset for examining year-to-year variability at individual sites. Twenty-nine sites have 30-38 years of data, 13 sites have 20-29 years of data and 21 sites have 10-19 years of data. Figure 20 shows the locations of the long-term monitoring sites. We examined distributions of MMI scores at reference and stress sites with 10 or more years of data. Overall, both reference and stress sites had similar amounts of variability (Figures 21 & 22, respectively); the mean range of MMI scores in both datasets was 46 (Tables 22 & 23). In the reference dataset, the Caney River (SB217, SP811) had the most variable MMI scores, ranging from 13.6 to 77.6 over its 15 year period of record. Thompson Creek (SB354) was least variable, with MMI scores ranging from 59 to 80 (Table 22, Figure 21). In the stress dataset, the Wolf River SB363 site had the most variable MMI scores, ranging from 9 to 88 over its 18-year period of record. Indian Creek (SB204) was least variable, with MMI scores ranging from 0 to 12 (Table 23, Figure 22).

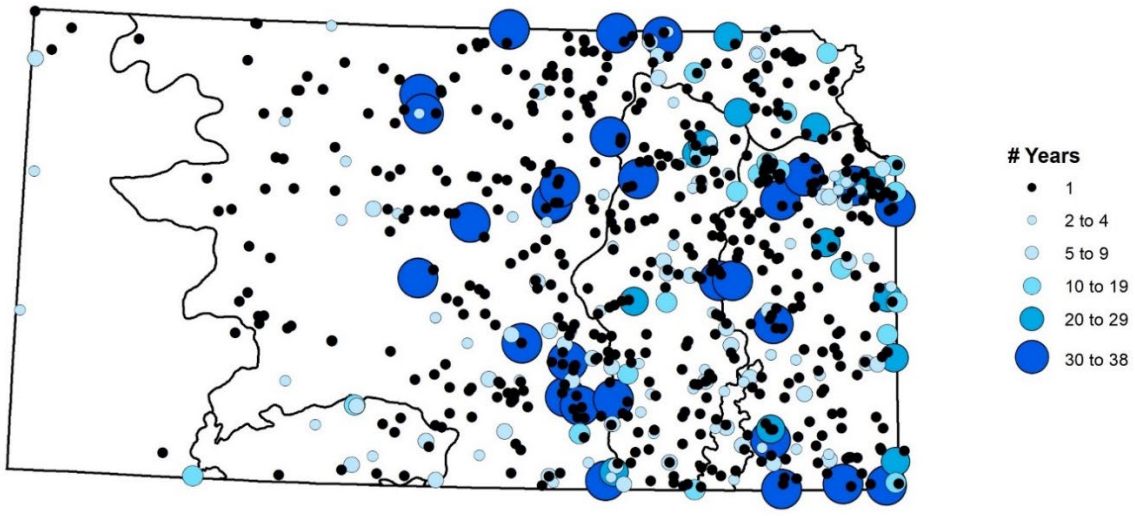


Figure 20. Sites in the KDHE macroinvertebrate dataset, color-coded and proportionally sized based on number of years of data.

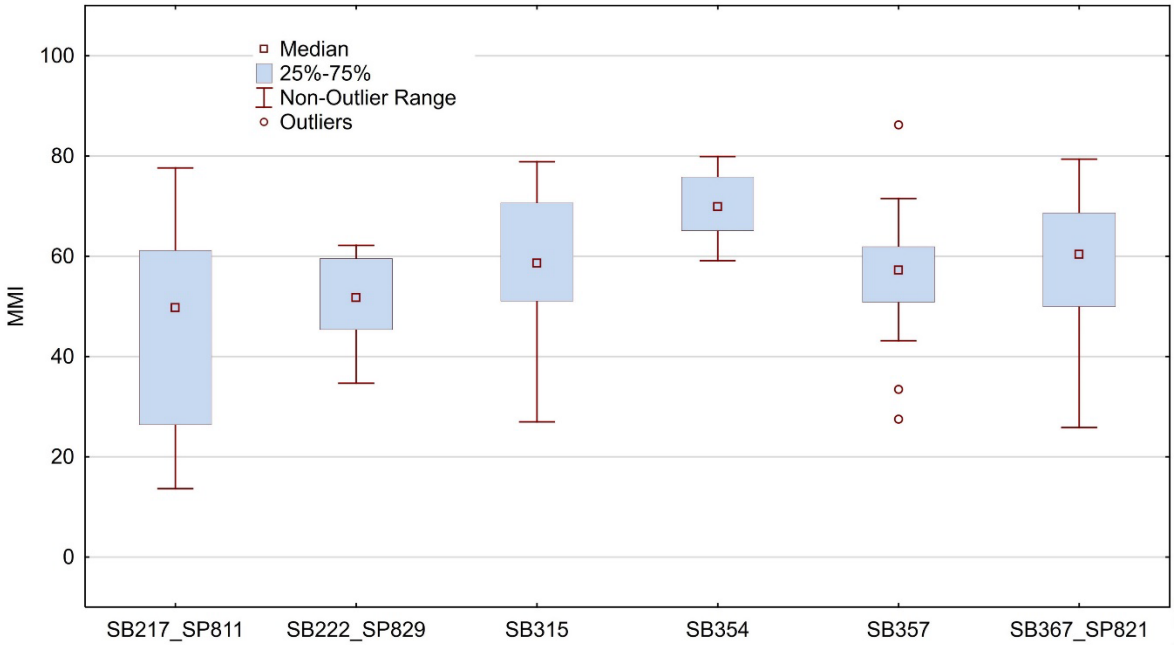


Figure 21. Box plots showing distributions of MMI scores at reference sites with 10 or more years of data. For more information on these sites, see Table 22.

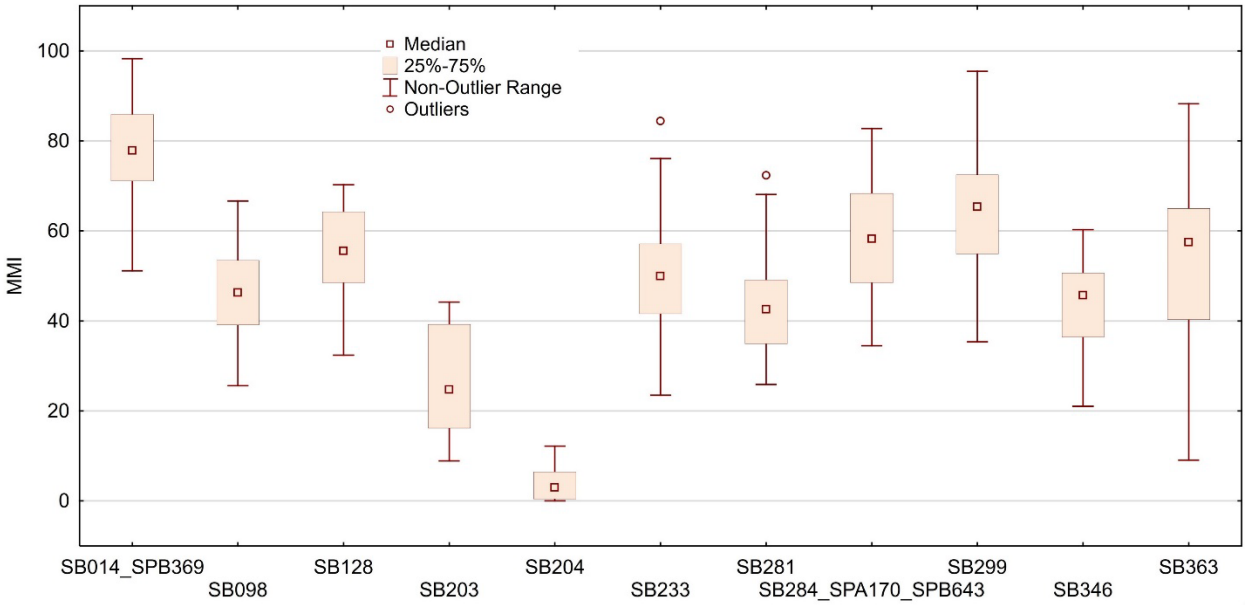


Figure 22. Box plots showing distributions of MMI scores at stress sites with 10 or more years of data. For more information on these sites, see Table 23.

Table 22. Reference sites with 10 or more years of data, with site information and MMI statistics. CGP = Central Great Plains.

Level 3 ecoregion	Median flow (cfs)	Latitude	Longitude	Site	Waterbody Name	Program	Min Year	Max Year	Num Years	Mean MMI	Min MMI	Max MMI	Range MMI
SW Tablelands	2.6	37.5165	-99.1653	SB354	Thompson Creek	SB	1990	2005	14	70.2	59.1	79.9	20.8
CGP	241.0	37.1119	-96.9866	SB315	Walnut River	SB	1980	2003	24	58.9	27.0	78.9	51.9
Flint Hills	20.0	38.2267	-96.8317	SB367, SP821	Cedar Creek	SB, SP	1990	2016	21	58.3	25.8	79.4	53.6
Flint Hills	20.1	38.2250	-96.5608	SB357	South Fork Cottonwood River	SB	1985	2001	14	56.5	27.5	86.2	58.7
High Plains	43.2	37.0113	-100.4917	SB222, SP829	Cimarron River	SB, SP	1980	2016	13	50.9	34.7	62.1	27.5
Cross Timbers	65.2	37.0036	-96.3161	SB217, SP811	Caney River	SB, SP	1982	2016	15	45.6	13.6	77.6	64.0
									Mean	***56.8***	***31.3***	***77.4***	***46.1***

Table 23. Stress sites with 10 or more years of data, with site information and MMI statistics. CGP = Central Great Plains, CIP = Central Irregular Plains, WCBP = Western Corn Belt Plains.

Level 3 ecoregion	Median flow (cfs)	Latitude	Longitude	Site	Waterbody Name	Program	Min Year	Max Year	Num Years	Mean MMI	Min MMI	Max MMI	Range MMI
CGP	80	39.5541	-98.6921	SB014, SPB369	North Fork Solomon	SB, SP	1980	2017	36	77.8	51.1	98.3	47.1
WCBP	5.6	39.4631	-95.9500	SB299	Soldier Creek	SB	1985	2020	25	65.5	35.4	95.5	60.1
CGP	185	38.3556	-98.6637	SB284, SPA170, SPB643	Arkansas River	SB, SP	1980	2017	34	57.7	34.5	82.8	48.3
WCBP	7.63	39.7258	-96.3287	SB128	North Fork Black Vermillion R	SB	1996	2017	13	55.0	32.4	70.3	37.9
WCBP	41.25	39.8494	-95.1802	SB363	Wolf River	SB	1982	2017	18	52.9	9.1	88.2	79.2
WCBP	229	39.9577	-96.6097	SB233	Big Blue River	SB	1980	2018	38	50.5	23.5	84.4	60.9
CIP	508	38.0836	-95.6560	SB098	Neosho River	SB	1980	2018	33	46.4	25.6	66.7	41.1
CGP	11.1	37.5996	-97.4044	SB346	Cowskin Creek	SB	1980	2019	38	44.1	21.0	60.3	39.2
CGP	673	37.5429	-97.2758	SB281	Arkansas River	SB	1980	2019	35	43.2	25.9	72.4	46.5
CIP	4910	39.1114	-94.6163	SB203	Kansas River	SB	1980	2015	12	26.3	8.9	44.2	35.3
CIP	21.11	38.9384	-94.6081	SB204	Indian Creek	SB	1980	2016	15	3.7	0.0	12.2	12.2
									Mean	***47.6***	***24.3***	***70.5***	***46.2***

5.3.3 Stream size

We used scatterplots to examine relationships between MMI scores, reference status and stream size, as measured by modeled median flow (Perry et al. 2004) and total watershed area. Patterns related to stream size were not evident in the reference samples. Stressed samples had noticeably higher MMI scores when modeled median flows were in the ~50-500 cfs range (Figure 23), with some sites receiving higher MMI scores than the top-scoring reference sites (>80). The same pattern occurred with total watershed area, with some of the stressed samples having noticeably higher MMI scores in the ~5,000 to 50,000 km2 range (Figure 24). A list of the stressed sites that had MMI scores > 60 in the MMI calibration dataset and median flow statistics ranging from 50-500 cfs can be found in Table 24.

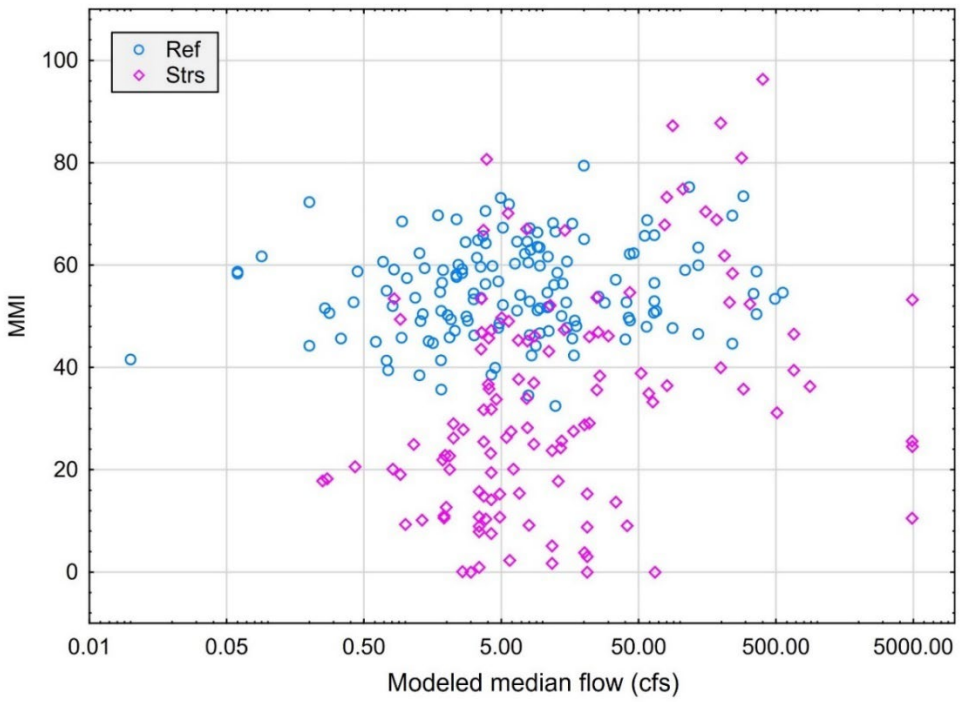


Figure 23. Scatterplot of MMI scores vs. modeled median flow (cfs; Perry et al. 2004), based on the reference and stressed samples in the MMI calibration dataset (Ref = 145; Strs = 122). The x-axis is log-transformed.

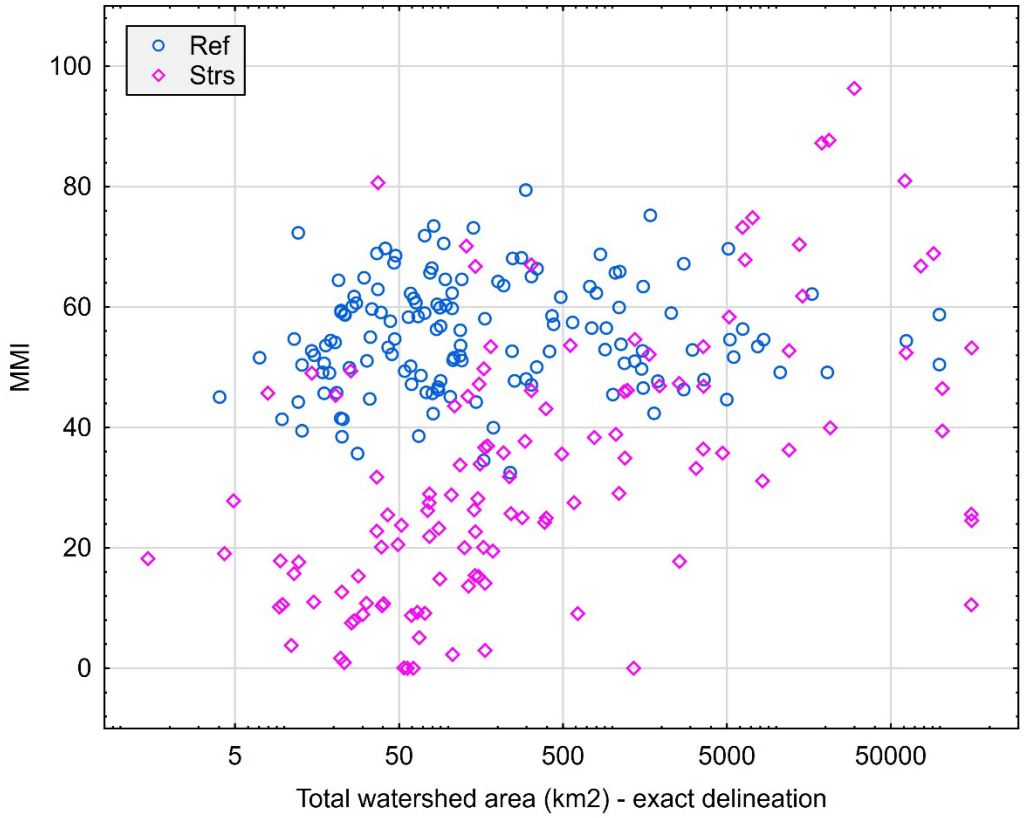


Figure 24. Scatterplot of MMI scores vs. total watershed area (km^2; based on exact watershed delineations) for reference and stressed samples in the MMI calibration dataset (Ref = 145; Strs = 122). The x-axis is log-transformed.

Table 24. Stressed sites that had MMI scores > 60 in the MMI calibration dataset and median flow statistics ranging from 50-500 cfs. CGP=Central Great Plains. Sites with blue shading had MMI scores > 80. Also shown are mean (min-max) MMI scores for sites with multiple years of data.

Latitude	Longitude	SITE	Waterbody Name	Coll Date	Source	Level 3 ecoregion	Median flow (cfs)	MMI	# Years of data	Mean MMI (Min-Max), all years of data
38.8636	-97.4833	SB265	Smoky Hill River	09-Aug-10	SB	CGP	400	96.3	1	--
38.7404	-97.5807	SB514	Smoky Hill River	31-Jul-96	SB	CGP	198	87.7	4	87.7 (81.8-93.6)
38.7823	-98.4119	SPA142	Smoky Hill R	05-Sep-07	SP	CGP	88.54	87.2	1	--
39.5891	-97.6583	SB003	Republican River	27-Jun-17	SB	CGP	281	81.0	9	63.1 (44.2-81)
39.0065	-97.9286	SPB006	Saline R	18-Aug-10	SP	CGP	105	74.8	1	--
39.5541	-98.6921	SB014	North Fork Solomon River	20-Sep-17	SB	CGP	80	73.3	37	78.6 (61.8-98.3)
39.4637	-98.1967	SPB210	Solomon R	21-May-12	SP	CGP	154	70.4	1	--
38.3557	-98.6638	SPB643	Arkansas River	08-Aug-17	SP	CGP	185	68.9	1	--
39.0252	-98.1887	SPA069	Saline R	07-Jul-06	SP	CGP	77.5	67.9	1	--
39.4093	-98.0234	SPB258	Solomon R	05-Aug-13	SP	CGP	210	61.9	1	--

5.3.4 Habitat type (riffle-run vs. glide-pool)

We used box plots to compare distributions of MMI scores in Riffle-Run vs Glide-Pool sites, as designated by the SP field staff during Rapid Habitat Assessment (RHA) surveys (Barbour et al. 1999). The analysis was limited to MMI calibration samples. Patterns related to habitat type were not evident in the reference samples (Figures 25). There was, however, a noticeable pattern in the stressed samples. The median MMI scores of Glide-Pool stressed vs reference sites were very close (53.4 vs. 55.7, respectively), with some of the Glide-Pool stressed sites scoring higher than the top-scoring reference sites. The MMI was more effective at discriminating between reference and stressed Riffle-Run sites vs. Glide-Pool sites (Figure 25).

As a next step, we examined whether the Glide-Pool stressed sites with high MMI scores had modeled median flows in the ~50-500 cfs range (Section 5.3.3). We found overlap; five of the seven stressed sites listed in Table 24 were Glide-Pool sites. We then explored differences in substrate composition between reference and stressed Glide-Pool sites. Substrates at Glide-Pool sites consist primarily of fines (silt, clay, muck), sand (>0.06-2 mm) and small gravel (>2-16 mm). We found that stressed Glide-Pool sites had noticeably lower percent fines when modeled median flow reached 50 cfs (≤ 20% vs. 40% or higher) (Figure 26). This makes sense, given how higher flows will flush fine sediments downstream. The lower percent fines at higher CFS Glide-Pool sites might partly account for the better MMI scores, but further investigation is warranted.

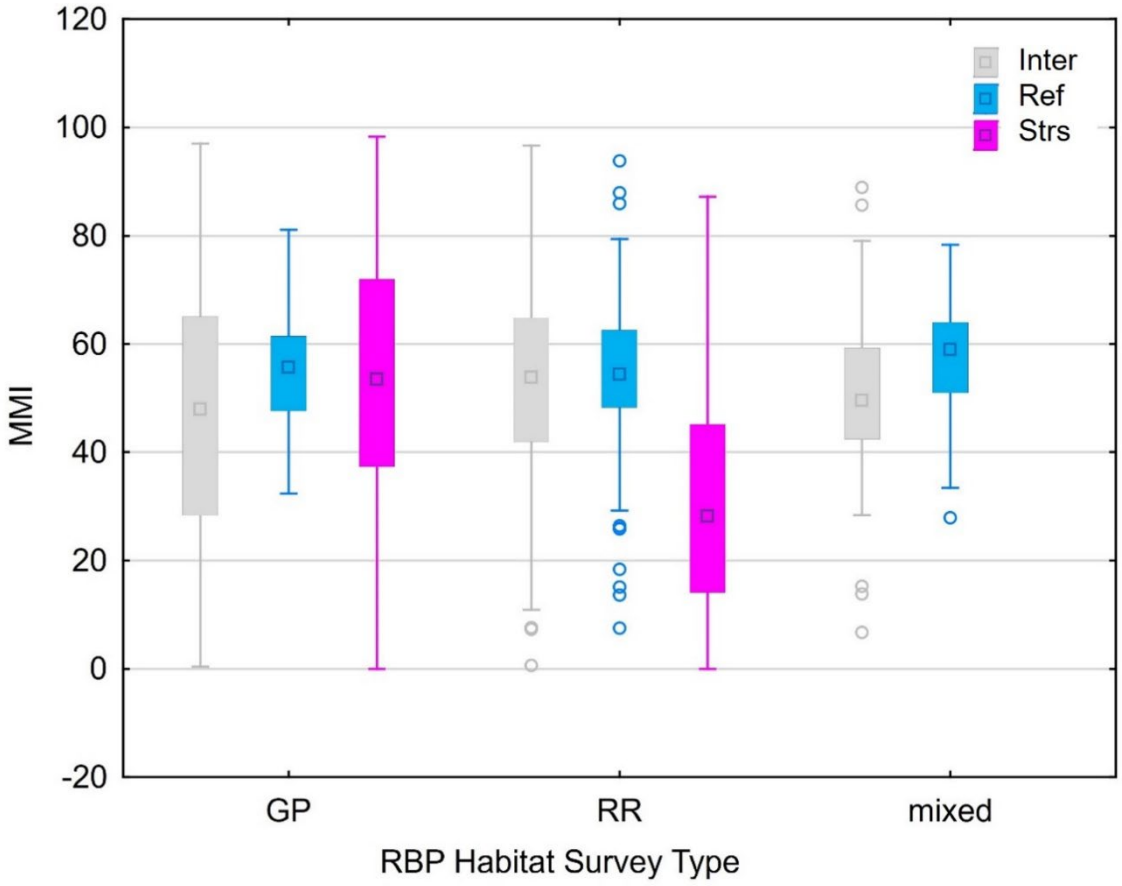


Figure 25. Box plot showing distributions of MMI scores at Riffle-Run (RR), Glide-Pool (GP) and mixed sites, grouped by disturbance category (reference, stressed, intermediate). Habitat type was designated by the SP field staff during Rapid Habitat Assessment (RHA) surveys. Mixed sites share characteristics of both RR and GP systems and were assessed as both RR and GP sites over time. This plot is based on MMI calibration samples.

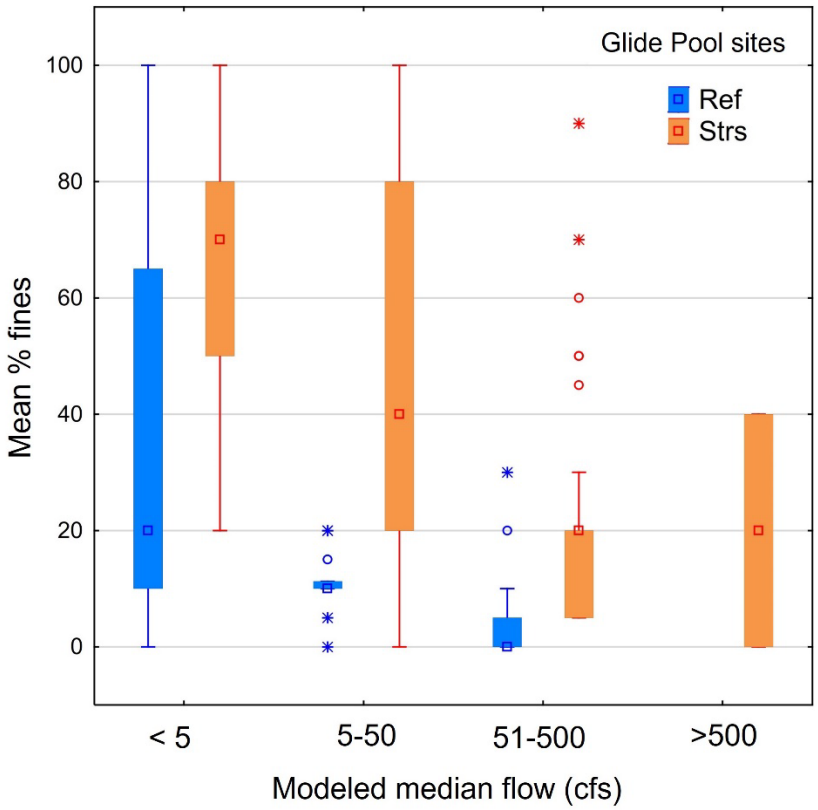


Figure 26. Box plot showing distributions of mean % fines at reference and stressed Glide-Pool sites, grouped by median flow category (cfs).

5.3.5 Level 3 ecoregion

Figure 27 shows all the BMI sampling sites color-coded by mean MMI scores, overlaid on Omernik Level 3 ecoregions. Sites with the highest MMI scores were concentrated in central Kansas. The worst-scoring sites were in eastern Kansas, with clusters around Kansas City and the southeastern portion of the Central Irregular Plains (CIP) ecoregion. We generated box plots to examine distributions of MMI scores at reference sites in each region, based on the MMI calibration dataset. All eight of the Level 3 ecoregions had at least one reference site, but most had small sample sizes (< 10 reference sites, except for the Flint Hills and Central Great Plains ecoregions). Overall, distributions of MMI scores were similar at reference sites across the ecoregions, with median scores mostly ranging from 55-60. The Western Corn Belt Plains was an exception, with a median reference site MMI score of ~50 (Figure 28). Where stressed sites were available[13], we also examined how well the MMI discriminated between reference and stress sites. The MMI was best at discriminating between reference and stress sites in the Flint Hills

[13]Three ecoregions – the Ozark Highlands, Cross Timbers and Southwestern Tablelands - did not have sites in the stressed (High Stress) dataset.

and Central Irregular Plains and least effective in the Central Great Plains and Western Corn Belt Plains ecoregions (Figure 28). The Central Great Plains covers the largest area and had the most number of sites (n=247), including 30 higher flow (>50 cfs) Glide-Pool streams. The stressed sites in the MMI calibration dataset that received high MMI scores (> 60) and had median flow statistics ranging from 50-500 cfs were all located in the Central Great Plains (Table 24). The Western Corn Belt Plains had a mix of Glide-Pool (n=14) and Riffle-Run (n=21) sites, with mostly lower flows (CFS < 50).

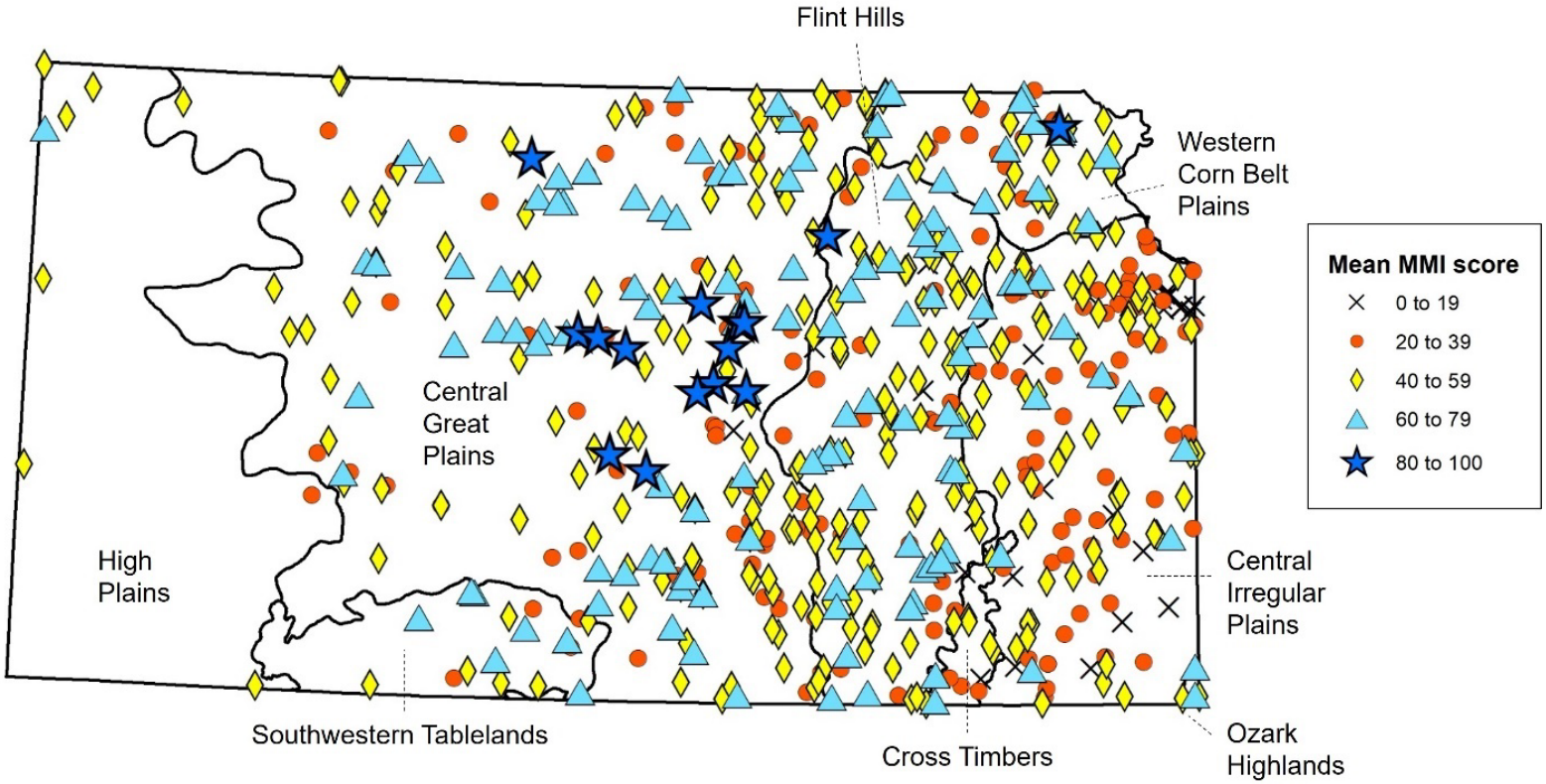


Figure 27. BMI sampling sites color-coded by mean MMI scores, overlaid on Omernik Level 3 ecoregions (delineated with black lines).

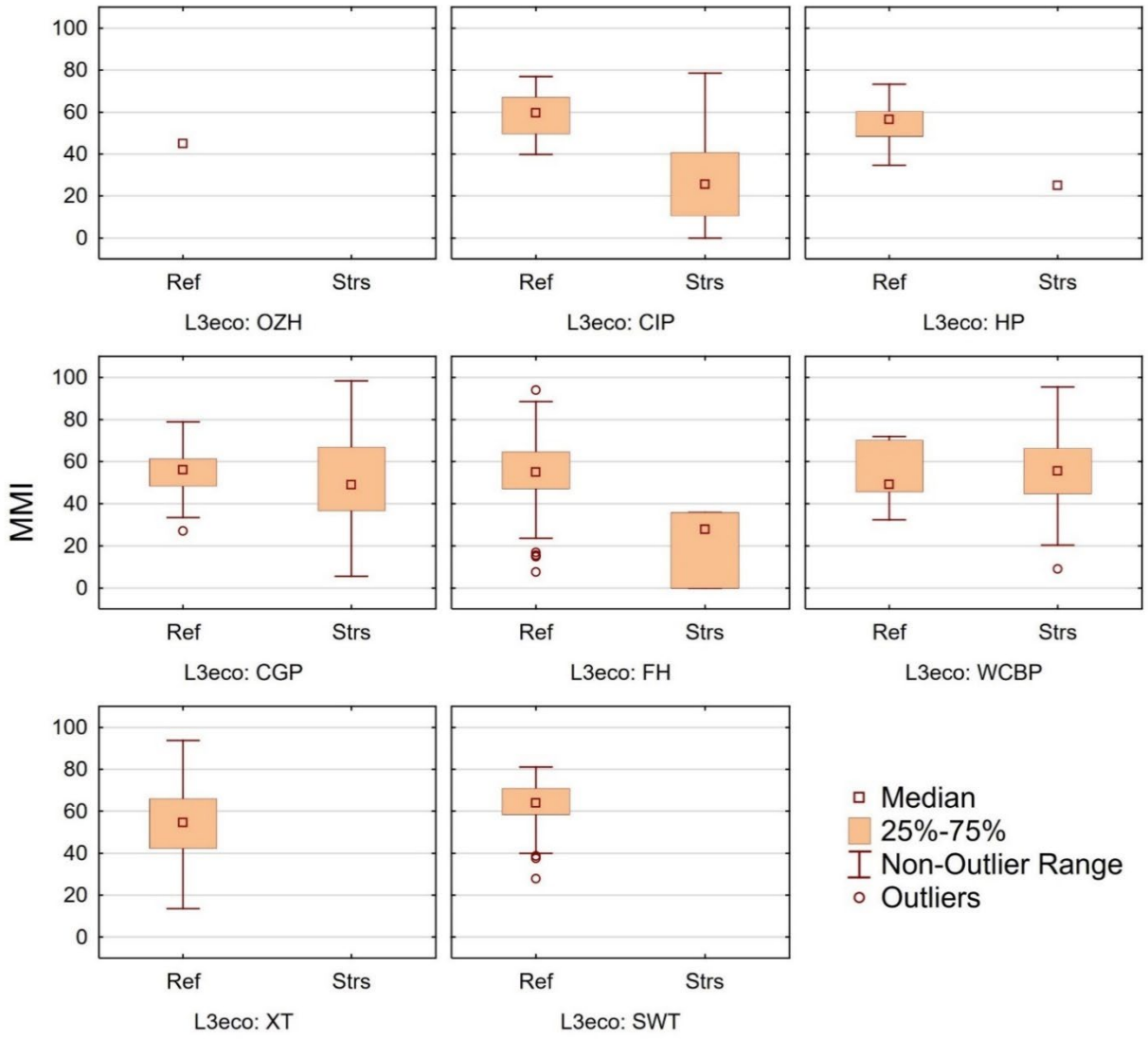


Figure 28. Box plots showing distributions of MMI scores in reference and stressed samples, grouped by Omernik Level 3 ecoregion, based on the MMI calibration dataset. CGP=Central Great Plains, CIP=Central Irregular Plains, FH=Flint Hills, HP=High Plains, OZH=Ozark Highlands, SWT=Southwestern Tablelands, WCBP=Western Corn Belt Plains, XT=Cross Timbers

5.3.6 Hydrologic Unit Code (HUC)

We generated box plots to examine distributions of MMI scores at reference sites in each Hydrologic Unit Code (HUC) 6 basin (Figure 29), based on the MMI calibration dataset. Of the 12 HUC6 basins, all but the Lower Missouri-Blackwater (which includes Kansas City) had at least one reference site. The Middle Arkansas had the most number of reference sites (n=31). Overall, distributions of MMI scores were similar at reference sites across the HUC6s with median scores mostly ranging from 50-60 (Figure 30). The MMI was effective at discriminating between reference and stress sites (where available) in most basins, with the best separation (least overlap between the 25th percentile of reference and 75th percentile of stress) occuring in the Kansas, Osage and Neosho basins. The most noticeable exceptions were the Smoky Hill and Missouri-Nishnabotna HUC6 basins (Figure 30). Median MMI scores at stressed sites in the Smoky Hill basin were slightly higher than median scores at the reference sites, with MMI scores at stressed sites reaching as high as 96.

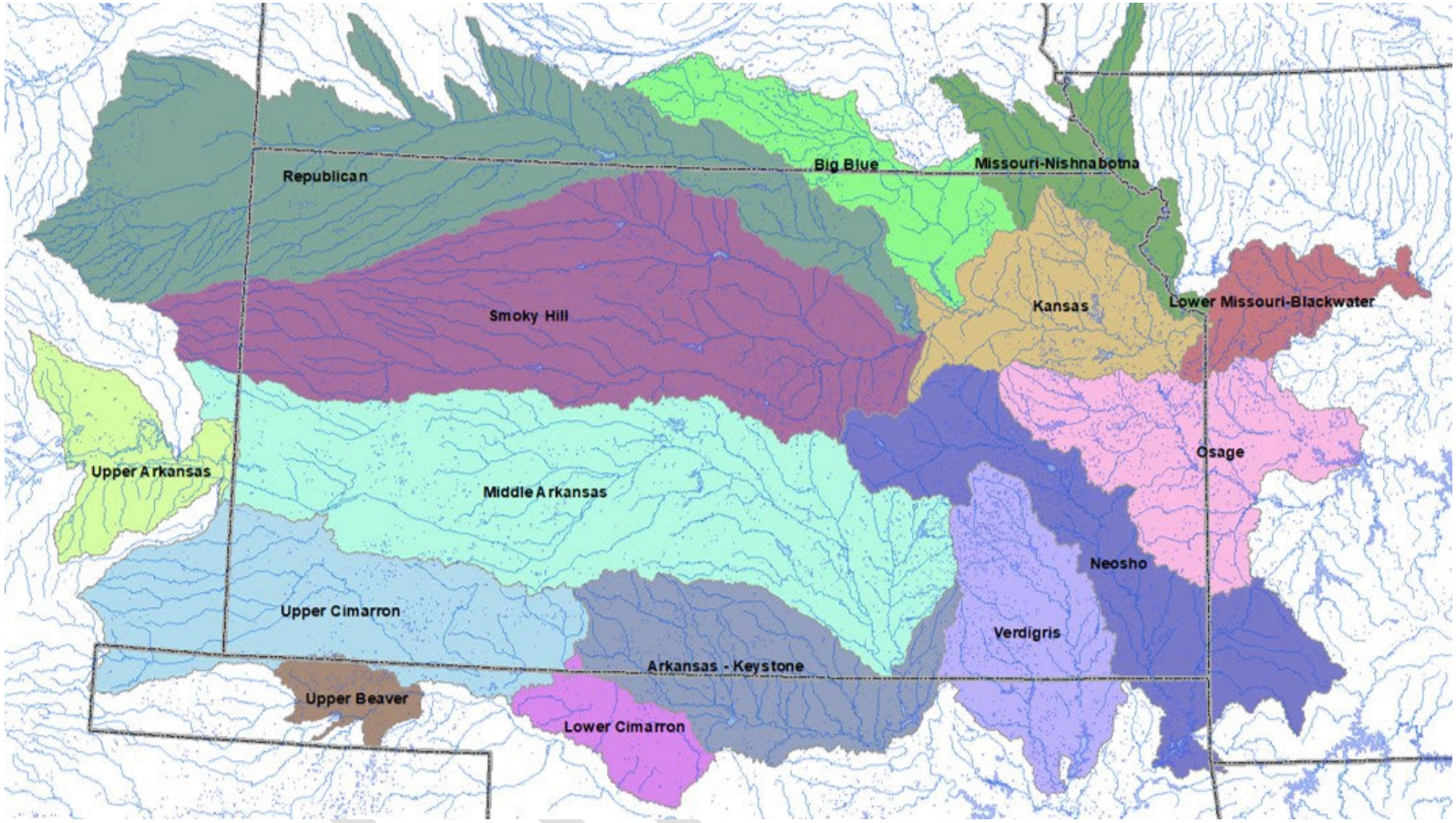


Figure 29. Hydrologic Unit Code (HUC) 6 basins in Kansas.

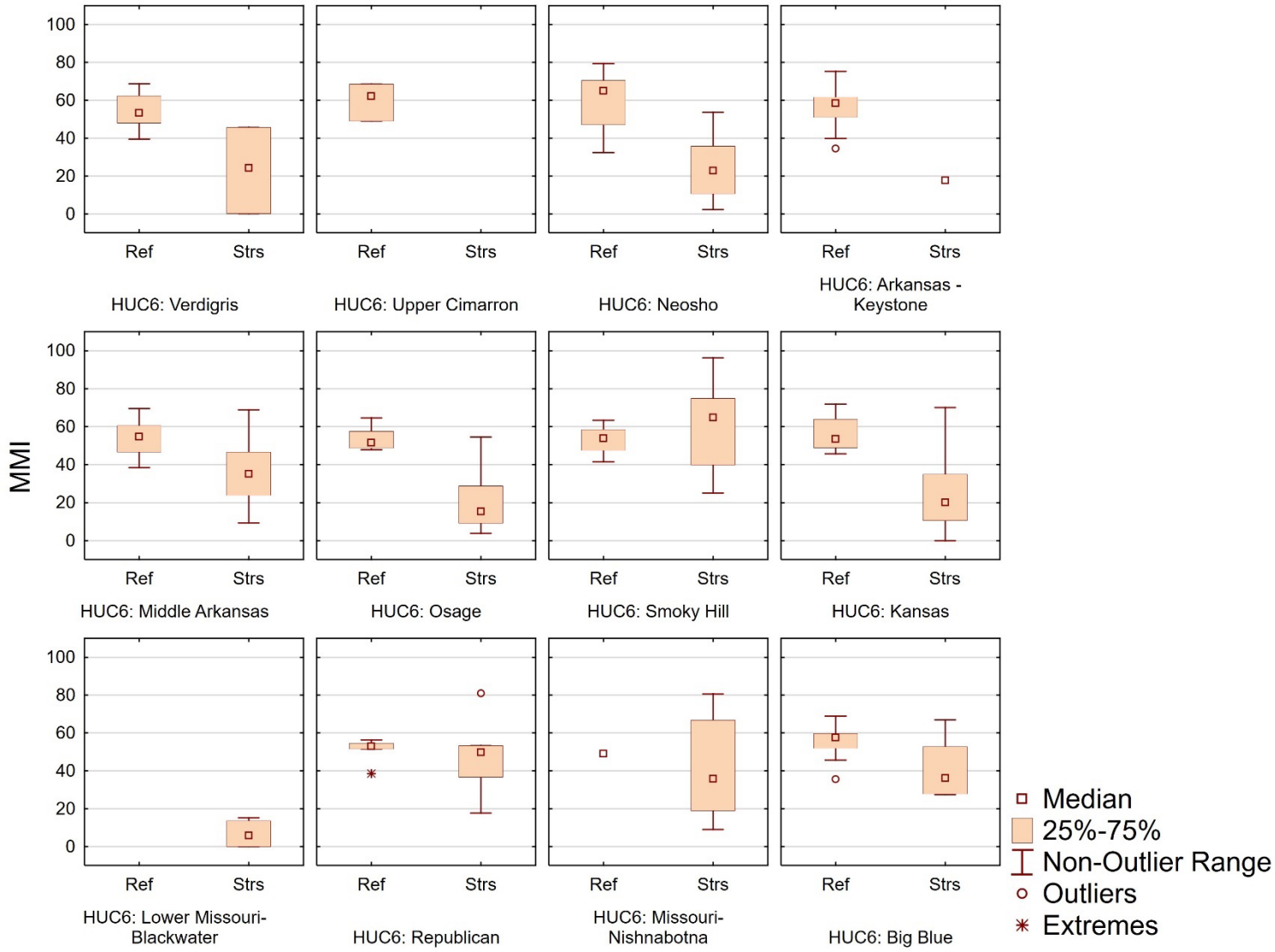


Figure 30. Box plots showing distributions of MMI scores in reference and stressed samples, grouped by HUC6, based on the MMI calibration dataset.

5.3.7 Specific conductance

In some areas, Kansas has streams with naturally high conductivity due to natural geological influences by known underlying salt deposits. These are reflected in Kansas's EPA-approved Surface Water Quality Standards for Natural Background Concentrations of chloride, sulfate, selenium and flouride (Table 1h, KDHE 2022), which influence specific conductance concentrations. The dominant chloride/sulfate ion (based on specific conductance) are:

- Specific conductance 0 – 500 µmohs/cm (85% sulfate, 15% chloride ion dominant)
- Specific conductance 500 – 1000 µmohs/cm (80% sulfate, 20% chloride ion dominant)
- Specific conductance 1000 – 1500 µmohs/cm (75% sulfate, 25% chloride ion dominant)
- Specific conductance 1500 – 5000 µmohs/cm (43% sulfate, 57% chloride ion dominant)
- Specific conductance 5000 – 11, 500 µmohs/cm (0% sulfate, 100% chloride dominant)

Areas with high observed specific conductance are shown in Figure 31. High concentrations can be driven by both natural and anthropogenic factors and it can be challenging to apportion contributions across sources. Human influences include inputs from point and non-point source discharges, water use and irrigation returns, disturbance of vegetative cover, soils, and geology and changes to evapotranspiration and the water table leading to mobilization of deep mineral deposits. We explored using Predicted Background Conductivity Data from Olson & Cormier (2019)[14] to tease out areas affected by anthropogenic influences but found that the modeled values did not match closely enough with observed values to use with the predicted data with confidence.

Because of food web and osmoregulation implications (e.g., standard 860 mg/L chloride for acute aquatic life), increasing chloride (and specific conductance) concentrations are likely a driver of loss in species richness and abundance. We explored whether this was evident in the BMI data by examining the relationship between MMI scores and median and maximum specific conductance at reference sites in the MMI calibration dataset. In particular, we were interested in the MMI response once concentrations reached 1500 µmohs/cm (when chloride comprises >50% of the ions) and 5000 umohs/cm (when chloride comprises 100% of the ions). There were limited numbers of reference sites in these higher ranges (10 sites in the 1501-5000 range and four sites > 5000). We examined scatterplots and Spearman correlation analyses and did not find significant relationships ($p < 0.05$) with the MMI, nor with the random forest adjusted input metrics (Figure 32).

[14]Interactive map: U.S. EPA Freshwater Explorer - Overview (arcgis.com)

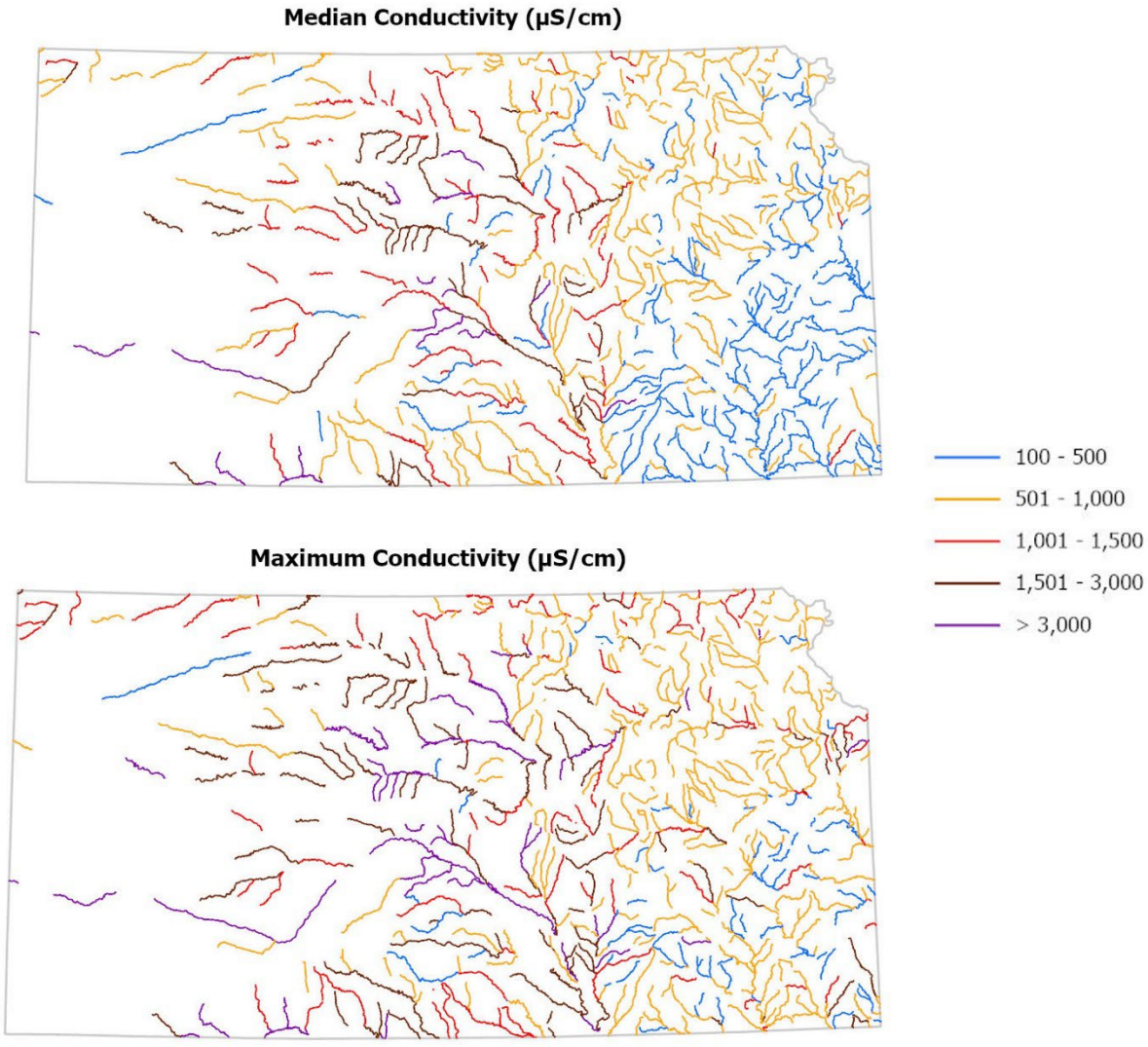


Figure 31. Observed median and maximum specific conductance concentrations (µs/cm), based on all available samples for channel unit segment (CUSEGA) reaches. Maps were provided by Katlynn Decker (KDHE).

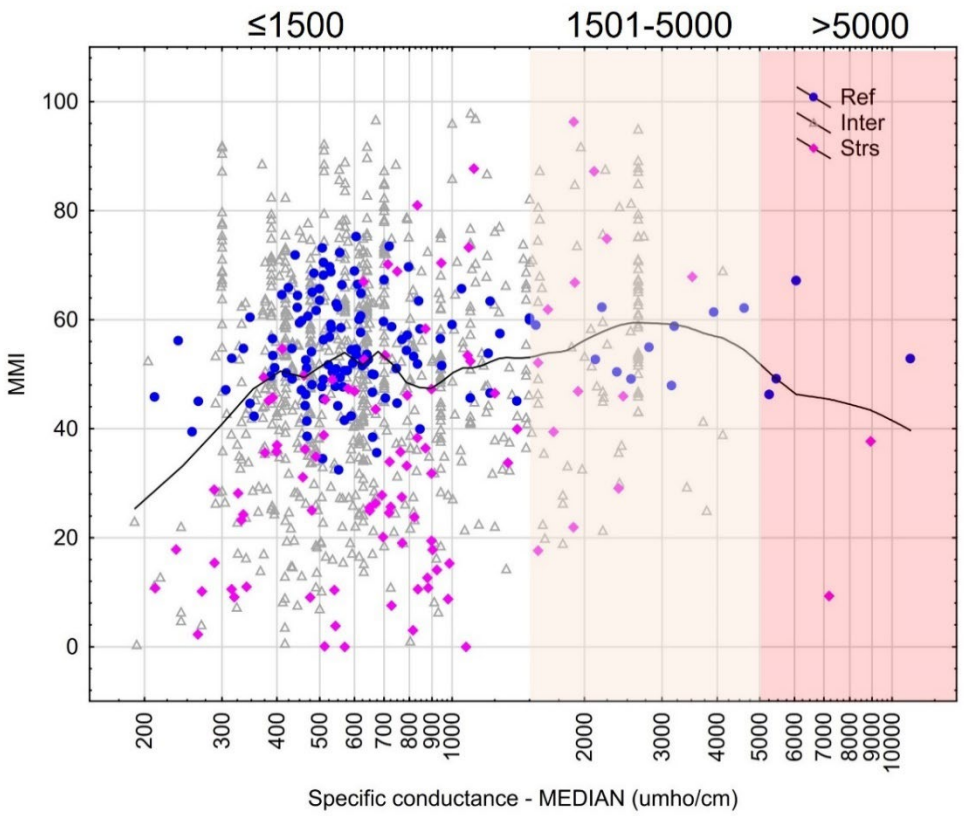


Figure 32. Scatterplot of MMI scores vs. observed median specific conductance (based on all available grab samples per site), based on the MMI calibration dataset and fit with a Lowess line. The x-axis is log-transformed. Conductivity ranges are color-coded at ≤ 1500, 1501-5000 and >5000 µmohs/cm. 1500 and 5000 correspond to thresholds at which chloride comprises >50% and 100% of the dominant ion.

5.4 Anomalous samples

To help assess performance of the MMI, we did a targeted search for 'anomalous' samples, which had disturbance designations and MMI scores that were furthest out of agreement. It also served as a check of the disturbance index. We defined 'anomalous' as reference samples that had MMI scores < 35 ('underachievers') and stressed samples with MMI scores > 80 ('overachievers'). Figure 33 shows the locations of the anomalous samples. For each of the samples, we compiled information on the BMI community (MMI scores and input metrics, percent air breather individuals, number of total taxa and individuals, etc.), and sample and site information. This information can be found in Attachment 4. We then went through the following checklist to determine whether the sample had any unusual characteristics that might account for the discrepancy:

- Was it collected outside the normal sampling period?
- Was it collected before 2000? (most of the KS MMI calibration samples were post-2000)
- Were there changes in field staff and/or taxonomists? Level of taxonomic resolution?
- Were there any notes on the field form about beaver activity, pooling (drying), portions of the sampling reach being too deep to sample, or fish? (fish can impact BMI communities in headwater streams in particular)

- If there were multiple years of data, how variable were the MMI scores? Was the low or high score a single occurrence or did it occur consistently over time?
- Was the sample likely affected by an extreme weather event (drought or flooding)?
    - What was the Standardized Precipitation Index (SPI) during the month/year of the sampling event? What was it for the year prior? (to check for lag year effects)
    - What was the stream size? (smaller headwater streams are more prone to drying whereas larger streams are likely more resilient)
    - Were percent air breather individuals higher than normal? (higher values (≥ 25%) indicate more potential for disturbance such as drought)
    - What was the predominant substrate? (soft, unstable sandy substrate is more vulnerable to scour from high flow events)
- Was specific conductance high (> 1500 umohs/cm)? If so, were sources natural or anthropogenic?
- Were any of the random forest predictor values outside the range of what was represented in the MMI calibration reference dataset?

We found likely explanations for some of the anomalous samples, but not for others. For example, Nescatunga Creek in Comanche County was a 'Best Reference' headwater site (modeled CFS = 3.85) that was in the bin of 'underachievers.' This was likely due to impacts from drought. From 2007-2010, MMI scores ranged from 62-65. In 2011, when there was an extreme drought (100% of the county rated D4 =exceptional drought), the MMI score dropped to 28 and percent air breather individuals increased to 40% (up from 0-13%). In 2012, the MMI score improved (38) but was still lower than normal, likely due to lag year effects from the drought. By 2015 & 2018, MMI scores had rebounded to 64 and 78, respectively.

On the other end of the spectrum, we looked into some of the 'overachieving' Smoky Hill River sites, which were 'High Stress' sites in larger systems (modeled CFS ≥ 198) that were receiving high MMI scores. Based on feedback from KDHE staff, the Smoky Hill River is sandy-bottomed, fairly wide and braided, and not deeply incised (at least not the first terrace) which allows for the development of a lot of edge habitat. It tends to flow clear, and the edge habitat has been pretty stable through time, providing refugia for several orders of insects (Figure 34). Being wide and in an area with less rainfall than to the east, blowout flows are not frequent. It is possible that the braided channel and often heavily vegetated sand/gravel bars and low/wide first terrace simulate a floodplain connection during modestly elevated flows. The clear water and elevated nutrients result in a lot of filamentous algae which probably enhances midge production (and serves as a food supply for accompanying predators). It is also possible that the disturbance index is not well calibrated for larger systems (and the sites should not be in the 'High Stress' category). The 'overachieving' Smoky Hill River sites have total watershed areas of more than 20,000 km^2 and most of the disturbance index variables are based on spatial scales of 5-km^2 or less[15], which is a very small proportion of the much larger watershed.

[15] At the time we developed the disturbance index, which was done collaboratively with KDHE, the consensus was to weight the more localized scales more heavily. We felt the more localized scales would better capture stressors affecting the BMI sampling reaches, and higher vs. lower quality riparian areas.

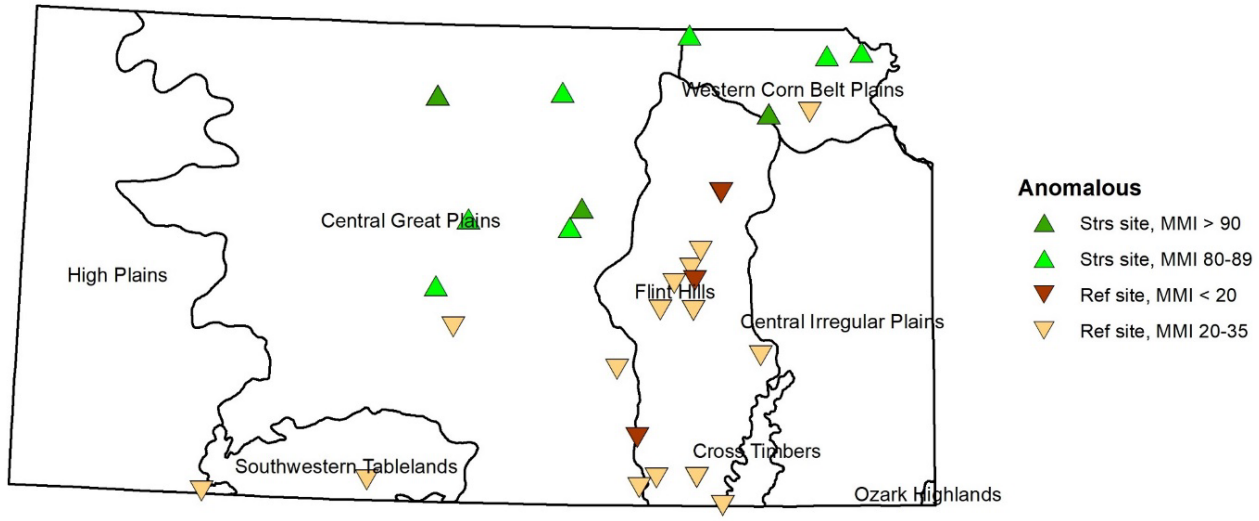


Figure 33. Locations of anomalous samples.

Figure 34. The Smoky Hill River has a number of 'overachieving' sites. It is sandy-bottomed, fairly wide and braided, and not deeply incised (at least not the first terrace) which allows for the development of a lot of edge habitat. This photo of Site SB268 was provided by KDHE.

6 Precision statistics and evaluation of sampling method and inter-annual variability

6.1 Overview

There are multiple potential sources of variability in bioassessments, including those associated with field sampling, laboratory sample processing, and temporal and spatial factors (Stribling et al. 2008, Mazor et al. 2009). Variability can contribute to inconsistencies in bioassessment results and should be taken into consideration when interpreting MMI results. A commonly used approach for quantifying variability in bioassessments is to calculate precision statistics based on the likeness of samples collected from the same site or stream reach (Cao et al. 2003). Precision of metrics allow attributing index variability to particular aspects of the macroinvertebrate assessment tools, and allow assessors to more confidently determine whether differences between MMI results at the same sites over time are due to variability associated with typical interannual differences and processing versus a significant shift in biological condition that might be attributed to another source of variability, such as a new stressor or remediation of stressor effects.

6.2 Methods

We performed a precision analysis to quantify the variability of the KS MMI and its input metrics based on sampling and temporal effects. We assessed five different types of variability, as summarized in Table 25 using all available data. Same site/same day samples were available at 173 sites. Same day replicates provide measures of variability and error in the sampling protocols, which includes microsite differences, sample crew variability, protocol application errors, and all aspects of sample processing and taxa list generation (taxonomic identification error, data entry error, etc.), as well as calculation errors for metrics and the index (e.g., taxa trait association error, site characteristic errors, etc.). The other types of variability had temporal components: seasonal and interannual (two versions: all months and summers only), and seasonal and programmatic (SP versus SB program). The multiple sources of variability are not mutually exclusive and cannot be completely separated from one another (and should not be interpreted additively).

Table 25. Five types of variability were analyzed in the KDHE BMI dataset.

Type of variability	Description	# Rep Sets	Total # Samples
Same site/same day	Samples collected on the same day from two adjacent reaches	173	351
Same site/same year/different day (seasonal)	Samples collected on different days during the same year	75	184
Same site/multiple years (interannual)	Samples collected during different years, including all months	179	2382
	Samples collected during different years, limited to summer only (June – Sept)	163	1499

Same site/same year/different program (seasonal/programmatic)	Samples collected on different days during the same year by different programs (SB versus SP)	17	38

The precision analysis focused on the KS MMI and component metrics (both raw metric values and adjusted metric scores). Precision statistics were generated using the R statistical environment (R Core Team, 2023) using an ANOVA with unique site IDs as the grouping variable for replicates. All sites (and samples within each replicate set) were analyzed together to assess the overall variability found in the dataset.

The statistics derived from the ANOVA include:

- **Mean-squared-error (MSE)**, which measures the variability within replicate sets (i.e., within sites). Higher MSE = greater variability (less desirable for bioassessment applications). Thus, MSE is an estimate of the variance within replicate sets.
- **Root-mean-squared-error (RMSE)**, which is the square root of the MSE and an estimate of the standard deviation of values within sets. RMSE tells you how concentrated the data are around the mean within sets. Higher RMSE = greater divergence (less precision). Thus, RMSE is an estimate of the standard deviation of metric values or index scores within sites.
- **Coefficient of variation (CV)**, which is the ratio of the standard deviation to the mean, where the mean is derived from the replicated values in the analysis. The CV is the percentage of the mean represented by the RMSE (100*RMSE/Mean). Higher CV = greater variability. CV is a standardized value that can be compared among metrics and indices even when the scales of metrics and indices may differ.
- **90% confidence interval (CI90)**, which is a measure of uncertainty. CI90 is the range around the observation within which the true mean value is expected to fall in 90% of cases. CI90 is calculated as the RMSE*1.645. The constant multiplier is derived from standard normal tables for the area under the normal curve at 90% confidence. For example, an index score of 70 points for an index with a CI90 of ± 10 points is expected to be between 60-80 points with 90% confidence. This range of possible scores comes from the intrinsic variability that causes MMI scores to fluctuate. Thus, the true mean MMI score is the measured value ± 10 points (from the example). The true mean is the mean that would be calculated from multiple observations. Thus, a higher CI90 = greater degree of uncertainty. The CI90 is best applied when comparing two values to detect changes over time or between sites. It is not appropriate for comparison of an index score to an index threshold.

For comparing variability among metrics, the CV is the most relevant statistic because it suggests which metrics might be contributing to index score variability more than others. Metric CVs of <20% are considered to be precise for repeated measures (rule of thumb). For indices, the CI90 indicates the range within which comparisons could be considered similar among sites or within sites over time.

6.3 Results

The MMI scores were most precise in the same site/same day replicates, with a CV of 13% and a CI90 of ±12.75 index points (Table 26). These statistics are within the range of precision statistics cited for other multimetric indices. A Montana study by Stribling et al. (2008) showed lowest variability in mountainous streams (CI90 = ±5.7) and highest in plains streams (CI90 = ±8.5). The measurement quality objectives recommended by Stribling et al. (2008) were CVs of 10-15% for mountains and plains indices. In

Michigan streams, Tetra Tech (2020) found CVs ranging from 13-32% and CI90s ranging from 14-20 index points across multiple site classes.

Introduction of temporal variability decreased MMI precision. CVs increased by ~10% (to 23-26%) and CI90s ranged from ±19 to 23 index points. Seasonal variability was slightly lower than interannual variability, and 'summer only' replicates had slightly lower variability than 'all month' replicates (but were within 1% CV and ±1 CI90 point of each other). When programmatic variability was added to the seasonal variability, it further decreased MMI precision, adding ~3% CV and ±4 CI90 index points (Table 26). Compared to a study of index temporal variability in Massachusetts streams (Block et al. 2021), the variability calculated for the KS MMI was high. In the Massachusetts study, interannual CVs for multiple IBIs ranged from 9.5 – 13% and CI90s ranged from 11 – 16 index points. The KDHE biologists have hypothesized that high interannual variability might be attributed in part to hydrologic variability that sometimes includes drought conditions. Drought years were not distinguished in the current precision analysis.

Table 26. Precision statistics for the MMI, for five types of variability.

Type of variability	# Rep Sets	Total # Samples	MSE	Mean	RMSE	CV	CI90
Same site/same day	173	351	60	59.33	7.75	13.06	12.75
Same site/same year/different day (seasonal)	75	184	137	51.2	11.71	22.87	19.26
Same site/multiple years (interannual) - summer only	163	1499	180.8	55.23	13.45	24.34	22.12
Same site/multiple years (interannual) - all months	179	2382	200.3	54.87	14.15	25.79	23.28
Same site/same year/different day & different program (seasonal/programmatic)	17	38	206.7	55.33	14.38	25.99	23.65

Among the MMI metrics, across all five types of variability, the HBI, the number of climber and clinger taxa, and the number of EPT taxa were least variable (had lower CVs). The HBI was especially precise in the raw form (CV = 1.8%). CVs for adjusted scores were mostly higher than for raw metric values (which was expected, given how raw values had a smaller range of values than the 0-100 scoring scale). Percent sensitive taxa was an exception, with most of its raw values having higher CVs than the scores (the highest being 56% for 'interannual variability – all months'). Number of semi-voltine taxa also had relatively high CVs, with seasonal variability being highest (54% for the adjusted score) (Table 27).

Table 27. Precision statistics for repeated measures of the KS MMI metrics.

Metric	Mean	Same site/ same day		Seasonal		Interannual - summer only		Interannual - all months		Seasonal/ programmatic	
		RMSE	CV	RMSE	CV	RMSE	CV	RMSE	CV	RMSE	CV
MMI	59.33	7.75	13.06	11.71	22.87	13.45	24.34	14.15	25.79	14.38	25.99
HBI	5.38	0.1	1.84	0.2	3.81	0.21	3.79	0.24	4.47	0.15	2.73
HBI (RF adj score)	62.63	9.26	14.79	18.5	30	20.44	37.28	22.29	38.06	15.81	27.91
# climber + clinger taxa	21.38	2.65	12.38	4.41	22.56	4.11	19.41	4.27	20.93	4.87	22.65
# climber + clinger taxa (RF adj score)	60.93	13.33	21.88	21.1	42.26	20.66	35.97	21.45	39.22	25.82	45.81
# semivoltine taxa	5.81	1.5	25.76	2.11	41.51	2.04	35.18	2.1	38.15	2.18	40.81
# semivoltine taxa (RF adj score)	57.01	18.02	31.61	24.48	53.62	23.85	43.33	24.89	47.77	26.39	47.43
% sensitive taxa*	6.97	2.71	38.93	2.76	38.28	3.16	50.55	3.72	56.01	3.55	52.45
% sensitive taxa* (RF adj score)	53.87	17.35	32.21	18.29	41.13	21.49	45.57	23.41	46.63	25.55	51.37
# EPT taxa	10.99	1.46	13.32	2.83	26.39	3.04	24.5	3.18	26.92	3.47	27.27
# EPT taxa (RF adj score)	62.2	9.4	15.11	19.57	36.11	19.69	31.92	21	35.72	24.92	42.79

*sensitive = BCG attribute III + IV_better

7 Exploration of assessment thresholds

In addition to having narrative biocriteria, some state biomonitoring programs have integrated numeric biocriteria into their Surface Water Quality Standards (SWQS). With numeric biocriteria, management actions can be triggered or prioritized based on assessments relative to a threshold (or thresholds). States like Maine and Minnesota use numeric biocriteria to evaluate Aquatic Life Use Attainment decisions and to designate different categories of biological condition. At this time, KDHE does not have plans to try and integrate numeric biocriteria into their SWQS. That would warrant additional analyses, a rule-making process that includes a period for public review and comment, as well as approval by the U.S. Environmental Protection Agency (EPA) following promulgation. Our objective here is only to explore potential MMI thresholds for multiple biological condition categories (e.g., Very Good, Satisfactory, Moderately Degraded, and Severely Degraded). We used three approaches to explore potential MMI thresholds: distribution statistics, balancing Type I and Type II errors, and crosswalking MMI scores with results from the Great Plains Biological Condition Gradient (BCG) model (Stamp et al., in progress).

7.1 Distribution statistics

The distribution statistics, or percentile-based, approach (in particular, the reference condition (RC) approach) is commonly used by states for setting numeric biocriteria thresholds (Hughes et al. 1986, Gibson et al. 1996). With the RC approach, MMI scores are calculated from a least-disturbed reference site dataset, and then a percentile of the MMI scores is chosen to represent the RC. Typically, the 25th or 10th percentile is used for the satisfactory condition threshold (e.g., Yoder and Rankin 1995, DeShon 1995, Barbour et al. 1996, Roth et al. 1997), but this varies across datasets. Having a sound, well-documented, reference dataset is critical to this approach.

When selecting percentiles, it is important to consider level of disturbance. Using a higher percentile in situations where the reference dataset has higher levels of disturbance (and likely includes some sites that are not truly of reference quality) provides a degree of safety. The 10th percentile of RC is generally used where there is greater confidence that the reference sites are of high quality (for example, with datasets that have many “minimally disturbed” sites). In reference datasets with only “least disturbed” sites, the 25th percentile is typically used (Stoddard et al. 2006). The 5th percentile of reference datasets is not commonly used since it is prone to the effect of outliers as well as variability and potential error in reference designations. Even when reference sites are carefully selected from a robust pool of minimally disturbed locations, natural variability and sampling error precludes the assumption that every reference sample is representative of biological integrity goals, so using percentiles less than 10% would likely underestimate impairment (or the departure from the desired reference condition). In heavily disturbed areas or regions where a stream class has overall poor condition (i.e., poorer than least disturbed), thresholds based on the 25th or 10th percentile are likely to be under-protective so an alternative or modified approach (such as the 75th or 90th percentile of all sites) is sometimes used. For example, in the Huron/Erie Lake Plains ecoregion, Ohio based their threshold on the 90th percentile of all sites (Ohio EPA 1987, 1989).

Table 28 shows distribution statistics for KS MMI scores for two datasets: 1) reference sites only; and 2) all samples.

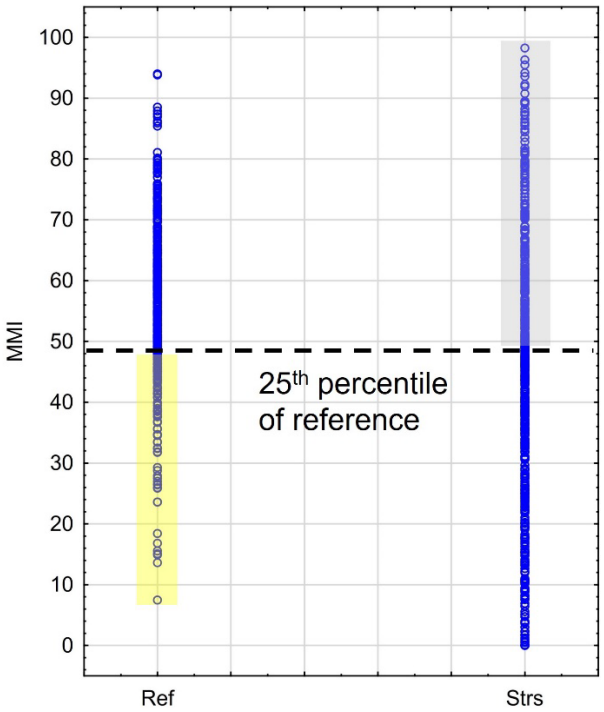
Table 28. KS MMI scores for multiple percentiles based on two datasets: reference only (n=383 samples); and all samples (n=2952).

Dataset	Percentiles	MMI score
Reference sites	5th	31.9
	10th	39.1
	15th	43.9
	20th	46.2
	25th	48.1
	50th	57.3
	75th	64.9
	90th	73.5
	95th	78.3
All sites	5th	16.6
	10th	25.0
	15th	31.0
	25th	39.6
	50th	53.3
	75th	66.8
	90th	78.1
	95th	84.8

7.2 Balancing Type I and Type II Error

For the second approach, we examined ratios of Type I and Type II error. Type I error is known as a "false positive" finding (in this case, falsely calling a site disturbed when it is not). Type II error captures "false negative" findings (or falsely calling a disturbed site undisturbed). When you decrease the probability of one error, it increases the probability of the other. A consequence of having a high Type I error rate is a higher likelihood of mistakenly subjecting undisturbed sites to potentially costly management actions, whereas having a high Type II error rate increases the likelihood of not detecting degradation. Most biomonitoring programs try to simultaneously minimize Type I and Type II errors (Breine et al. 2007), but approaches vary across entities and depend on acceptable error rates.

We measured Type I error as the percentage of reference sites that fell below the various percentiles/potential thresholds and Type II error as the percentage of stressed sites that had scores greater than or equal to the thresholds. Type I and II error rates and the number and percentages of samples that fell above or below the various thresholds are summarized in Figure 35. The 25th percentile of reference sites (which corresponds with a MMI score of 48.1) had the smallest difference between Type I and II errors (Type I = 24.8%; Type II = 43.9%; difference 19.1%). If 48 was used as the Satisfactory Biological Condition threshold, and the full dataset were considered, 1,781 samples (60.3%) would score greater than or equal to the threshold, and 1,171 samples (39.7%) would score below the threshold.



Type I error
- Erroneously calling a site disturbed
- Calculated as the % of Ref sites with MMI scores < the threshold. For example, if 48.1 is the threshold, 95 out of 383 reference samples (24.8%) have MMI scores < 48.1
- Type I error rates increase as the threshold increases

Type II error
- Erroneously calling a disturbed site un-disturbed
- Calculated as the % of Strs sites with MMI scores ≥ the threshold. For example, if 48.1 is the threshold, 219 out of 499 reference samples (43.9%) have MMI scores ≥ 48.1
- Type II error rates decrease as the threshold increases

Percentile (ref only)	MMI	Number (%) samples in each disturbance group				Type I error	Type II error	Dif'ce
		≥ Threshold		< Threshold				
		Ref (n=383)	Strs (n=499)	Ref (n=383)	Strs (n=499)			
5th	31.9	363 (94.8%)	363 (72.7%)	20 (5.2%)	136 (27.3%)	5.2%	72.7%	67.5%
10th	39.1	344 (89.8%)	306 (61.3%)	39 (10.2%)	193 (38.7%)	10.2%	61.3%	51.1%
15th	43.9	326 (85.1%)	264 (52.9%)	57 (14.9%)	235 (47.1%)	14.9%	52.9%	38%
20th	46.2	307 (80.2%)	243 (48.7%)	76 (19.8%)	256 (51.3%)	19.8%	48.7%	28.9%
25th	**48.1**	**288 (75.2%)**	**219 (43.9%)**	**95 (24.8%)**	**280 (56.1%)**	**24.8%**	**43.9%**	**19.1%**
50th	57.3	192 (50.1%)	142 (28.5%)	191 (49.9%)	357 (71.5%)	49.9%	28.5%	21.4%
75th	64.9	96 (25.1%)	96 (19.2%)	287 (74.9%)	403 (80.8%)	74.9%	19.2%	55.7%
90th	73.5	39 (10.2%)	54 (10.8%)	344 (89.8%)	445 (89.2%)	89.8%	10.8%	79%
95th	78.3	20 (5.2%)	38 (7.6%)	363 (94.8%)	461 (92.4%)	94.8%	7.6%	87.2%

Figure 35. Distribution of KS MMI scores across the reference (Ref) and stressed (Strs) disturbance categories. The table summarizes Type I and II error rates and the number and percentages of samples in each disturbance category that fell above or below the various thresholds. Cells highlighted in yellow show Type I error rates; gray cells show Type II error rates.

7.3 Biological Condition Gradient (BCG) crosswalk

During the development of the KS MMI, a concurrent yet independent project has been going on to calibrate quantitative Biological Condition Gradient (BCG) models for macroinvertebrate and fish assemblages in the Great Plains region, which includes streams in three Omernik Level 3 ecoregions in Kansas (Central Great Plains, Central Irregular Plains and Western Corn Belt Plains). Several KDHE biologists have been participating in the BCG project, along with biologists from Minnesota, Iowa, Nebraska, Missouri and Oklahoma.

7.3.1 Background on the BCG

The BCG is a conceptual model describing changes in aquatic systems with increasing levels of human disturbance (Davies and Jackson 2006, USEPA 2016, Hausmann et al. 2016, Gerritsen et al. 2017). It includes predicted changes in structural and functional characteristics of stream systems as they degrade in response to human disturbance (Figure 36). These measurable characteristics are defined as "attributes" of the biological communities and the physical habitat that reflect the condition of an aquatic ecosystem (USEPA 2016). The attributes include properties of the system and communities (e.g., taxa richness, structure, abundance, system functions) and organisms (e.g., tolerance, rarity, native range, physical condition).

The BCG has been divided into six levels of biological condition in response to increasing levels of stress (natural and undisturbed (BCG level 1) to completely biologically and ecologically disrupted (BCG level 6)). The six levels provide a flexible framework for a state to determine the number of levels that can be implemented. The number of levels realized will be influenced by both natural and programmatic reasons. For example, in some places, it may be possible to discriminate 6 different levels of condition. In other places, states may only be capable of discriminating three or four levels.

The BCG model describes changes in the biota that have been commonly observed by aquatic scientists in different regions across the country (Davies and Jackson 2006). The BCG provides a common language to interpret and communicate current ecological conditions relative to baseline conditions that are anchored in BCG level 1 ("as naturally occurs"). In this way, the scoring scale is intended to be universal (USEPA 2016, Davies and Jackson 2006), but descriptions of communities, taxa, and their responses to the anthropogenic stress gradient are calibrated to be specific to the conditions and communities found in the study area (in this case, Kansas). Calibration of the BCG for a specific region is a collective exercise among biologists to develop consensus assessments of sites by reviewing biological community data in sample worksheets, assigning the samples to a BCG level, and then eliciting the rules that the biologists use to assess the sites (Davies and Jackson 2006). The process involves assembling data, making taxon attribute assignments, assigning samples to BCG levels based on the framework shown in Figure 36, and, through a series of webinars and/or in-person meetings, developing narrative and quantitative BCG rules. The BCG calibration process uses an explicit reliance on professional judgment and development of consensus, supported with targeted analyses. Professional judgment is part of all biological condition assessments, including all common biotic indexes, even though such judgment may be hidden in apparently objective, quantitative approaches (e.g., Steedman 1994, Borja et al. 2004, Weisberg et al. 2008).

Numerous states have incorporated the BCG into their biological assessment programs to complement existing biological indices or other bioassessment tools and approaches. Examples include Minnesota, Alabama, Connecticut and Californa. The BCG provides a powerful approach for an operational monitoring and assessment program, for communicating resource condition to the public and for management decisions to protect or remediate water resources. The BCG and the calibrated decision system allow practical and operational implementation of multiple aquatic life uses in a state's water quality criteria and standards. Adoption of the BCG as an assessment tool in the context of multiple Aquatic Life Uses (Tiered Uses) yields the technical tools for protecting the state's highest quality waters, as well as developing realistic restoration goals for urban and agricultural waters. By refining these uses, good and exceptional quality water bodies can be protected while establishing attainable goals for water bodies degraded by legacy activities such as channelization. In doing so, the resulting management actions can be better tailored to attainable goals thereby improving effectiveness and efficiency of these actions. More detailed information on applications can be found in the US EPA BCG Practitioners Guide (2016).

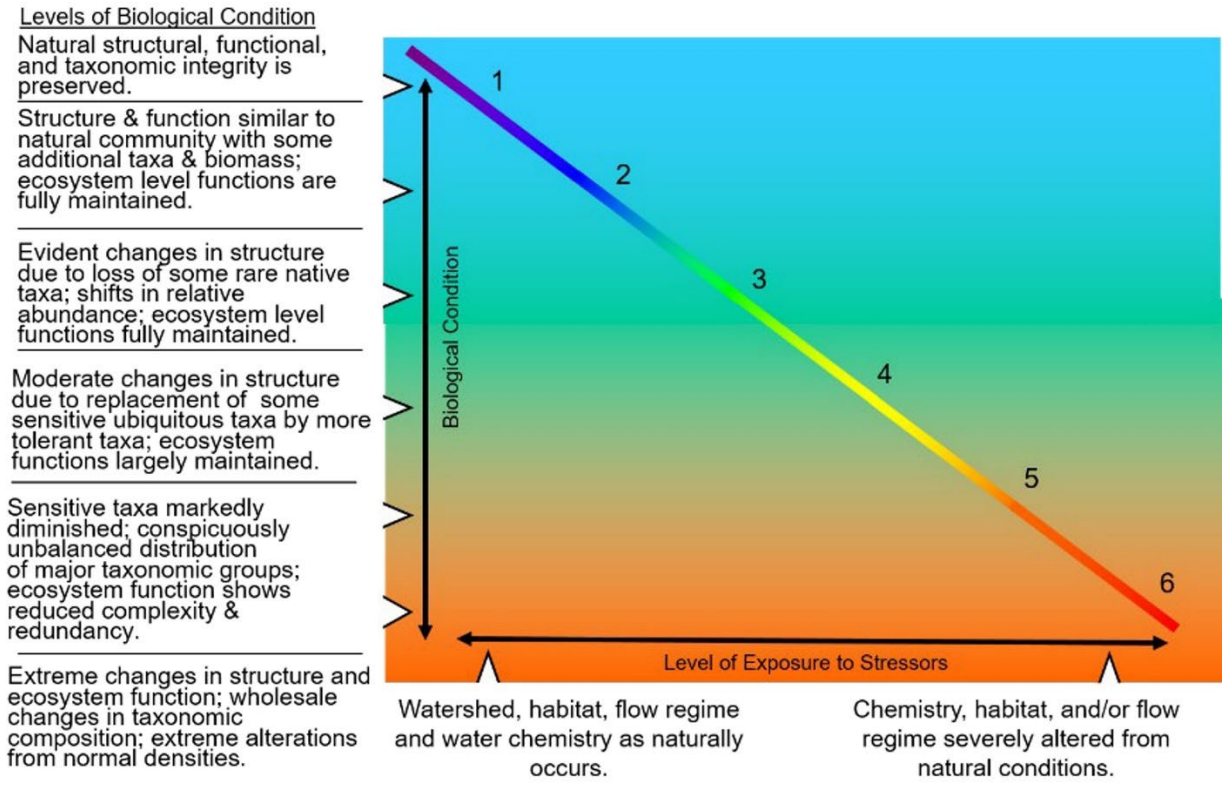


Figure 36. Conceptual model of the BCG (from US EPA 2016; modified from Davies and Jackson 2006). Six levels of condition (Y axis) along a gradient of increasing stress (X axis) ranging from naturally occurring to severely altered conditions are narratively described using biological information. Although in reality the relationship between stressors and their cumulative effects on the biota is likely nonlinear, the relationship is presented as such to illustrate the concept.

7.3.2 KS MMI – BCG Crosswalk

The KS Great Plains BCG model for macroinvertebrates is close to completion and is performing well[16] (Stamp et al., in progress). The BCG workgroup developed narrative and quantitative rules for four BCG levels (BCG level 3-6). Panelists concluded that BCG level 1 was not represented in the Kansas dataset, and the rules for BCG level 2 are conceptual since there weren't enough Level 2 samples to develop and test quantitative rules.

Preliminary narratives for BCG levels 3-6 are:

> **Level 3. Evident changes in structure of the biotic community and minimal changes in ecosystem function**. Good diversity and balance with high numbers of total taxa, good representation of Coleoptera, Odonata, Ephemeroptera and Trichoptera (COET) and low proportions of Chironomidae taxa and noninsect individuals. Sensitive taxa are present (as more than just single individuals) and highly tolerant taxa comprise only a small proportion of the individuals. There are high numbers of taxa representing habitat structure and diversity (hard substrate - snags or rocky, root mats or overhanging vegetation) and no obvious signs of nutrient enrichment.
>
> **Level 4. Moderate changes in structure of the biotic community with minimal changes in ecosystem function.** Good diversity and balance of all expected major groups, with moderate to high numbers of total taxa (including COET taxa), and low to moderate proportions of Chironomidae taxa, noninsect individuals and highly tolerant individuals. Some signs of nutrient enrichment may be present but nutrient indicator taxa don't dominate the assemblage.
>
> **Level 5. Major changes in structure of the biotic community and moderate changes in ecosystem function.** Moderate to poor diversity and balance. At least one EPT taxa must be present, along with low to moderate numbers of COET taxa. Highly tolerant individuals may comprise up to 80% of the individuals.
>
> **Level 6. Severe changes in structure of the biotic community and major loss of ecosystem function.** Extreme alterations from normal densities and distributions (e.g., less than 100 total individuals), with a conspicuously unbalanced or altered distribution of major groups (e.g., might be missing EPT taxa).

The KS Great Plains BCG model has 14 different metrics, two of which overlap with MMI metrics (number of climber + clinger taxa and number of EPT taxa). Both the BCG and MMI also utilize at least one metric based on sensitive taxa (Attribute III+IV_better). Unlike the MMI, which is calculated by averaging five metric scores, the BCG works like a logical cascade; if any of the BCG rules for a given level aren't met, the sample gets bumped down to the next level, with Level 6 being the lowest.

[16] To evaluate the performance of quantitative BCG models, the number of samples are assessed where the BCG decision model's nominal level exactly match the panel's median ("exact match") and the number of samples where the model predicted a BCG level that differed from the median expert opinion ("mismatch" samples). Currently, the quantitative model is 96% accurate in replicating the panel assessments within ± 0.5 BCG level and no BCG model outputs were more than 1 level off from the panelist median calls.

We used box plots to explore concordance between the MMI and preliminary results from the KS Great Plains BCG model and found good agreement (Figure 37), which is encouraging given how the two indices were developed independently. The box plots showed an expected downward pattern, with MMI scores decreasing as BCG levels increased from Level 2 to 6. There was no obvious bias in the direction of disagreement among models (as shown by the similar number of MMI assessments that were better or worse than the corresponding BCG assessments).

Some states have used the BCG to help inform thresholds for numeric indices for tiered aquatic life uses. As an example, Minnesota used the BCG to help set thresholds for several use categories (General, Exceptional, Modified) for their Indices of Biological Integrity (IBI). They calculated median IBI scores for each BCG level and used the median score for BCG level 4 to represent their General Use threshold (Bouchard et al. 2016). In the KDHE dataset, the median MMI score for BCG level 4 is 50.8, which is close to the 25th percentile of reference (MMI = 48), and also corresponds well with the scoring scheme of the MMI (MMI metrics that meet natural expectations receive scores in the 50-60 range). The median BCG level 3 MMI score (69) could be potentially be used as a threshold for identifying sites that exemplify the best biology, while the median BCG level 5 MMI score (31) could potentially be used to draw the line between moderately and severely degraded sites (Figure 37). Minnesota (and other states) have also used the BCG to link narratives to the numeric thresholds, which gives the numeric scores greater meaning.

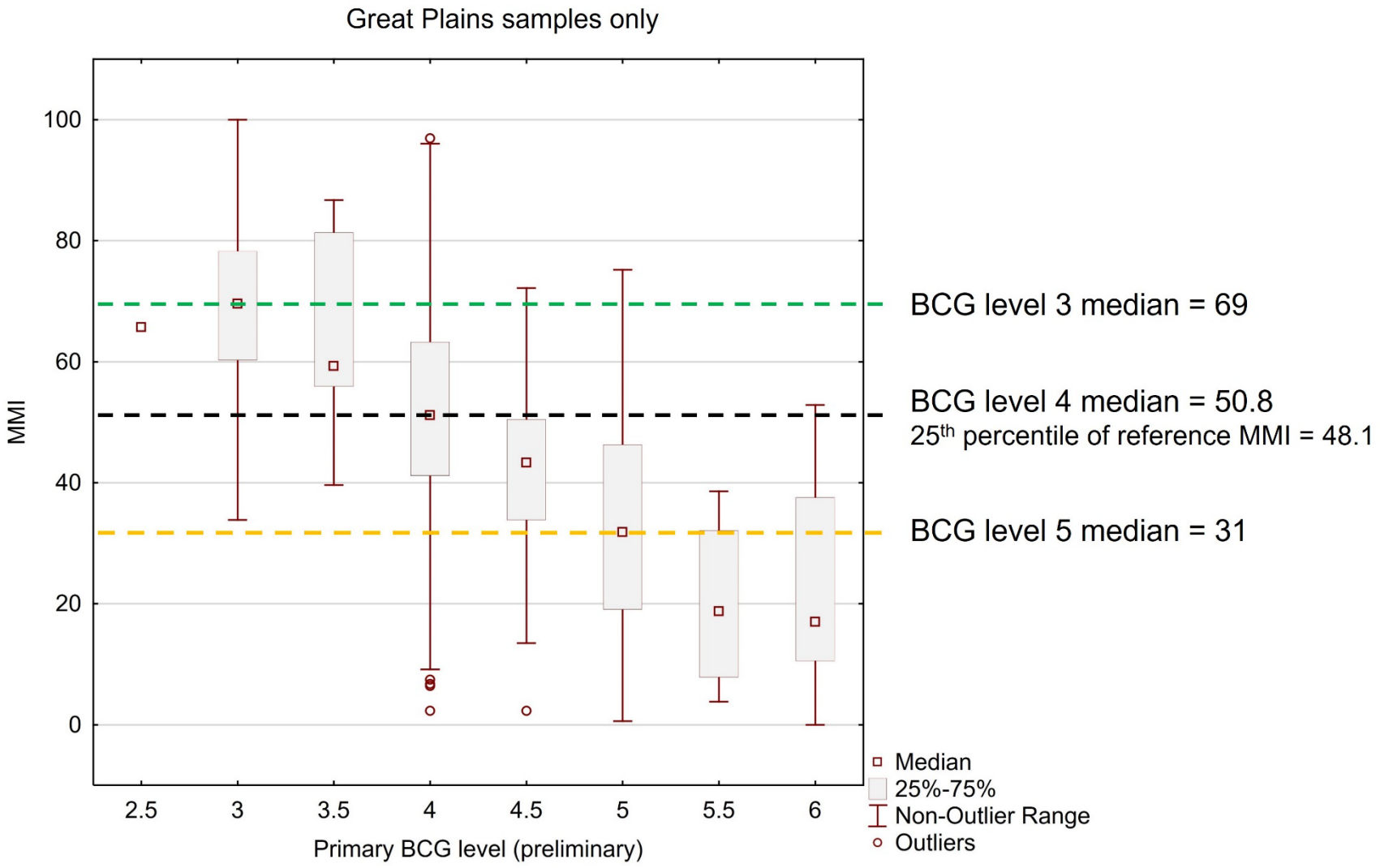


Figure 37. Box plot showing distributions of KS MMI scores by BCG levels for samples from the Central Great Plains, Central Irregular Plains and Western Corn Belt Plains ecoregions. Sample sizes: BCG Level 2.5 (2/3 tie) = 1, Level 3 = 90, Level 3.5 (3/4 tie) = 9, Level 4 = 455, Level 4.5 (4/5 tie) = 30, Level 5 = 113, Level 5.5 (5/6 tie) = 4, Level 6 = 15.

7.4 Combining multiple lines of evidence

If KDHE someday decides to establish numeric thresholds for the MMI, and decides to establish multiple biological condition categories (e.g., Very Good, Satisfactory, Moderately Degraded, and Severely Degraded), results from the three lines of evidence suggest potential MMI breakpoints of ~30/50/70. The breakpoint of 50 equals the preliminary median MMI score for BCG level 4 and is close to the 25th percentile of reference, which is commonly used to establish numeric thresholds for the satisfactory condition threshold in reference datasets with only "least disturbed" sites (Stoddard et al. 2006). The threshold of 50 also aligns well with the metric scoring scale. Metric values that are equal to the expected values (based on the natural predictor variables) receive scores in the 50-60 range.

KDHE could use these MMI breakpoints (30/50/70) to:

- Identify sites with the best biology, which might make good candidates for protection
- Document incremental changes (for example, improvement from severely to moderately degraded, or moderately degraded to satisfactory; or a decline in condition from very good to satisfactory)
- Develop realistic restoration goals for water bodies degraded by legacy activities such as channelization

Once the Great Plains BCG model is finalized, we recommend that KDHE recalculates the median MMI scores for BCG levels 3, 4 and 5 and confirms that they still align well with these thresholds. We also recommend that they investigate anomalous samples, where the BCG and MMI are most divergent. Because the Great Plains BCG model does not cover all the Level 3 ecoregions in Kansas, KDHE may decide to place more weight on the other lines of evidence. The best thresholds and threshold derivation method will depend on the goals and priorities of KDHE, as well as ecosystem considerations, such as acceptable deviations from reference conditions.

8 MMI calculator

The KS MMI can be calculated with a free R-based tool (referred to as a Shiny app) that can be accessed via this weblink: [placeholder]

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. Even though the Shiny apps are linked to R, users do not need R on their computers to run them, nor do they need to have any familiarity with R. They just need an internet connection. For those who prefer to work with R, the R code can be downloaded for free from this weblink: [placeholder].

The MMI calculator 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 MMI calculator and get a result. However, if samples do not meet the criteria listed below, results should be interpreted with caution.

Criteria:

- Geographic area: Kansas
- Stream type: freshwater, wadeable

- Subsample size: ≥ 200-count samples are recommended for best performance; results for samples with < 100 total individuals should be interpreted with caution
- Taxonomic resolution: genus or species-level as allowed by available keys, specimen condition, and specimen maturity
- Collection gear: 500-µm mesh D-frame net and fine point forceps
- Collection period: April 15– ~October 15 (leaf out)
- Collection method: KDHE Stream Probabilistic (SP) program or Stream Biological (SB) Monitoring Program protocols[17]. Organisms are composited from multiple habitats. Field staff perform active sampling for 60 minutes, trying to maximize the diversity of organisms being captured (targeting macrohabitats like riffles, leafpacks, undercut banks and rootmats, aquatic vegetation, and large woody debris). Organisms are picked in the field and identified in the lab.
- Predictor variables: data should come from the sources listed in Table 13. Results for sites with predictor values outside the range of values represented in the MMI reference calibration dataset should be interpreted with caution.

9 Discussion

We developed a MMI for macroinvertebrate assemblages in wadeable streams in Kansas, using data from 700 sites that spanned several decades. The new MMI improves the diagnostic ability of KDHE and other practitioners to identify degradation in biological integrity and water quality. The MMI is modernized compared to past assessment indices used in Kansas and makes use of a random forest model approach in which metric expectations for each site are predicted from multiple natural environmental variables at reference sites (thus, the classification is continuous, not in discrete site classes). The MMI is comprised of biological metrics that were found to be responsive to a general stressor gradient, are ecologically meaningful, diverse in response mechanisms and can be readily communicated.

The KS MMI was calibrated using the Reference Condition approach, which bases biological expectations on reference sites (which, in Kansas, represent least-disturbed conditions). If a site receives a MMI score that does not resemble reference scores, it indicates that there might be stressors influencing the biological condition at that site. Overall, the KS MMI was effective at discriminating between reference and stressed sites, in particular in small to medium-sized streams (<50 cfs and <5,000 km2) in the Flint Hills and Central Irregular Plains ecoregions. When we examined relationships with individual stressors, the MMI decreased noticeably when percent impervious exceeded 1% and percent hay/pasture exceeded ~30%. The MMI did not show a decipherable relationship with percent row crop.

In the Central Great Plains and Western Corn Belt Plains ecoregions, the KS MMI was less effective at distinguishing reference from stressed sites. Particularly noticeable were some 'overachieving' larger, higher flow (>50 cfs), glide-pool, stressed sites in the Central Great Plains that had high MMI scores. The anomalous patterns could be due to limitations with the MMI (e.g., larger, glide-pool sites had limited representation in the reference dataset) as well as potential shortcomings with the disturbance index. For example, we know that the percent land cover metrics are oversimplified and do not capture agricultural BMPs, such as tilling practices, crop rotations and installation of fencing to keep livestock out of the streams. It is also possible that the disturbance index needs to be recalibrated for larger

[17] Stream Probabilistic | KDHE, KS; Stream Biological Monitoring Program | KDHE, KS

systems. Most of the 'overachieving' sites had total watershed areas of more than 20,000 km^2. All but one of the primary disturbance variables were based on a spatial scale of 5-km^2 or less, which is a very small proportion of the these larger watersheds. Another possible contributing factor is elevated nutrients bumping up taxa richness. At one of the 'overachieving' Smoky River sites, a KDHE biologist noted that elevated nutrients result in a lot of filamentous algae which probably enhances midge production and serves as a food supply for accompanying predators.

Moving ahead, we recommend that KDHE consider doing the following:

Set up a formal process for evaluating MMI performance. We recommend using a targetted approach, starting with investigation of the anomalous samples listed in Attachment 4. Use of a structured process, such as following the checklist in Section 5.4, is recommended. Do the MMI scores or reference/stressed designations seem 'off'? If so, are there common themes - e.g., at underachieving sites, are there stressors or site characteristics that appear to be impacting the BMI community that aren't being picked up in the disturbance index? The MMI alignment with the BCG categories (including the Central Great Plains and Western Corn Belt Plains ecoregions, which had lower DEs) shows that the MMI is a fair representation of the biological values described by a panel of biological experts from the Great Plains. The concordance suggests that the MMI disturbance gradient described using landscape characteristics is not effectively describing stressors on BMI communities in some ecoregions.

Explore ways to reduce variability. The MMI has fairly high varibility. If this can be reduced, it would allow assessors to more confidently determine whether differences between MMI results at the same sites over time are due to variability associated with typical interannual differences and processing versus a significant shift in biological condition that might be attributed to another source of variability, such as a new stressor or remediation of stressor effects. Some components of MMI variability, such as interannual variability stemming from extreme weather events, cannot be reduced. But there may be opportunities in other areas, such as standardization of taxonomic efforts and nomenclature across the SP and SB programs, and narrowing of the index period used by the SB program.

Collect additional data. We recommend that the SB and SP programs achieve greater consistency in the type of habitat data they are collecting. Moving ahead, it would be helpful to have Rapid Habitat Assessment survey data and estimates of flow habitat (riffle, run, glide, pool) and substrate composition at all the sites to allow for further exploration of the differences in the discrimination efficiency of the MMI in riffle-run vs glide-pool streams. We also recommend that during biological sampling events, both programs collect (and make accessible) *in situ* water quality measurements. Of particular interest are conductivity data (samples with values >1500 µmohs/cm should be flagged for further evaluation to determine the source (natural vs. anthropogenic) and to evaluate whether the elevated chloride (and specific conductance) concentrations are likely driving a loss in species richness and abundance). Also of interest are water temperature data, which can signal the presence of localized groundwater inputs that can also impact the BMI community.

Also, if not already doing so, KDHE may also want to consider collecting diatoms. Diatom responses to disturbance differ from the responses of other aquatic groups and might be better suited for detecting certain types of disturbance (Johnson et al. 2006, Hering et al. 2006, Justus et al. 2010). For example, algae are known to be sensitive to nutrients. In fact, they have been called the aquatic life "first responders" to nutrient impacts and have been used effectively in developing nutrient targets (Stevenson and Smol 2003, Stevenson et al. 2010, Charles et al. 2019).

Do more targeted sampling to broaden the disturbance gradients and types of sites represented in the reference dataset. From this project, we know that there are patterns in the BMI community related to size, habitat (riffle-run vs. glide-pool) and ecoregion, as well as from natural geological influences. Adding to the reference dataset will be important for future MMI recalibrations (which are recommended periodically; e.g., Jessup and Stribling 2008, Stribling et al. 2016).

Explore the use of multiple biological condition categories. Our preliminary analyses suggest potential MMI breakpoints of ~30/50/70 for establishing four biological condition categories (e.g., Very Good, Satisfactory, Moderately Degraded, and Severely Degraded). This should be revisited once the Great Plains BCG (Stamp et al., in progress) is completed. The BCG is helpful for not only informing MMI thresholds, but also adds meaning to the numeric thresholds due to the narratives that accompany the BCG levels. There is precedent for this approach. Minnesota used the BCG to help inform thresholds for several use categories (General, Exceptional, Modified) for their IBIs. They calculated median IBI scores for each BCG level and used the median score for BCG level 4 to represent their General Use threshold (Bouchard et al. 2016). We recommend that KDHE consider eventually expanding the BCG to include all of Kansas.

10 Literature cited

Angelo, R.T., G.L. Knight, K.T. Olson and T.C. Stiles. 2010. Kansas reference streams: selection of suitable candidates, impending threats to reference stature, and recommendations for longterm conservation. Bureau of Environmental Field Services and Bureau of Water, Kansas Department of Health and Environment, Topeka, Kansas. 61 pp. Available online: https://www.kdhe.ks.gov/DocumentCenter/View/13768/Kansas-Reference-Streams-PDF

Bailey, R.C., R.H. Norris, and T.B. Reynoldson. 2004. Bioassessment of freshwater ecosystems: using the reference condition approach. Kluwer Academic Publishers, New York.

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 bioassement 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. 339 pp. Full text available at www.epa.gov/owow/monitoring/rbp.

Block, B., J. Stamp, and B. Jessup. 2021. Interannual Variability in Metric Values and Index Scores for Wadeable Streams in Massachusetts. Prepared for the Massachusetts Department of Environmental Protection, Worcester, MA.

Blocksom, K.A. 2003. A Performance Comparison of Metric Scoring Methods for a Multimetric Index for Mid-Atlantic Highlands Streams. Environmental Management, 31, 670-682.

Borja, A., Franco, J., Muxika, I., 2004. The Biotic Indices and the Water Framework Directive: the required consensus in the new benthic monitoring tools. Marine Pollution Bulletin 48 (3–4), 405–408

Breine, J., Maes, J., Quataert, P., Van den Bergh, E., Simoens, I., Thuyne, G., and C. Belpaire. 2007. A fish-based assessment tool for the ecological quality of the brackish Schelde estuary in Flanders (Belgium). Hydrobiologia 575: 141-159.

Carlisle, D.M., S.A. Spaulding, M.A. Tyree, N.O. Schulte, S.S. Lee, R.M. Mitchell, and A.A. Pollard. 2022. A web-based tool for assessing the condition of benthic diatom assemblages in streams and rivers of the conterminous United States. Ecological Indicators, 135, p.108513.

Cao, Y., C. P. Hawkins, and M. R. Vinson. 2003. Measuring and controlling data quality in biological assemblage surveys with special reference to stream benthic macroinvertebrates. Freshwater Biology 48:1898–1911.

Cao, Y., C.P. Hawkins, J. Olson, and M.A. Kosterman. 2007. Modeling natural environmental gradients improves the accuracy and precision of diatom-based indicators. Journal of the North American Benthological Society 26(3), 566-585.

Charles D.F., A.P. Tuccillo and T.J. Belton. 2019. Use of diatoms for developing nutrient criteria for rivers and streams: a biological condition gradient approach. Ecological Indicators 96: 258-269.

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

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

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

Flotemersch, J.E., J.B. Stribling, and M.J. Paul. 2006. Concepts and Approaches for the Bioassessment of Non-wadeable Streams and Rivers. US Environmental Protection Agency, Office of Research and Development.

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

Gibson G. R., M. Barbour, J. B. Stribling, J. Gerritsen & J. R. Karr. 1996. Biological criteria: technical guidance for streams and rivers - revised edition. EPA 822-B-96-001. U.S. Environmental Protection Agency, Washington, D.C.

Hastie, T.J. and R.J. Tibshirani. 1999. Generalized additive models. Washington, DC: Chapman & Hall/CRC.

Hausmann, S., D.F. Charles, J. Gerritsen and T.J. Belton 2016. A diatom-based biological condition gradient (BCG) approach for assessing impairment and developing nutrient criteria for streams. Science of the Total Environment 562: 914-927.

Hering, D., R.K. Johnson, S. Kramm, S. Schmutz, K. Szoszkiewicz, and P.F.M. Verdonschot. 2006. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. Freshwater Biology, 51(9), pp.1757-1785

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. https://doi.org/10.1899/12-009.1

Hill, R.A., M.H. Weber, S.G. Leibowitz, A.R. Olsen, and D.J. Thornbrugh, 2016. The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. Journal of the American Water Resources Association (JAWRA) 52:120-128. DOI: 10.1111/1752-1688.12372.

Hughes, R.M., D.P. Larsen, and J.M. Omernik. 1986. Regional reference sites: a method for assessing stream potentials. Environmental Management 10:629–635.

Huggins, D.G. and M. Moffett. 1988. Proposed Biotic and Habitat Indices for use in Kansas Streams Report No. 35 of the Kansas Biological Survey.

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

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

Johnson R.K., D. Hering, M.T. Furse and R. Clarke. 2006. Detection of ecological change using multiple organism groups: metrics and uncertainty. Hydrobiologia, 566, 115–137.

Justus, B. G., J.C. Petersen, S.R. Femmer, J.V. Davis, and J. E. Wallace. 2010. A comparison of algal, macroinvertebrate, and fish assemblage indices for assessing low-level nutrient enrichment in wadeable Ozark streams. Ecological Indicators, 10(3): 627-638

Kansas Department of Health and Environment (KDHE). 2020. Part III: Stream Biological Monitoring Program Quality Assurance Management Plan KDHE Stream Biological Monitoring Program: https://www.kdhe.ks.gov/1233/Stream-Biological-Monitoring-Program

Kansas Department of Health and Environment (KDHE). 2021. Part III: Stream Probabilistic Monitoring Program Quality Assurance Management Plan
https://www.kdhe.ks.gov/DocumentCenter/View/14395/Stream-Probabilistic-Monitoring-Program---Part-III-PDF

Karr, J.R., K.D. Fausch, P.L. Angermeier, P.R. Yant, and I.J. Schlosser. 1986. Assessing biological integrity in running waters: a method and its rationale. Illinois Natural History Survey Special Publication 5.

KDHE Stream Probabilistic Program: https://www.kdhe.ks.gov/1230/Stream-Probabilistic Quality-Assurance-Management-Plan-PDF (ks.gov)

Kansas Surface Water Quality Standards (KDHE). 2022. Prepared by The Kansas Department of Health and Environment. Available online: https://www.kdhe.ks.gov/1381/Kansas-Surface-Water-Quality-Standards

Leppo, E.W. 2021. BioMonTools R package GitHub, web - https://leppott.github.io/BioMonTools/ GitHub, code, issues and discussion - https://github.com/leppott/BioMonTools Website/Shiny app – weblink is available upon request. Contact Jen.Stamp@tetratech.com

Liaw, A. and M. Wiener 2002. Classification and Regression by randomForest. R News 2(3), 18--22.

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., A. H. Purcell, and V. H. Resh. 2009. Long-term variability in bioassessments: A twenty-year study from two northern California streams. Environmental Management 43:1269-1286.

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.

Ohio EPA. 1987. Biological criteria for the protection of aquatic life: Volume I: The role of biological data in water quality assessment. Ohio EPA division of Water Quality Planning and Assessment, Columbus, OH.

Ohio EPA. 1989. Biological Criteria for the Protection of Aquatic Life: Vol. III. Standardized Biological Field Sampling and Laboratory Methods for Assessing Fish and Macroinvertebrate Communities. Ohio EPA division of Water Quality Planning and Assessment, Columbus, OH.

Olson, J.R. and Cormier, S.M., 2019. Modeling spatial and temporal variation in natural background specific conductivity. Environmental Science and Technology. doi: https://dx.doi.org/10.1021/acs.est.8b06777

Perry, C.A., Wolock, D.M., and Artman, J.C., 2004. Estimates of median flows for streams on the Kansas surface water register: U.S. Geological Survey Water Resources Investigations Report 2004-5032, 219 p., https://doi.org/10.3133/sri20045032

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://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, J.C. Seibel. 1997. Maryland Biological Stream Survey: Ecological status of non-tidal streams in six basins sampled in 1995. Maryland Department of Natural Resources, Chesapeake Bay and Watershed Programs, Monitoring and Non-tidal Assessment, Annapolis, Maryland. CBWP-MANTA-EA-97-2.

Stamp et al., in progress. Calibration of the Biological Condition Gradient (BCG) for Macroinvertebrate Assemblages in Freshwater Wadeable Streams in the Great Plains Region of Kansas, Missouri, Iowa and Nebraska.

Steedman, R.J. 1994. Ecosystem health as a management goal. Journal of the North American Benthological Society. 13(4):605–610

Stevenson, R.J. and J.P. Smol. 2003. Use of algae in environmental assessments. Academic Press, San Diego, CA.

Stevenson, R.J., Y. Pan, & H. van Dam. 2010. Assessing environmental conditions in rivers and streams with diatoms. The Diatoms: Applications for the Environmental and Earth Sciences, 2nd ed. Cambridge University Press, Cambridge, 5785.

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

Stribling, J. B., B. K. Jessup, and D. L. Feldman. 2008. Precision of benthic macroinvertebrate indicators of stream condition in Montana. J. N. Am. Benthol. Soc. 27:58-67.

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

Tang, T., R.J. Stevenson, and D.M. Infante. 2016. Accounting for regional variation in both natural environment and human disturbance to improve performance of multimetric indices of lotic benthic diatoms. Science of the Total Environment, 568:1124-1134

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

Tetra Tech. 2020. Calibration of a Multi-metric Macroinvertebrate Index for Assessment of Wadeable Michigan Streams. Prepared for the Michigan Department of Environment, Great Lakes, and Energy (EGLE), Lansing, MI and U.S. Environmental Protection Agency, Region 5, Chicago, IL.

Tetra Tech, Inc. 2021. Wisconsin Disturbance Index. Prepared for EPA Region 5.

Tetra Tech 2023. Taxa tolerance analyses on the KDHE dataset [in progress]

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.

Twardochleb, Laura & Hiltner, Ethan & Pyne, Matthew & Zarnetske, Phoebe. 2021. Freshwater insects CONUS: A database of freshwater insect occurrences and traits for the contiguous United States. Global Ecology and Biogeography. 30. 10.1111/geb.13257.

U.S. Environmental Protection Agency (USEPA). 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 (USEPA). 2016. A Practitioner's Guide to the Biological Condition Gradient: A Framework to Describe Incremental Change in Aquatic Ecosystems. Office of Water, Washington DC. EPA 842-R-16-001.

Weisberg, S.B., B. Thompson, J.A. Ranasinghe, D.E. Montagne, D.B. Cadien, D.M. Dauer, D. Diener, J. Oliver, D.J. Reish, R.G. Velarde, and J.Q. Word. 2008. The level of agreement among experts applying best professional judgment to assess the condition of benthic infaunal communities. Ecological Indicators 8:389–394.

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

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.