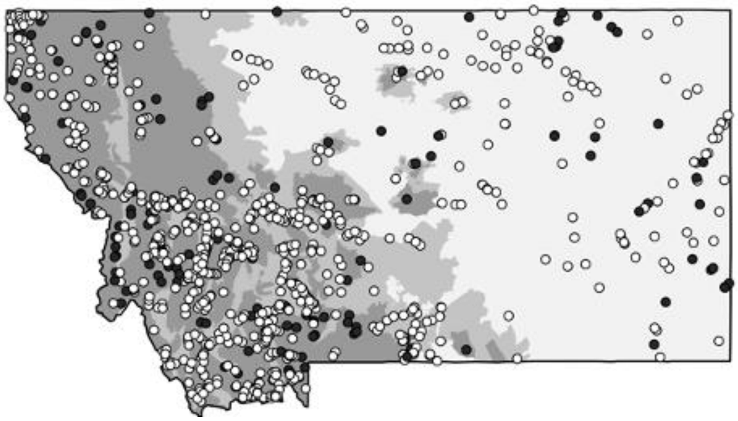
Eutrophication thresholds associated with benthic macroinvertebrate condition in Montana streams



Prepared for:

Montana Department of Environmental Quality
Helena, MT

Prepared by:

Nicholas O. Schulte and Joseph M. Craine
Jonah Ventures
5485 Conestoga Ct #210
Boulder, CO 80301

Prepared on:

October 5, 2023

Table of Contents

1.0 Introduction

The Montana Department of Environmental Quality (DEQ) is developing a translation process for its narrative nutrient standards (Administrative Rules of Montana 17.30.637(1)(e)) that uses the responses of benthic macroinvertebrate community characteristics (i.e., metrics) to causal eutrophication indicators (nitrogen, phosphorus, benthic algal chlorophyll *a*, and benthic algal ash-free dry weight) as part of the process of interpreting the standards.

Benthic macroinvertebrates are often considered to be secondary indicators of nutrient enrichment in wadeable streams (Mazor et al. 2022). Nitrogen and phosphorus are the most common causes of eutrophication in freshwater ecosystems, which often leads to an excess of benthic algae (or periphyton) on the streambed (Poikane et al. 2021). Such algal growth can reduce the quality of food, available habitat, and oxygen availability for macroinvertebrates (Bowman et al. 2007). The community composition of macroinvertebrates (i.e., the relative numbers of taxa and individuals at a location) reflects these responses to nutrient enrichment over time (Chambers et al. 2006). Therefore, macroinvertebrates can be used as robust, integrative indicators of eutrophication and biological condition (Heiskary and Bouchard 2015).

This report documents the analysis of thresholds, or change points, in the relationships between macroinvertebrate metrics and eutrophication indicators to support the translation of Montana's narrative nutrient standards relative to macroinvertebrate condition. This analysis follows a weight-of-evidence, or multiple-lines-of-evidence, approach that is recommended by the U.S. Environmental Protection Agency (EPA) for development of nutrient criteria, which integrates reference site distributions, predictive relationships, existing thresholds, and best professional judgment. The specific objectives of the present study were to:

- Curate water quality data for co-analysis with existing macroinvertebrate metric data,
- Characterize the macroinvertebrate metrics that are most responsive to eutrophication indicators (total nitrogen, total phosphorus, benthic algal chlorophyll *a*, and benthic algal ash-free dry weight),
- Identify candidate thresholds in macroinvertebrate metrics and eutrophication indicators for each of three macroinvertebrate regions in Montana (Mountains, Low Valleys and Transitional, and Plains) using multi-model selection and reference site distributions,
- Determine additional effects of covariates (e.g., temperature, flow, pH, specific conductance) on candidate thresholds after accounting for the influence of eutrophication indicators,
- Test whether multimetric indices (MMIs) yielded higher explanatory power or substantially different causal variable changepoints than single metric models in threshold analysis of macroinvertebrate condition.

2.0 Data preparation

2.1 Macroinvertebrate metrics

Prior to the present analysis, Rhithron Associates, Inc. downloaded all benthic macroinvertebrate count data from the Water Quality Portal (WQP) and also identified all relevant data from its own database that were collected in Montana by DEQ, EPA, and other collaborators. Taxonomy was harmonized, and samples were curated according to macroinvertebrate and site selection criteria: adequate target count, consistent field and laboratory methods, wadeable streams/medium rivers only, and an index period between 2005 and 2021. From the count data, 191 metrics were generated using BioMonTools (Leppo et al. 2021) across 577 harmonized taxa and 1606 curated samples. Most metrics were calculated in four ways: number of taxa (prefix "nt_"), percent of taxa ("pt_"), number of individuals ("ni_"), and percent of individuals ("pi_"). Metrics that represent diversity or tolerance indices were calculated according to the respective formula. This sample-by-metric matrix defined the site and date ranges of the present analysis.

All data processing and analysis in the present work was conducted using R v.4.2.0 (R Core Team 2022). For samples from the same site and date (e.g., from field replicates or methods comparisons), values from each metric were averaged to reduce the influence of patterns caused by the geographic proximity of samples or the date of measurement (i.e., spatial and temporal autocorrelation). Repeat visits to sites between 2005 and 2021 were retained to account for variable water quality conditions and macroinvertebrate assemblages over this 17-year time period. Following deduplication, 1415 samples remained.

2.2 Macroinvertebrate regions

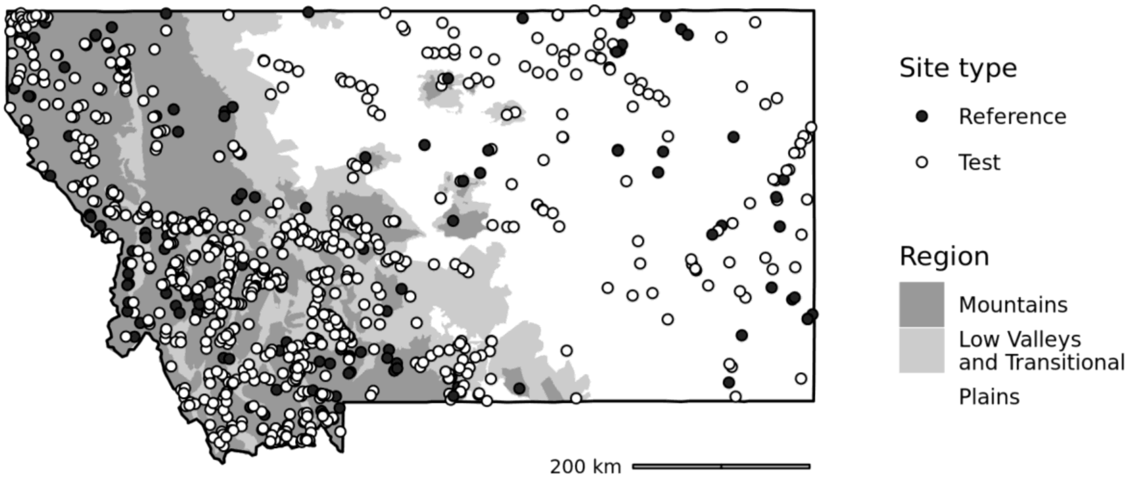


Figure 1. Map of Montana macroinvertebrate regions with reference and test sites used in the present analysis.

During construction of MMIs and observed/expected models for Montana, Jessup et al. (2006) determined that Montana stream macroinvertebrates were best classified according to three site classes that parallel previously defined physiographic and ecological regions: Mountains, Low Valleys, and Plains. Community composition of reference sites strongly differed among these macroinvertebrate regions. Subsequent work showed that transitional regions on the eastern side of the Rocky Mountain front are biologically more similar to western Montana than to the plains further to the east (Teply and Bahls 2007); this pattern is consistent with earlier ecoregion maps which describe a "Montana Valleys and Foothill Prairies" ecoregion (Bahls et al. 1992). Therefore, the present analysis focused on region-specific analyses of Mountains, Low Valleys and Transitional, and Plains – hereafter referred to as macroinvertebrate regions. Jessup et al. (2006) listed a subset of ecoregions and other geographic characteristics belonging to each macroinvertebrate region, but this list was incomplete given the site list used in the present study. In consultation with DEQ, macroinvertebrate regions were defined according to current Level III and IV Ecoregions (Woods et al. 2002) as listed in Table 1.

Table 1. Level II and IV Ecoregions associated with each Montana macroinvertebrate region.

Macro-invertebrate region	Ecoregions
Mountains	15. Columbia Mountains/Northern Rockies (excl. 15c Flathead Valley) 16. Idaho Batholith 17. Middle Rockies (excl. Level IV Ecoregions in Low Valleys and Transitional) 41. Canadian Rockies
Low Valleys and Transitional	15c. Flathead Valley 17s. Bitterroot-Frenchtown Valley 17u. Paradise Valley 17w. Townsend Basin 17aa. Dry Intermontane Sagebrush Valleys 17ac. Big Hole 17ak. Deer Lodge-Philipsburg-Avon Grassy Intermontane Hills and Valleys 42l. Sweetgrass Uplands 42n. Milk River Pothole Upland 42q. Rocky Mountain Front Foothill Potholes 42r. Foothill Grassland 43s. Non-calcareous Foothill Grassland 43t. Shield-Smith Valleys 43u. Limy Foothill Grassland 43v. Pryor-Bighorn Foothills 43o. Unglaciated Montana High Plains
Plains	18. Wyoming Basin 42. Northwestern Glaciated Plains (excl. Level IV Ecoregions in Low Valleys and Transitional) 43. Northwestern Great Plains (excl. Level IV Ecoregions in Low Valleys and Transitional)

2.3 Water quality data

Water quality data were downloaded from the WQP using the following search parameters: *State: Montana*, *Site Type: Stream*, *Date Range: 01-01-2005 to 12-31-2021*, *Data Profiles: Sample Results*. A total of 1606559 observations from 6669 sites were reported. Data were filtered to sites in the metric dataset (column: *MonitoringLocationIdentifier*) and water quality variable (column: *CharacteristicName*), targeting variables associated with four primary eutrophication indicators (nitrogen, phosphorus, benthic algal chlorophyll *a*, and benthic algal ash-free dry weight [AFDW]) and an assortment of background variables or other stressors known to influence macroinvertebrates (alkalinity, aluminum, chloride, dissolved oxygen, flow, hardness, iron, magnesium, mercury, pH, sodium, sulfate, temperature, solids, specific conductance, turbidity, and zinc). Water quality variables were separated by media (e.g., water or sediment) and fraction (e.g., total or dissolved) and converted each variable to consistent units.

For most samples across the index period, benthic chlorophyll *a* and AFDW were sampled from multiple transects using either template, hoop, or sediment core collection methods (DEQ 2021). Biomass values were then calculated as weighted averages of each method (excluding sediment cores for AFDW). In the WQP, most biomass values were reported as final weighted averages, but some were reported as individual transect values or method-specific composite values, each of which required further analysis for harmonization. Processing of benthic algal biomass records was as follows:

- Pre-calculated weighted averages or composite measurements with only a single collection method across all transects (chlorophyll *a*) or non-sediment core transects (AFDW). No further analysis. 68% of samples.
- Individual transect values by collection method (i.e., template, hoop, or core). Weighted average site means were calculated ignoring non-detect transects (including 0.5 * detection limits yielded mean values that were highly correlated to those from ignoring non-detects, Pearson *r* = 0.99) and excluding core transects from AFDW calculations. 20% of samples.
- Chlorophyll *a* measurements from the surface area of a single rock. All records were from 2005. No further analysis. 6% of samples.
- No method reported. All records were from the 2019 National Aquatic Resource Surveys. No further analysis. 6% of samples.

Prior to further data processing, water quality measurements were averaged across multiple samples taken at the same site on the same day, as was done for macroinvertebrate metrics. Accordingly, a sample was defined as a unique site-by-date.

To account for multiple detection limits and/or reporting limits for a given water quality variable across the index period, the 5th percentile of each water quality variable was calculated across all samples, excluding non-detects. If 0.5 * detection limit was less than or equal to the 5th percentile, 0.5 * detection limit was used as the measured value. If 0.5 * detection limit was greater than the 5th percentile, the observation was removed. Among eutrophication indicators, 2% of TN, 0% of TP, 4% of benthic chlorophyll *a*, and 0% of benthic AFDW observations were removed with this approach.

Since macroinvertebrate responses to water quality are integrative over time and water quality measurements were not always collected the day of macroinvertebrate sampling, the water quality values for each sample-by-variable from up to 30 days before and 7 days after macroinvertebrate sampling were averaged. This increased the number of samples with data for eutrophication indicators by ~5 – 20%, depending on the variable.

The most commonly observed fraction of each eutrophication indicator was selected: total nitrogen (TN), total phosphorus (TP), benthic algal chlorophyll *a* (corrected for pheophytin), and benthic algal AFDW. Six other variables had observations in at least 50% of samples and were selected for further analysis: water temperature, flow, pH, hardness, specific conductance, and total suspended solids.

Extreme outliers for each variable were removed based on manual inspection of distributions and consultation with DEQ regarding anomalous events, nontarget sites, and possible equipment malfunction. During outlier removal, 5 sites (each with one sample) were removed, as they represented large rivers. As a result, 1410 samples with macroinvertebrate metric and water quality data were used in further analysis (Table 2, Table S1).

The distribution of each water quality variable-by-region was assessed using histograms (Figure S1). All variables except temperature and pH followed a log-normal distribution. That is, after log10(x) transformation, the variable was approximately normally distributed. Otherwise, each variable was strongly right skewed, with the vast majority of observations clustered at the low end of the variable range and few observations of very large values. Transforming log-normal variables was necessary prior to data analysis to (1) stabilize the variance across the entire range of values, (2) ensure that normality assumptions of the statistical methods used were met, (3) allow models to more sensitively determine relationships between metrics and water quality variables across dynamic response ranges, and (4) decrease the influence of rare, extreme observations. Therefore, all variables except temperature and pH, which were already normally distributed, were log10(x) transformed prior to data analysis. Flow values were log10(x + 1) transformed due to observations of 0 cfs.

Table 2. Water quality variables selected for data analysis. Values are based on samples remaining after outlier removal. Values for each region are mean (standard deviation) across all samples and reference samples.

Variable	Outlier threshold	Samples		Mountains		Low Valleys and Transitional		Plains	
		All	Ref.	All	Ref.	All	Ref.	All	Ref.
Samples		1410	319	689	206	461	47	260	66
Total nitrogen, TN (mg/L)	2	929	298	0.16 (0.25)	0.10 (0.15)	0.38 (0.54)	0.14 (0.07)	0.99 (0.79)	1.21 (0.87)
Total phosphorus, TP (mg/L)	5	1067	297	0.02 (0.04)	0.01 (0.02)	0.04 (0.08)	0.02 (0.02)	0.11 (0.20)	0.10 (0.12)
Benthic chlorophyll *a* (mg/m^2)	1300	733	232	17.02 (20.35)	12.02 (9.79)	29.33 (35.10)	44.59 (42.09)	38.19 (67.42)	31.81 (27.49)
Benthic ash-free dry weight, AFDW (g/m^2)	300	422	168	13.29 (21.66)	5.77 (9.90)	22.53 (35.27)	19.96 (18.26)	17.67 (16.54)	13.67 (12.45)
Water temperature (°C)	*na*	1190	261	11.50 (3.57)	10.42 (3.20)	14.26 (3.38)	13.75 (3.10)	21.46 (4.23)	21.52 (5.00)
Flow (cfs)	2000	922	253	19.16 (65.38)	13.52 (18.60)	49.16 (157.82)	11.25 (17.73)	78.74 (182.16)	2.08 (4.72)
pH	*na*	1095	253	7.92 (0.68)	7.78 (0.62)	8.21 (0.48)	8.14 (0.34)	8.40 (0.53)	8.44 (0.54)
Hardness (mg/L)	2000	750	259	97.74 (89.58)	66.02 (54.46)	164.30 (92.16)	121.60 (67.13)	355.94 (280.80)	382.45 (306.84)
Specific conductance (µS/cm)	11000	1116	215	186.18 (173.65)	150.71 (212.14)	323.56 (268.41)	219.91 (111.52)	1616.84 (1473.60)	2088.10 (1284.90)
Total suspended solids (mg/L)	5000	886	262	4.70 (9.95)	1.06 (1.92)	7.67 (9.34)	3.84 (4.57)	60.10 (102.15)	50.04 (77.14)

2.4 Dataset summary

Following data curation, 1410 discrete samples from 983 wadeable streams and medium rivers were retained for data analysis. Of these, 319 represented reference sites. Nine metrics had 0 standard deviation across these samples and were removed, leaving 182 metrics from 6 categories: diversity, phylogeny, tolerance, functional feeding group, habit, and life history (voltinism). 1291 samples had an observation for at least one eutrophication indicator or water quality variable. For the 4 target eutrophication indicators, TP had the most observations (1067 samples), followed by TN (928), benthic chlorophyll *a* (732), and benthic AFDW (422). Reference sites had significantly lower TN and TP than test sites in the Mountains and Low Valleys and Transitional, but not in the Plains (Figure 2). Meanwhile, benthic chlorophyll *a* in reference sites was lower than that in test sites only in the Mountains, while benthic AFDW was lower in reference sites in the Mountains and Plains.

3.0 Correlation analysis

Correlations among water quality variables, metrics, and water quality variable-metric pairs were calculated to (1) select water quality variables that represented distinct gradients of local conditions, (2) identify groups of highly similar metrics to screen metrics during threshold analysis, and (3) identify metrics that responded strongly to eutrophication indicators.

3.1 Methods: correlation analysis

For each macroinvertebrate region, Spearman rank correlations with significance tests were calculated between each pair of log-transformed water quality variables, macroinvertebrate metrics, and metrics-water quality variables. Samples with missing values for any variable in a given pair were ignored. Correlations were calculated using 'cor.test' in the R *stats* package. For metric and metric-water quality correlations, clusters of highly correlated metrics were determined using k-means clustering to identify groups of both highly similar metrics and those metrics that responded similarly to water quality variables. For numbers of clusters between 2 and 10, the within-cluster sum of squares (WCSS) was calculated, which measures the variance within clusters. The optimal number of clusters was then determined as the point at which the rate of decrease in WCSS was lower than the average rate of change across the range of candidate numbers of clusters. This point represents the "elbow" at which adding more clusters does not substantially decrease WCSS, indicating a leveling off in the explained variance.

3.2 Results: correlation analysis

3.2.1 Water quality correlations

In general, eutrophication indicators were moderately correlated across regions, but none so strongly that they were considered to represent the same gradient (Figure 3, Table S2). TN and TP were positively correlated in each region ($\rho > 0.4$), with the strongest correlation in the Plains ($\rho = 0.72$). Benthic chlorophyll *a* was not significantly correlated with nutrients in any region, but was correlated with AFDW across regions (ρ between 0.48 and 0.59). Meanwhile, benthic AFDW was weakly, positively correlated with nutrients in the Mountains and Low Valleys and Transitional but more strongly and negatively correlated with nutrients in the Plains. With regards to other water quality variables, nutrients were moderately, positively correlated with

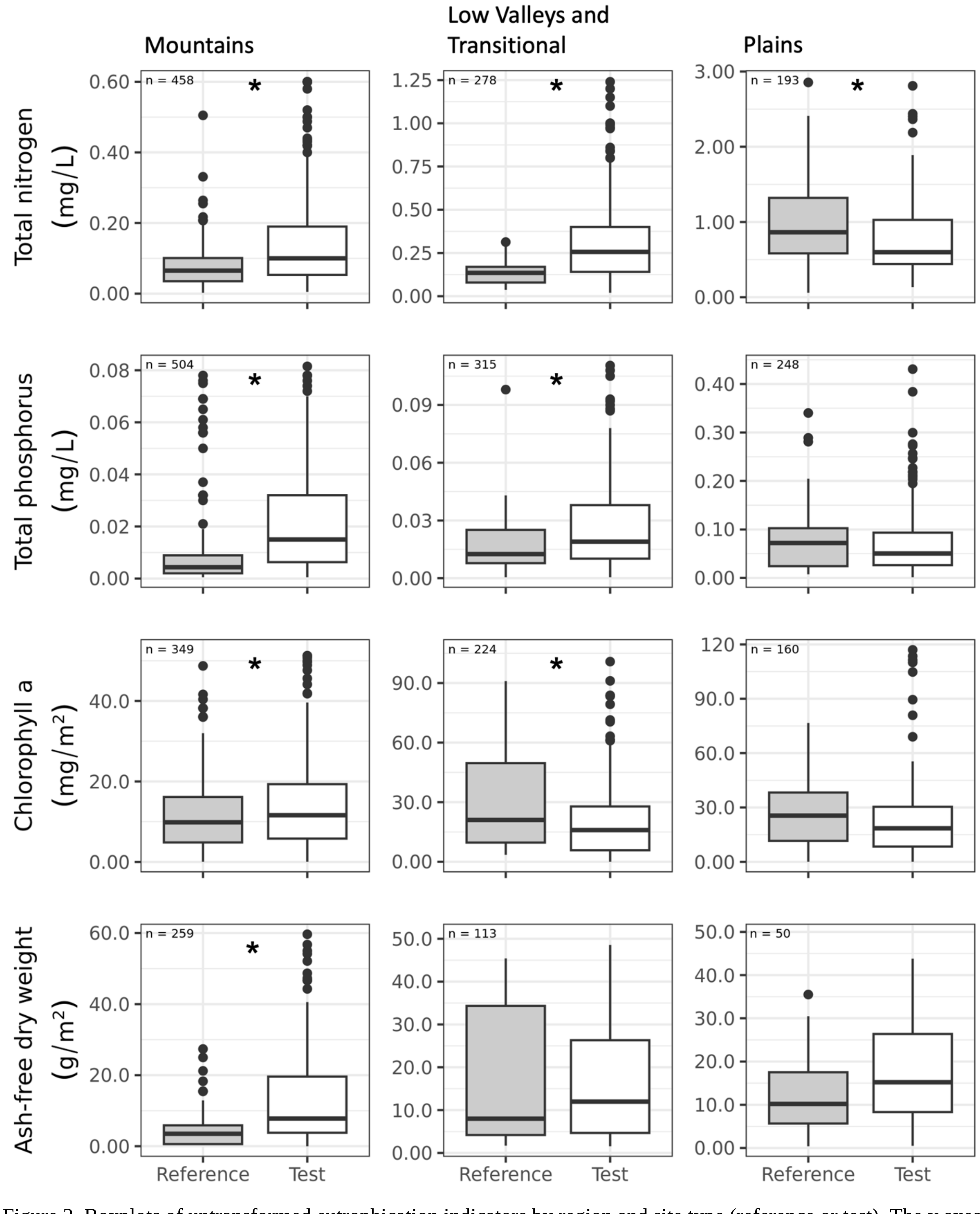


Figure 2. Boxplots of untransformed eutrophication indicators by region and site type (reference or test). The y axes extend only to the 95th percentile of observations to better visualize the contrast between reference and test sites. The n values in the upper left corner of each plot correspond to the number of samples with observations for the given eutrophication indicator-by-region. Asterisks indicate significant differences in means at $p < 0.05$ (Welch's t-tests).

specific conductance and total suspended solids in each region ($\rho > 0.3$). Among water quality variables, hardness and specific conductance were strongly, positively correlated – particularly in the Mountains and Low Valleys and Transitional. Since more observations were available for specific conductance across the dataset, hardness was removed from further analysis. Otherwise, few strong correlations were apparent among water quality variables.

3.2.2 Metric correlations

Spearman correlations among metrics were calculated to assess redundancy among metrics (Figure S2, Table S3). Variations of the same base metric (e.g., pi_EPT, pt_EPT, nt_EPT) were highly correlated regardless of region. In each region, highly correlated metrics were separated into 4 – 5 clusters. Each region contained clusters characterized by Ephemeroptera/Plecoptera/Trichoptera (EPT) taxa, intolerant taxa, Shannon diversity, and Beck's Biotic Index; proportion of dominant or tolerant taxa alongside the Hilsenhoff Biotic Index (HBI); non-insect taxa; and varying combinations of taxa, habits, and/or functional feeding groups. Importantly, correlations among metrics were not used to select metrics for further analysis, as a data-driven, all-metrics approach to threshold analysis was used to harness the power of the dataset. Correlations were used in later analyses to remove MMIs that contained highly correlated metrics. Given the number of comparisons, correlation matrices and heatmaps are provided as supplementary files.

3.2.3 Metric-water quality correlations

Overall, relationships between most metrics and water quality variables followed expected patterns based on historical responsiveness of metrics (Figure S3, Table S4). In general, highly correlated metrics in the metrics-only correlations clustered together in their relationships with water quality variables. In each region, water quality variables were generally clustered into three groupings: flow, total nitrogen and specific conductance with other miscellaneous variables, and temperature and benthic chlorophyll *a* with other miscellaneous variables (Table 3). Metrics that were most responsive to eutrophication indicators were EPT taxa, intolerant taxa, Beck's Biotic Index, and diversity indices (negatively correlated) and tolerant taxa, dominant taxa, and HBI (positively correlated).

Nevertheless, regions differed slightly in the specific groupings of variables and metrics, as well as in the strength of correlations. For example, in the Low Valleys and Transitional, metrics were most strongly correlated with flow, hardness, TN, and AFDW. Meanwhile, there were weaker correlations between EPT taxa and nutrients than in the Mountains. Similarly, there were fewer strong, positive relationships between metrics and nutrients in the Plains than in the Mountains or Low Valleys and Transitional, while metrics were more strongly related to flow and AFDW.

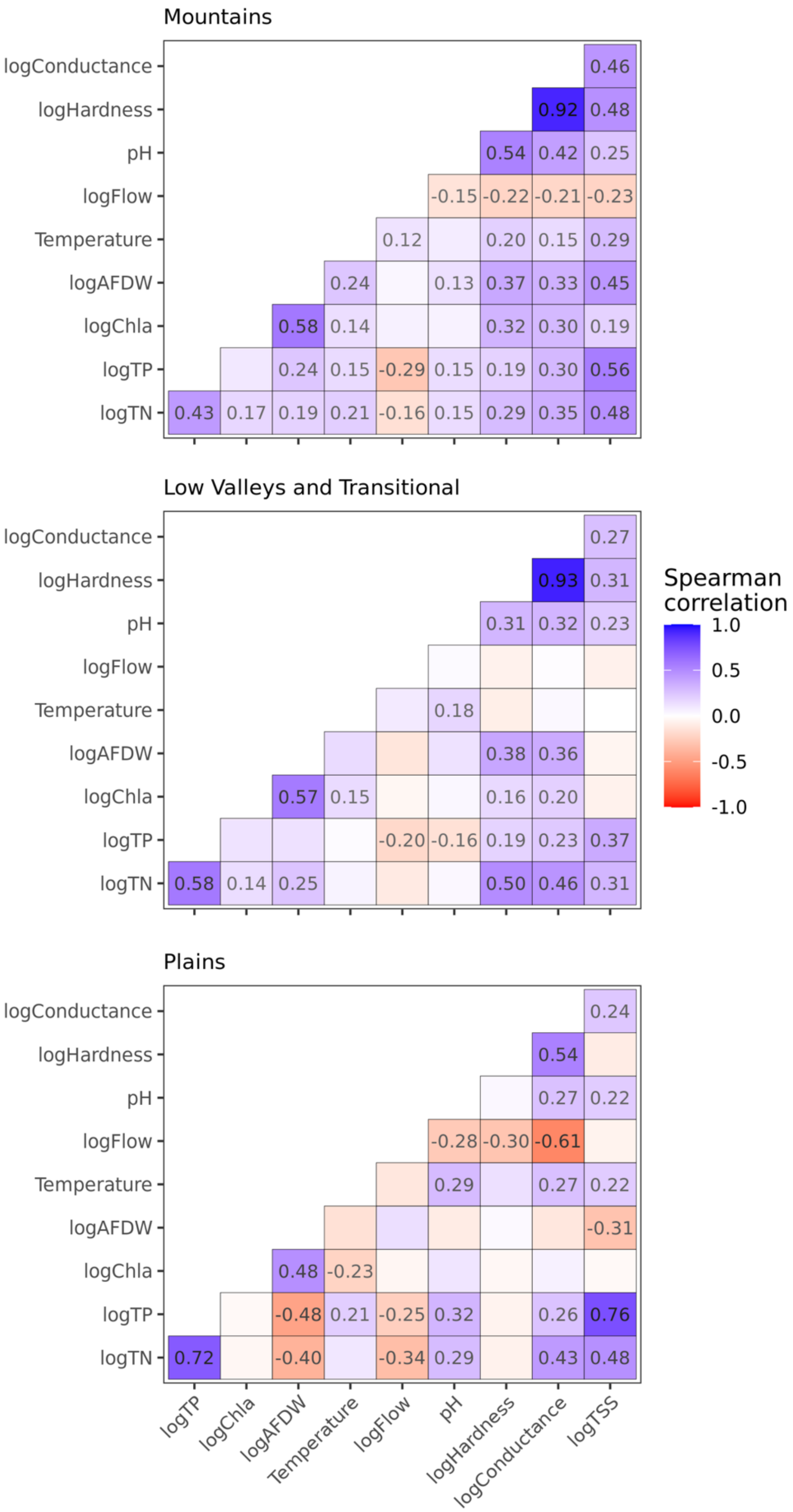


Figure 3. Spearman rank correlations among water quality variables by region. Significant correlations are shown with their correlation coefficients ($p < 0.05$).

Correlations between metrics and water quality variables were used as a preliminary screening of metric responsiveness to eutrophication indicators. However, linear and nonlinear modeling of all metrics was used as a more robust measure of these relationships. Correlation matrices and heatmaps are provided as supplementary files.

Table 3. General direction and strength of Spearman rank correlations between metric clusters and water quality variable clusters across regions. + refers to positive correlation, - to negative correlation. The number of symbols ranges from one (weak) to three (strong) to approximate the absolute strength of correlation.

Metric clusters	Flow	Total nitrogen/ Specific conductance/ Others	Temperature/ Chlorophyll *a*/ Others
Tolerant taxa, HBI	-	+++	+
Dominant taxa	-	++	+
EPT taxa, intolerant taxa, Becks, Shannon	++	- -	-
Total individuals, scrapers, omnivores, predators	+	-	-/+
Various groups, incl. non-insects, Hydropsychidae, Isopoda, Chironomidae, Coleoptera	-/+	+/-	+/-

4.0 Threshold analysis

4.1 Methods: threshold analysis

The goal of threshold analyses was to identify relationships between metrics and eutrophication indicators from which thresholds or change points could be determined using piecewise linear regression and/or nonlinear regression models. Analysis followed an all-metrics procedure with iterative model selection based on multiple lines of evidence. The next section provides an overview of the process. Subsequent sections describe each step in further detail.

4.1.1 Workflow

For each metric-eutrophication indicator pair, multiple models were computed - including linear regression, piecewise linear regressions (i.e., segmented regression), and nonlinear regressions. These models were univariable: that is, composed of a single metric as the response variable and a single eutrophication indicator as the explanatory variable. Rather than target a subset of metrics from correlation analyses, all metrics were considered in separate models. Models with multiple eutrophication indicators or water quality variables as explanatory variables (i.e., multiple regressions or multivariable models), were not considered because the focus was on determining thresholds associated with eutrophication. “Controlling” for variation in background variables at the outset can reduce the ability to detect relationships with target variables and reduce the sample dataset due to differential data collection. Potential independent effects of non-target variables like temperature or specific conductance were later accounted for via analysis of model residuals.

For each macroinvertebrate region, the top models of each eutrophication indicator were selected as those with the highest variation explained (R^2 values) and best model quality (Akaike Information Criterion, AIC). Across metrics, models from each of the four eutrophication indicators were compared, and the indicator with the highest variation explained was used to calculate thresholds of eutrophication impact. Thresholds were calculated as the regions of substantial change in the regression model. Each candidate threshold was validated by the distribution of reference sites.

Next, the extent to which other eutrophication indicators, background variables, and other stressors - together referred to as covariates - explained additional variation in the metric-eutrophication indicator relationships was assessed. This was done by calculating the residuals of the univariable model (i.e., the differences between observed metric values and the metric values predicted by the single eutrophication indicator model) and using the residuals as the response variable in univariable models that used each covariate as an explanatory variable.

4.1.2 Step 1: Selecting the strongest models between metrics and eutrophication indicators

First, the relationships between each metric and log-transformed eutrophication indicator (TN, TP, benthic chlorophyll *a*, and benthic AFDW) were characterized by six separate models (Figure 4):

- **Simple linear regression** - a straight line. If its R^2 was greater than or within 0.01 of another model, the relationship was considered linear, and no threshold could be determined. Calculated using ‘lm’ in *stats*.
- **Single breakpoint piecewise linear regression** - a “hockey stick” model with a single inflection point between two straight lines, each with different slopes. If its R^2 was within 0.05 of a nonlinear asymptotic or logistic regression, the nonlinear model was selected because of its relative simplicity of construction and interpretation. Calculated using ‘segmented’ in *segmented*, with npsi = 1, which forces the starting value of the breakpoint to be internally computed based on quantiles.
- **Double breakpoint piecewise linear regression** - a “broken stick” model with two inflection points between three straight lines, each

with different slopes. If its R^2 was within 0.05 of a nonlinear logistic regression, the logistic model was selected because of its relative simplicity of construction and interpretation. Calculated using 'segmented' in *segmented*, with npsi = 2, which forces the starting values of the breakpoint to be internally computed based on quantiles (Muggeo 2003).

- **Asymptotic regression** - a nonlinear model resembling a growth curve or exponential decay towards an asymptote. If its R^2 was within 0.05 of a logistic regression, the logistic model was selected because of its ability to characterize three potential thresholds (see below) instead of using the minimum or maximum value of the eutrophication indicator as a threshold. Calculated using 'SSasymp' in *stats*, a 'selfStart' model that internally calculates the starting values for model parameters (horizontal asymptote, response when input is 0, and natural log of the rate constant).
- **Four parameter logistic regression** - a nonlinear model resembling a sigmoid or S-shaped curve that has an upper and lower asymptote. If its R^2 was greater than that of linear regression and within 0.05 of any other model, this model was selected because of its ability to characterize three potential thresholds: initialization (change from asymptote to exponential change), maximum change (midpoint of curve representing linear change), and saturation (change from exponential change to another asymptote). Calculated using 'SSfpl' in *stats*, a 'selfStart' model that internally calculates the starting values for model parameters (left and right horizontal asymptotes, input value at the inflection point of the curve, and a numeric scale parameter).
- **Generalized additive models (GAMs)** - a nonlinear model that resembles a flexible, smooth curve that captures complex relationships. These models are superficially similar to the nonparametric locally weighted scatterplot smoothing (LOWESS) in that a "wiggly" line is fit to the relationship, but GAMs can be used to generate an R^2 value. GAMs were used to approximate the maximum amount of explicable variation between a metric and eutrophication indicator. If the R^2 of the GAM was greater than 0.05 of piecewise, asymptotic, and logistic regressions, the relationship was considered too complex for thresholds to be characterized, and the metric was removed from consideration for the given indicator. Given the complexity of GAMs relative to other modeling approaches, GAMs were not used to estimate

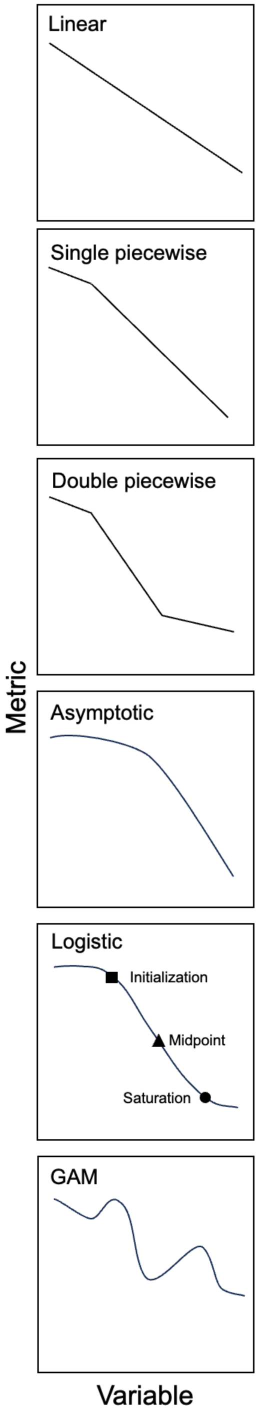


Figure 4. Conceptual plots of curves from linear and nonlinear models

potential thresholds of change. Calculated using ‘gam’ in *mgcv* with ‘family’ = Gaussian and the eutrophication indicator as a smooth term using the default number of knots, ‘k’ (Wood 2011).

For each eutrophication indicator, the 95th percentile of GAM R^2 values across metrics was used as the minimum R^2 required for a piecewise, asymptotic, and/or logistic model to be considered as sufficiently explanatory for the metric-indicator relationship. In each region, at least 10 metrics met this 95th percentile cutoff, so model selection for each metric proceeded.

For each region and each metric-by-indicator, the logistic model was selected as the most explanatory model (or had functionally equivalent explanatory power as other models).

4.1.3 Step 2: Selecting the most responsive eutrophication indicator

The second step was to determine which eutrophication indicator yielded the strongest relationship with candidate metrics. The 95th percentiles of R^2 values from logistic models for each eutrophication indicator were compared.

In each region, the relationships between metrics and TN were the strongest (i.e., the 95th percentile of R^2 for the top metrics and TN was greater than the 95th percentile of R^2 for the top metrics and TP, benthic chlorophyll *a*, or benthic AFDW). Therefore, TN was used as the explanatory variable in the initial models used to determine thresholds prior to modeling covariate relationships. For each region, all metric-TN logistic models with R^2 values within 75% of the top metric-TN logistic model R^2 were selected for further analysis.

4.1.4 Step 3: Determining metric and total nitrogen thresholds

For the third step, threshold values for the metric and TN were determined from logistic models. Three thresholds were estimated: initialization (the point of change from the first asymptote to exponential change), maximum change (the midpoint of the curve representing linear change), and saturation (the point of change from exponential change to the second asymptote). Initialization and saturation thresholds were calculated as the point on either side of the midpoint at which the slope was 50% of that at the midpoint.

Following consultation with DEQ and based on EPA guidance (EPA 2000), the distribution of reference sites was used to determine which, if any, threshold to set based on the model. If 75% of reference sites had TN concentrations below and metric values above (for metrics that decreased with TN) or below (for metrics that increased with TN) the initialization point of the curve, the initialization point would be the candidate threshold. If 75% of reference sites were between the initialization and midpoints of the curve, the midpoint of the curve would be the candidate threshold. If 75% of reference sites exceeded the midpoint, the metric was considered to be overly responsive and a poor indicator of the effects of eutrophication (i.e., reference sites were characterized by too high of TN and/or too low or high of metric values) (Figure 5). Additionally, the candidate threshold was considered to be the threshold value from the curve instead of the 75th reference site percentile point along the curve, because the 75th reference percentile value is being used primarily as a benchmark for threshold decision making and is more representative of the underlying dataset rather than the overall shape of the distribution.

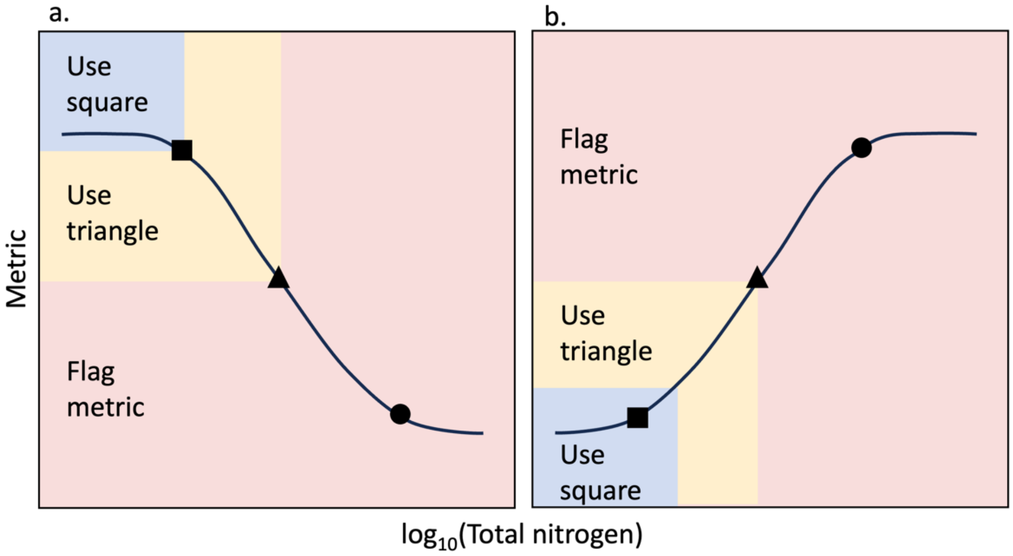


Figure 5. Conceptual plots of reference thresholds in logistic curves for (a) metrics that decrease with total nitrogen (i.e., high metric values are generally associated with good biological condition) and (b) metrics that increase with total nitrogen (i.e., low metric values are generally associated with good biological condition). If 75% of reference sites had metric and total nitrogen values in a given colored polygon (i.e., the point of intersection between hypothetical vertical and horizontal lines denoting the 75th percentile of reference sites for each axis), the denoted point of change in the curve was considered the candidate threshold point.

For illustrative purposes only, a top performing metric-TN model was selected from each region as a representative metric for which to report logistic model biplots, residual model biplots, and multimetric index (MMI) models.

4.1.5 Step 4: Estimating independent influences of other water quality variables

Following the calculation of region-specific thresholds in macroinvertebrate metrics relative to TN, the fourth step was to estimate additional variation of the metric that could be explained by other eutrophication indicators (TP, benthic chlorophyll *a*, and benthic AFDW) and other water quality variables (temperature, flow, pH, and specific conductance). To this end, the residuals of each metric-TN logistic model (i.e., the difference between observed metric values and predicted metric values) were calculated using 'residuals' in *stats*. These residuals were then used as the response variable in individual GAMs for each covariate (e.g., a GAM for residuals-by-TP, a GAM for residuals-by-AFDW, etc.). GAMs were used for this analysis because of their flexibility to model a variety of shapes between the residuals and covariates and, therefore, estimate the maximum amount of explicable variation. If the R^2 value of a residual-covariate

GAM was greater than 0.20, the covariate was considered to explain additional independent variation in the metric beyond that explained by TN alone.

4.1.6 Step 5: Comparing multimetric indices to single metric models

In addition to single metrics as the response variable in the initial metric-TN models, the extent to which MMIs increased the explanatory power of models over single metric models was tested. In general, MMIs operate on the principle that different metrics reflect different characteristics of the biological community, which in turn respond to different sources of water quality degradation. Therefore, MMIs are typically constructed and validated based on their ability to distinguish reference sites and disturbed sites, which are generally differentiated by a variety of stressors that represent general disturbance. Since the present analysis focused on the effects of eutrophication indicators on macroinvertebrate metrics, MMIs may have limited benefit over single metrics since multiple metrics must respond in complementary ways to only a single eutrophication indicator.

Nevertheless, MMI values were calculated using the methods of van Sickle (2010) - but MMI performance was not tested in the traditional way of differentiating reference and disturbed sites. Briefly, each metric was converted to a 0 - 10 scale, with values less than the 5th percentile set to 0 and values greater than the 95th percentile set to 10. For metrics with which reference sites had lower values, the metric was flipped (e.g., 10 became 0 and 0 became 10) so that all metrics shared the same scale and direction. It was expected that conducting region-specific analyses controlled for the strongest sources of variation in natural characteristics among sites. Therefore, so-called predictive MMIs were not generated, in which the influences of natural background variables like temperature, flow, or pH or landscape variables like watershed area, precipitation, soil lithology, and forest cover on a metric are “modeled out” (i.e., by using the residuals of a multivariable model between each metric and the landscape variables as the metric value).

For the representative metric of single metric models for each region, all 2- 4-metric combinations were determined regardless of metric category (e.g., Becks3 + nt_EPT, Becks3 + pt_ffg_pred, Becks3 + nt_EPT + pt_ffg_pred, etc.) - resulting in over 350000 MMI combinations for each region. For each of these MMIs, if the maximum correlation between scaled metrics was > 0.7 or < -0.7, the MMI was removed from consideration to reduce metric redundancy. For all other MMIs, scaled metric values were summed, divided by the number of metrics, and multiplied by 10 to get MMI scores that then spanned a 0 - 100 scale. Then, linear regressions, logistic regressions, and GAMs were calculated, with MMI scores as the response variable and TN (the top eutrophication indicator from single metric models) as the explanatory variable. The R^2, AIC, and TN threshold values from the top performing MMIs were compared to those of the top performing single metric models to determine if MMIs substantially increased the variation explained over single metric models and could be used to determine thresholds.

4.2 Results: threshold analysis

Detailed tables of model performance and logistic regression biplots for all top metrics are provided in supplementary files (Table S5, Figure S4).

4.2.1 Mountains

In the Mountains, 21 single metric models passed model selection and quality filtering, including removing logistic R^2 values less than 75% of that of the maximum. The 95th percentile of GAM R^2 values was 0.29, and logistic R^2 values ranged from 0.24 - 0.32. The maximum logistic R^2 was for the pt_tv_intol metric (0.32) and was greater than R^2 values of linear (0.26) and single breakpoint piecewise models (0.30) and comparable to double breakpoint piecewise (0.33) and GAM (0.34) values. Following the removal of redundant metrics (e.g., removing nt_tv_intol when pt_tv_intol had higher R^2), 8 metrics remained (Table 4). Of these, three metrics increased with TN (HBI, pt_tv_toler, pt_tv_stol).

Table 4. Top metrics and corresponding thresholds for the Mountains, arranged by logistic model R^2. Representative metric is bolded. Becks3 was selected as the representative metric instead of pt_tv_intol because it yielded comparable model performance and threshold values of TN and was also the top model in the Low Valleys and Transitional.

Metric	Description	Logistic R^2	Linear R^2	GAM R^2	TN threshold (mg/L)	Metric threshold
pt_tv_intol	Percent of intolerant taxa	0.32	0.26	0.34	0.155	42.16
Becks3	**Beck's Biotic Index v3**	**0.31**	**0.27**	**0.33**	**0.139**	**35.09**
nt_Pleco	Number of Plecoptera taxa	0.29	0.25	0.31	0.132	4.84
nt_EPT	Number of EPT taxa	0.28	0.24	0.29	0.139	18.13
HBI	Hilsenhoff Biotic Index	0.27	0.23	0.30	0.133	3.52
pt_tv_toler	Percent of tolerant taxa	0.26	0.21	0.29	0.159	12.31
nt_tv_ntol	Percent of mostly intolerant taxa	0.25	0.22	0.26	0.139	29.72
pt_tv_stol	Percent of semi-tolerant taxa	0.25	0.19	0.28	0.164	8.49

In each model except pt_tv_toler and pt_tv_stol, the intersection of the 75th percentile of TN (0.11 mg/L) and the 75th percentile of the metric was between the initialization point and midpoint of the curve (Figure 6). Therefore, based on the criteria discussed with DEQ, the TN and metric values at the midpoint of the curve represent the candidate threshold for these metrics. Across metrics, TN thresholds varied by no more than 0.032 mg/L.

Altogether, single metric logistic models in the Mountains meet all quality criteria and represent statistically viable and ecologically interpretable thresholds of eutrophication influences on macroinvertebrate condition. Since Becks3 was a top model in the Mountains and it is also the top model in the Low Valleys and Transitional (see *Section 4.2.2*), Becks3 was selected as the representative metric and model for the Mountains (Figure 6).

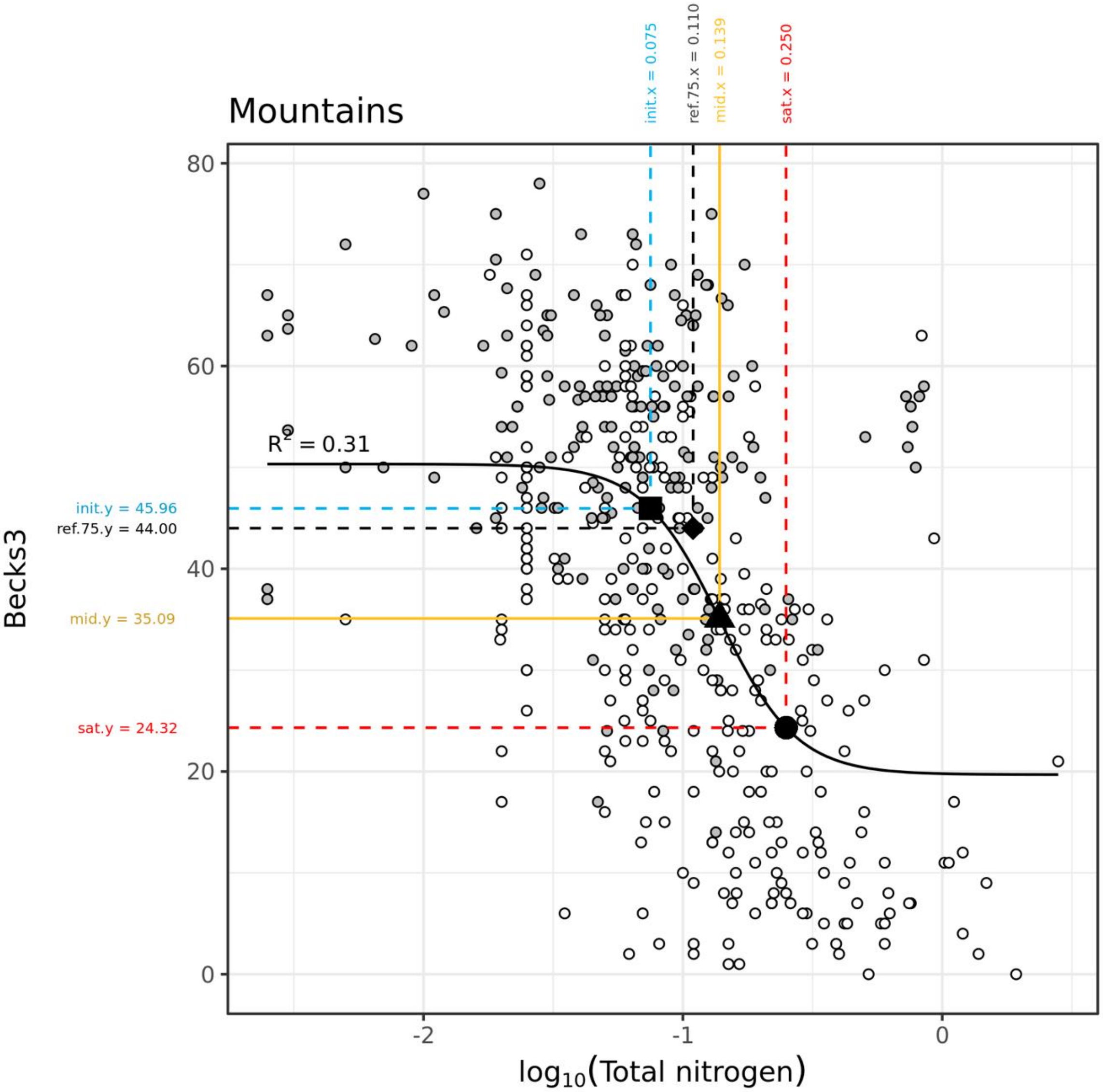


Figure 6. Biplot of reference (gray) and test (white) site values and logistic model curve for the representative single metric model for the Mountains: Becks3. init = initialization point, mid = midpoint, sat = saturation point, and ref.75 = 75th percentile of reference site values.

4.2.2 Low Valleys and Transitional

In the Low Valleys and Transitional, six single metric models passed model selection and quality filtering. The 95th percentile of GAM R^2 values was 0.19, and logistic R^2 values ranged from 0.21 - 0.26. The maximum logistic R^2 was for the Becks3 metric (0.26) and was greater than the linear R^2 (0.24) and comparable to the GAM R^2 (0.26). Following the removal of redundant metrics, three metrics remained - each of which decreased with increasing TN (Table 5).

For the top two models (Becks3 and nt_tv_intol), the midpoint of the curve represented the candidate threshold. For pt_Insect, the TN initialization value was the same as the 75th reference percentile of TN, thus making it unclear whether to select the initialization point or midpoint as the candidate threshold. To be conservative, the midpoint was selected as the candidate threshold for pt_Insect.

Table 5. Top metrics and corresponding thresholds for the Low Valleys and Transitional, arranged by logistic model R^2. Representative metric is bolded.

Metric	Description	Logistic R^2	Linear R^2	GAM R^2	TN threshold (mg/L)	Metric threshold
Becks3	**Beck's Biotic Index v3**	**0.26**	**0.24**	**0.25**	**0.199**	**18.68**
pt_Insect	Percent of insect taxa	0.21	0.18	0.22	0.300	84.22
nt_tv_intol4_EPT	Number of intolerant EPT taxa	0.21	0.19	0.22	0.238	10.64

Since Becks3 was the top model in the Low Valleys and Transitional and also a top model in the Mountains, Becks3 was selected as the representative metric and model for the Low Valleys and Transitional (Figure 7).

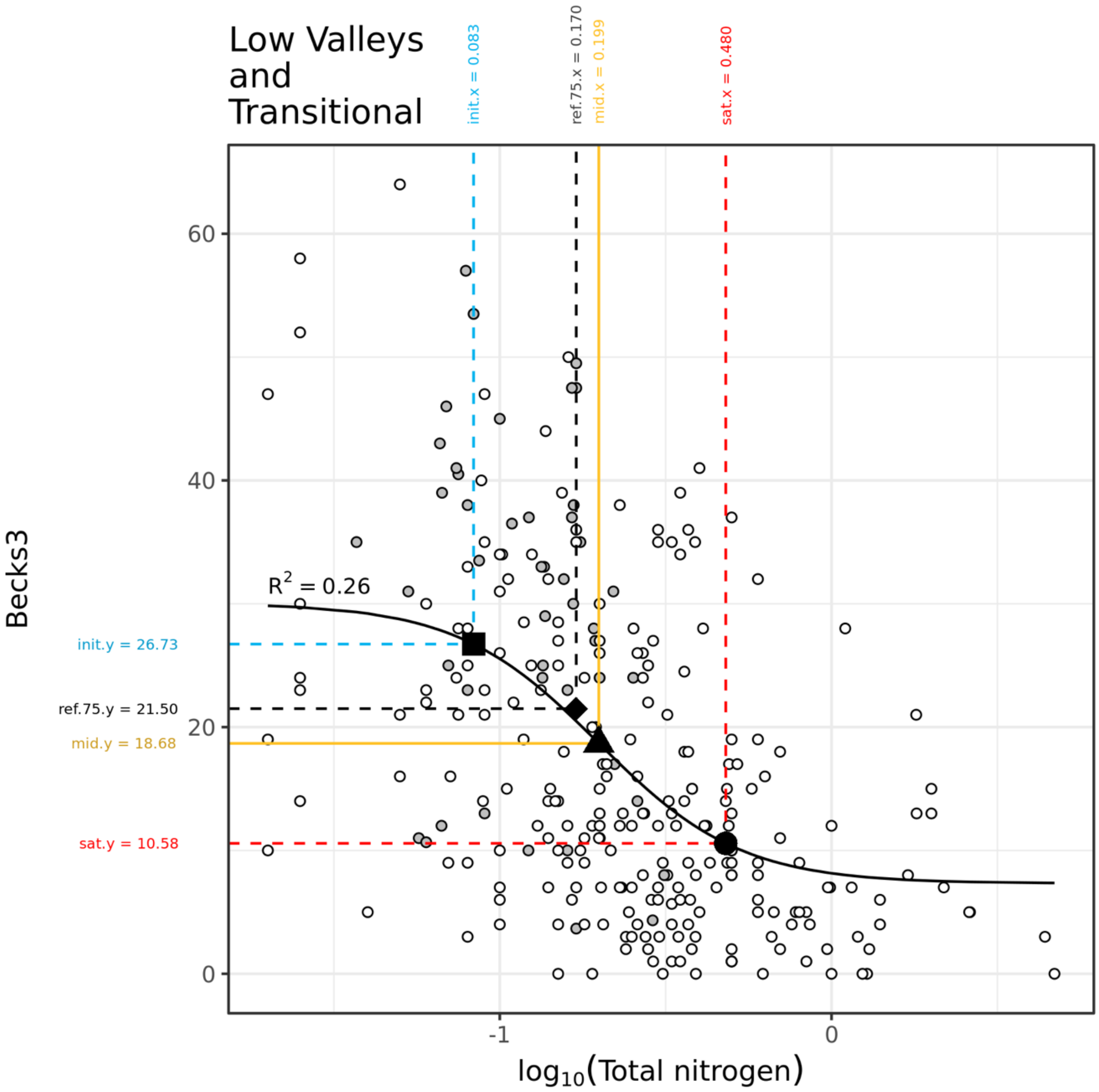


Figure 7. Biplot of reference (gray) and test (white) site values and logistic model curve for the representative single metric logistic model for the Mountains: Becks3.

4.2.3 Plains

In the Plains, no single metric models passed model selection and quality filtering because of at least one of the following: the 75th reference percentile of TN (1.47 mg/L) exceeded all candidate thresholds (initialization, midpoint, and saturation), the 25th or 75th percentile of the metric was above or below the logistic curve, and/or the logistic R^2 was less than the 95th percentile of GAM R^2 values (0.39) (Table 6). However, GAM R^2 values were inflated by a small number of metrics with very high R^2 caused by little variation in the metrics. Therefore, the better measure of model performance is likely the difference between logistic and GAM R^2 for a single model, and all but the top model met the previously defined criteria of the logistic R^2 being no more than 0.05 less than the GAM R^2. The distribution of reference sites, meanwhile, suggests that reference sites in the Plains represent site condition that is controlled by variables other than nutrients: more than 25% of reference sites had TN values greater than the midpoint of logistic curves, and reference sites with high TN had metric values indicative of poor condition. As seen in the boxplots of TN distributions in Figure 2, there was no difference in eutrophication indicators between reference and test sites in the Plains.

Table 6. Top metrics and corresponding thresholds for the Plains, arranged by logistic model R^2. Representative metric is bolded. Unlike the Mountains and Low Valleys and Transitional, both the midpoint and saturation point values are presented because the distribution of reference sites in the Plains exceeds even the saturation point in all models except nt_EPT, which is also why nt_EPT is selected as the representative metric.

Metric	Description	Logistic R^2	Linear R^2	GAM R^2	TN midpoint (mg/L)	TN saturation point (mg/L)	Metric midpoint	Metric saturation point
nt_ECT	Number of ECT taxa	0.34	0.27	0.39	0.885	1.300	8.45	5.48
nt_EPT	**Number of EPT taxa**	**0.32**	**0.28**	**0.37**	**0.937**	**1.490**	**6.29**	**3.18**
pi_tv_toler	Percent of tolerant individuals	0.31	0.28	0.33	0.835	1.240	43.27	58.01
pi_tv_stol	Percent of semi-tolerant individuals	0.30	0.27	0.33	0.791	1.100	39.99	52.44

In the most readily interpretable logistic model for the region (nt_EPT), the 25th percentile of nt_EPT values in reference sites was nt_EPT = 1, despite these sites ranging in TN from 0.58 mg/L to nearly 3.5 mg/L (Figure 8). If model performance alone is considered, nt_EPT, pi_tv_toler, and pi_tv_stol each had strong logistic relationships with TN, and candidate thresholds might be considered based on changepoints in the logistic curve without regard to reference site distributions. Given its common use in macroinvertebrate biomonitoring nationwide and straightforward interpretation, nt_EPT is presented as the representative model for the Plains. nt_EPT was also the only metric for which the saturation point of TN (1.49 mg/L) was slightly greater than the 75th percentile of reference TN (1.46 mg/L). While this value still invalidates the metric based on initial reference site criteria (i.e., the 75th percentile of reference TN must be below the midpoint TN, 0.94 mg/L), nt_EPT represents the top model for a threshold relationship between a macroinvertebrate metric and a eutrophication indicator in the Plains independent of reference site distributions.

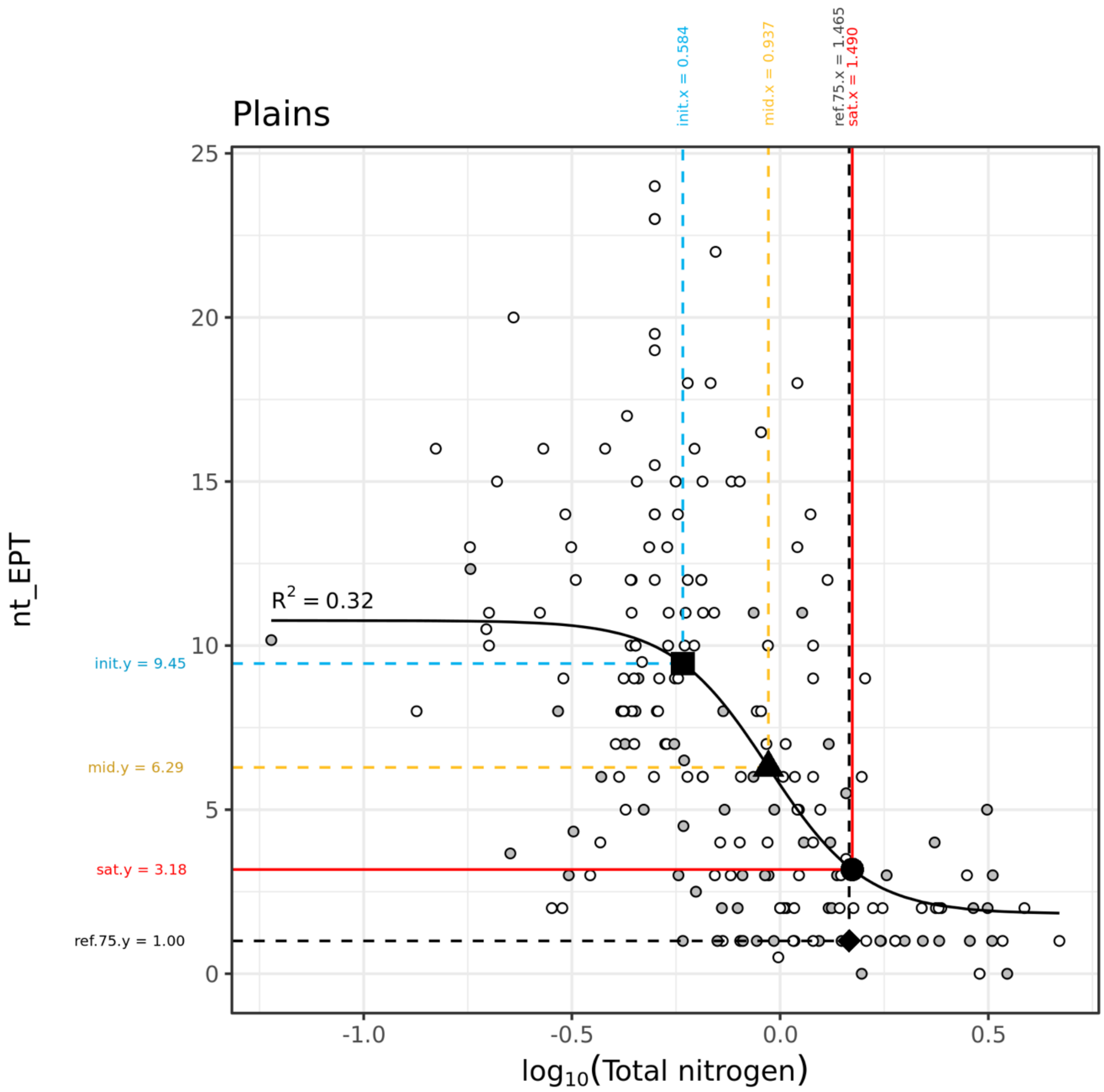


Figure 8. Biplot and curve for the representative single metric logistic model for the Plains: nt_EPT.

4.2.4 Residual influence of covariates

Since macroinvertebrate metrics might be sensitive to other variables beyond the influence of TN, the independent effects of other eutrophication indicators and water quality variables on each of the top metrics were examined. For the top single metric-TN logistic models for each region, the residuals of the metric were calculated. These residuals were then used as the response variable in individual GAMs in which the explanatory variable was each of the remaining eutrophication indicators and water quality variables.

In the Mountains, no covariates explained more than 20% of residual variation (Table 7). In both the Low Valleys and Transitional and Plains, residuals decreased with increasing specific conductance (R^2 = 0.26 and 0.25, respectively). Samples with specific conductance less than

Table 7. R^2 values for generalized additive models (GAMs) with the residuals of representative metric-by-log(TN) logistic models as the response and the water quality variable as the explanatory variable. For each variable, the approximate shape of the relationship is given as residuals decreasing with the variable (\), increasing (/), ∩-shaped, or no change (-).

Variable	Mountains (Becks3)	Low Valleys and Transitional (Becks3)	Plains (nt_EPT)
log(Total phosphorus)	0.12 \	0.00 \	0.07 ∩
log(Chlorophyll *a*)	0.08 /	0.02 ∩	0.15 \
log(Ash-free dry weight)	0.19 \	0.09 /	0.05 -
Temperature	0.12 \	0.04 \	0.05 \
log(Flow)	0.14 /	0.08 \	0.40 /
pH	0.18 ∩	0.08 \	0.02 \
log(Specific conductance)	0.17 \	0.27 \	0.25 \
log(Total suspended solids)	0.19 \	0.02 \	0.01 /

~200 μS/cm in the Low Valleys and Transitional and ~1500 μS/cm in the Plains had higher than expected Becks3 and nt_EPT values, respectively, than the threshold might indicate (Figure 9). Therefore, streams with higher specific conductance will likely have lower-than-expected metric values. For the Plains, residual nt_EPT was greater in streams with high flow, TN being equal (R^2 = 0.40). Therefore, samples from streams with higher flow are likely to have higher than expected nt_EPT values. From correlation analyses, specific conductance and flow were negatively correlated in the Plains (ρ = -0.61), indicating that sites with low specific conductance and high flow often co-occur.

Importantly, a weak relationship between residuals and covariates does not indicate no relationship between the metric and a given covariate, but rather no additional relationship to that between the metric and TN. For example, in each region, TN and TP were moderately correlated, and TP did not explain additional residual variation in metric scores. Therefore, when interpreting metric scores, TN and TP may both have causal effects. That is, metric scores may be influenced by changes in TN, TP, or a combination. This appears to be the case in each region, where logistic models between the representative metric and TP yielded similar patterns and metric thresholds as those with TN, despite weaker relationships with TP than those with TN (Figure 10).

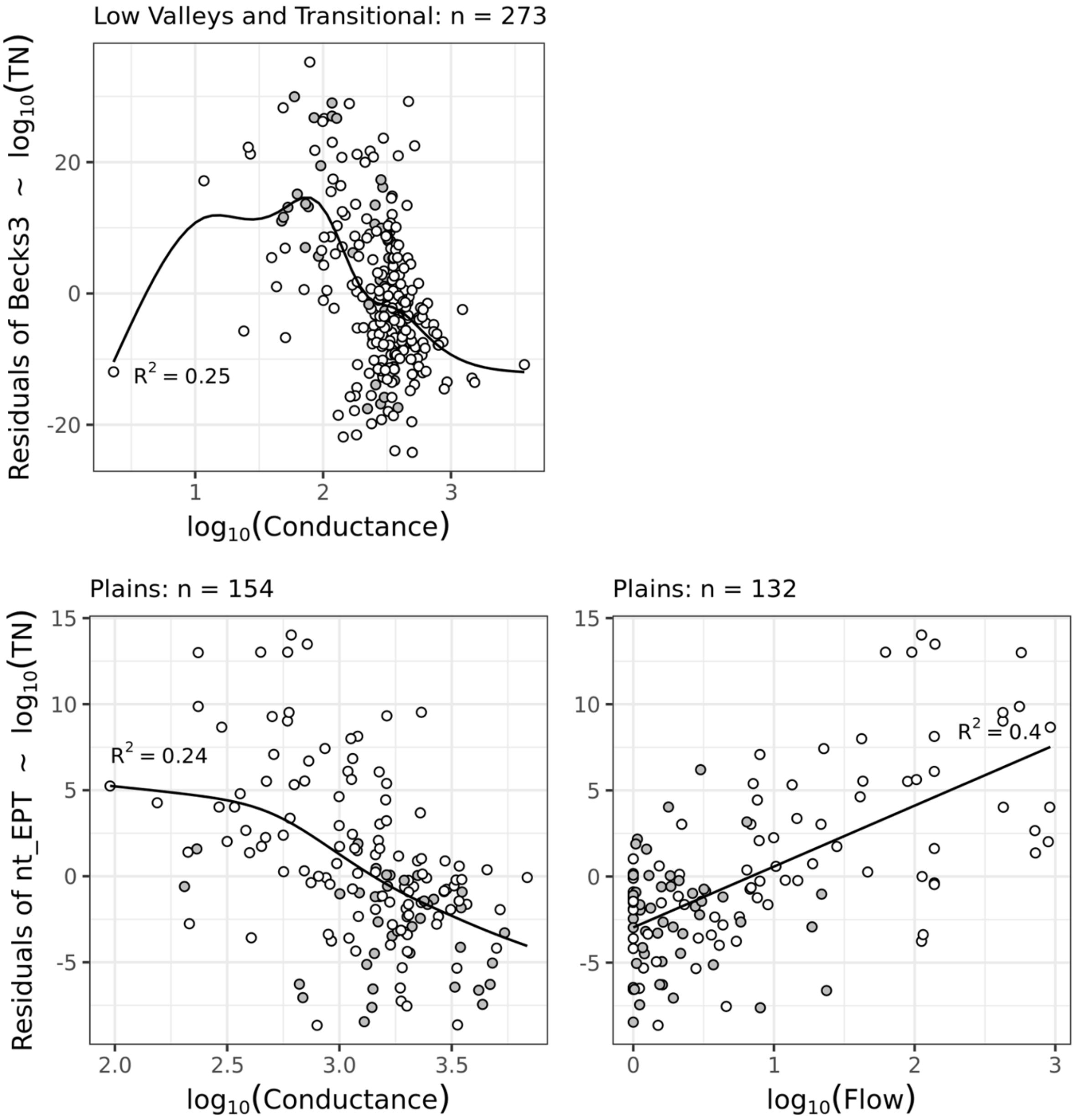


Figure 9. Biplots of residuals and covariates with GAM $R^2 > 0.2$ for each region.

A detailed table of residual model performance is provided in Table S6, GAM biplots between each pair of water quality variables in Figure S5, and biplots of water quality variables and residuals for all top metrics in Figure S6.

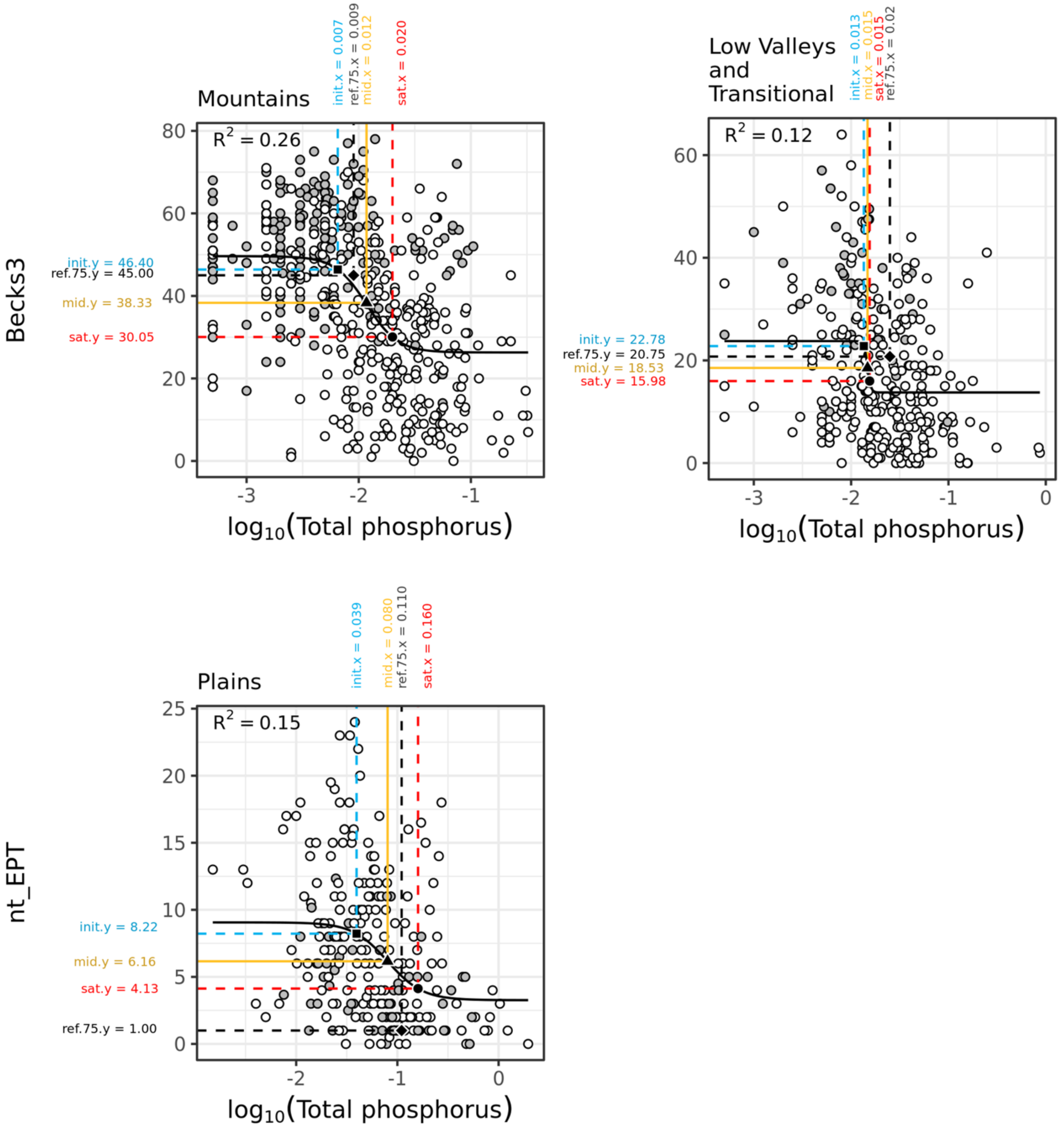


Figure 10. Biplots and curves for the logistic models of the representative metric for each region and total phosphorus.

4.2.5 Multimetric indices

Comparing the explanatory power of MMIs over representative single metrics, logistic model R^2 values in the Low Valleys and Transitional and Plains were 0.13 and 0.16 greater for the best MMIs, respectively (Table 8). Meanwhile, the best MMIs only marginally increased the R^2 by 0.05 in the Mountains over Becks3 alone. Since single metrics and MMIs that contain the single top metric of a region act as distinct metrics, even re-scaled thresholds in the single metrics cannot be directly compared to those of MMIs. Therefore, differences in TN thresholds are the best approximation of whether any increased explanatory power of MMI models affects candidate eutrophication thresholds. In each region the TN thresholds were similar between single metric and MMI models: single metric TN thresholds were 7% lower in the Mountains, 13% higher in the Low Valleys and Transitional, and 7% lower in the Plains.

Table 8. Comparison of top MMI logistic models to representative single metric models. MMIs were not selected for interpretability, though alternative metric combinations were similarly high performing. ^Threshold is initialization point. $^+$Threshold is saturation point.

Region	Single metric	Single metric R^2	Single metric TN threshold (mg/L)	MMI	MMI R^2	MMI TN threshold (mg/L)
Mountains	Becks3	0.31	0.139	Becks3 + nt_tv_toler + nt_volt_uni + pi_SimBtri	0.36	0.148
Low Valleys and Transitional	Becks3	0.26	0.199	Becks3 + li_total + pi_habit_cling_ PlecoNoCling + pi_tv_stol	0.39	0.175^
Plains	nt_EPT	0.32	1.490^+	nt_EPT + nt_habit_sprawl + pi_ffg_pred + pi_tv_stol	0.48	1.600^+

The present analysis shows that MMIs can have a higher percent of variation explained by logistic models than do single metrics, but modeled TN thresholds are not substantially altered. It can be noted that MMIs are arguably more difficult to interpret, as the complementary nature of each component metric is difficult to assess and may not necessarily explain more variation than would be expected by random chance.

A detailed table of MMI model performance for the top 10% of MMIs for each region is provided in Table S7.

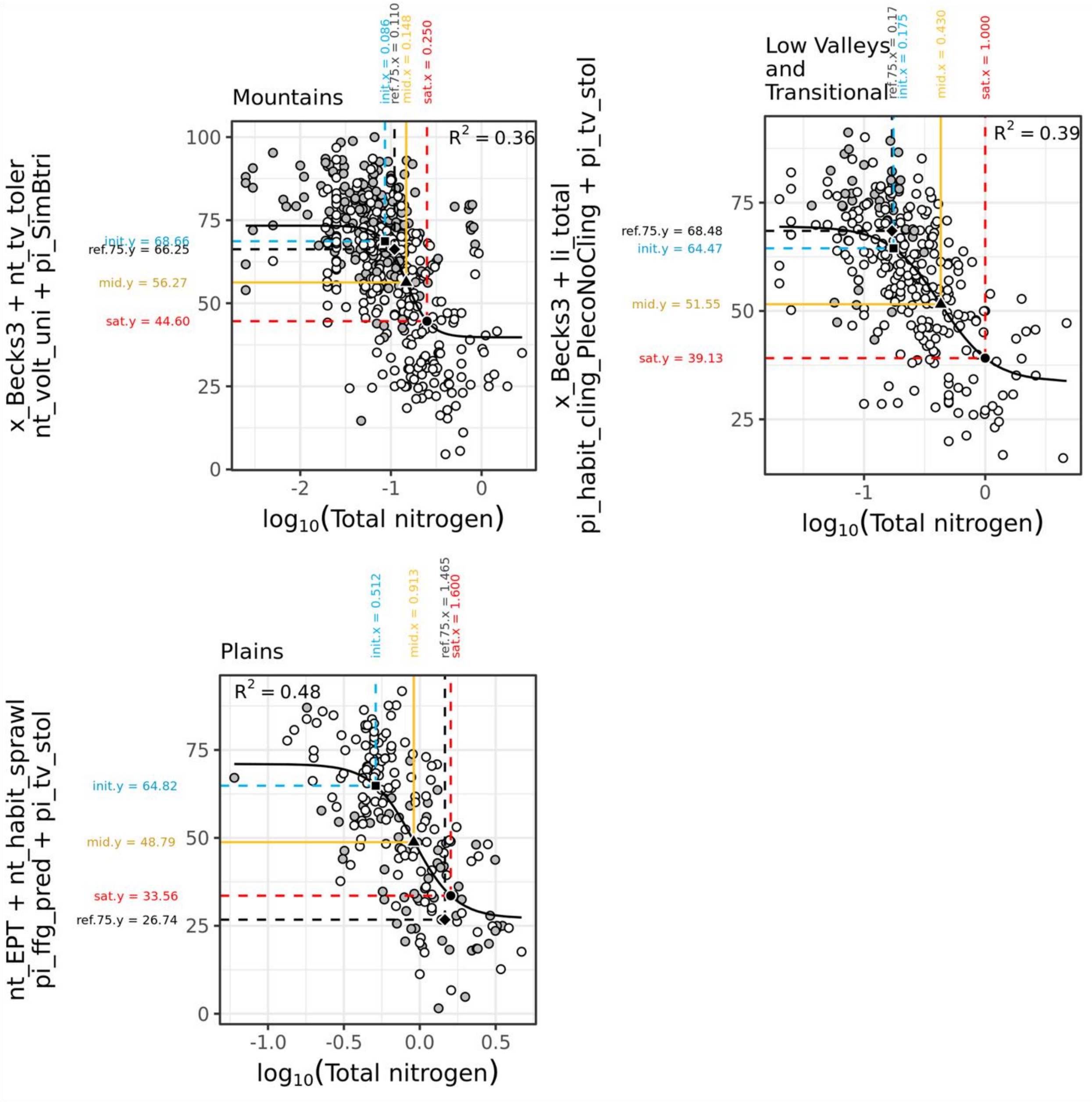


Figure 11. Biplots and curves for the logistic models of the top MMI for each region and total nitrogen.

5.0 Summary

Across three macroinvertebrate regions in the state of Montana, 1410 samples had macroinvertebrate metric data from 2005 to 2021, 1291 of which were associated with at least one water quality measurement. The present analysis revealed strong associations between metrics commonly linked to human disturbance and the eutrophication indicators of total nitrogen (TN), total phosphorus (TP), benthic algal chlorophyll *a*, and benthic algal ash-free dry weight (AFDW). Specifically, EPT taxa, intolerant taxa, Beck's Biotic Index, and diversity indices exhibited negative correlations, while tolerant taxa, dominant taxa, and HBI were positively correlated with these eutrophication indicators. In each region, metrics were more strongly associated with TN than with other eutrophication indicators.

To identify candidate thresholds of change in metrics relative to increasing TN, logistic nonlinear regressions were used to identify regions of change in each sigmoid, or S-shaped, metric-TN relationship. Representative metrics were selected from each region based on the model's explanatory power (R^2) as examples of candidate threshold selection. Becks3 – Beck's Biotic Index version 3, a weighted count of taxon-specific tolerance values whose values generally decrease with disturbance – was selected for the Mountains and Low Valleys and Transitional. The nt_EPT metric – the number of Ephemeroptera/Plecoptera/Trichoptera taxa, whose values generally decrease with disturbance – was selected in the Plains. In the Mountains, a Becks3 value of 35.09 corresponded to the point of maximum change at TN of 0.139 mg/L, which was greater than TN concentrations observed in 75% of Mountains reference sites. In the Low Valleys and Transitional, the point of maximum change in Becks3 was 18.68 at TN of 0.199 mg/L, which was also greater than that in 75% of the region's corresponding reference sites. In the Plains, a large number of reference sites had high TN and low nt_EPT. Ignoring the distribution of reference sites along the gradient of TN, a potential threshold of nt_EPT = 3.18 at TN of 1.490 mg/L could be identified in the sigmoidal relationship for the region.

In each region, neither TP, benthic chlorophyll *a*, nor benthic AFDW explained substantial variation in the observed metric values after accounting for TN. Nevertheless, while the thresholds herein were based on metric relationships with TN, TN and TP were moderately to strongly correlated to each other in each region, and logistic models between representative metrics and TP yielded similar patterns and thresholds to those between metrics and TN. Therefore, metric thresholds may reflect condition relative to TP as well as to TN, representing a general eutrophication effect. Additionally, in both the Low Valleys and Transitional and Plains, sites with increasing specific conductance exhibited lower than expected metric values suggesting an influence of conductance independent of TN on macroinvertebrate communities.

Finally, multiple metrics were combined into a single response variable, or multimetric index (MMI) for each region. Although some MMIs had greater explanatory power than single metrics in logistic regression models in the Low Valleys and Transitional and Plains, relationships between MMIs and TN did not strongly influence change points in TN over those identified by relationships with single metrics.

6.0 References

Bahls, L. L., Bukantis, B., & Tralles, S. 1992. Benchmark biology of Montana reference streams. Montana Department of Health and Environmental Science, Helena. December 1992.

Bowman, M. F., Chambers, P. A., & Schindler, D. W. 2007. Constraints on benthic algal response to nutrient addition in oligotrophic mountain rivers. *River Research and Applications*, *23*, 858-876.

Chambers, P. A., Meissner, R., Wrona, F. J., Rupp, H., Guhr, H., Seeger, J., ... & Brua, R. B. 2006. Changes in nutrient loading in an agricultural watershed and its effects on water quality and stream biota. *Hydrobiologia*, *556*, 399-415.

DEQ (Montana Department of Environmental Quality). 2021. Sample Collection and Laboratory Analysis of Chlorophyll-a Standard Operation Procedure WQPBWQM-011, Version 8.0. February 18, 2021.

EPA (U.S. Environmental Protection Agency). 2000. Nutrient criteria technical guidance manual: rivers and streams. United States Environmental Protection Agency, Office of Water and Office of Science and Technology, EPA 822-B-00-002.

Heiskary, S. A., & Bouchard Jr, R. W. 2015. Development of eutrophication criteria for Minnesota streams and rivers using multiple lines of evidence. *Freshwater Science*, *34*, 574-592.

Jessup, B., Hawkins, C., & Stribling, J. 2006. *Biological Indicators of Stream Condition in Montana using Benthic Macroinvertebrates*. Helena, MT: Montana Department of Environmental Quality. Oct 4, 2006.

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

Mazor, R. D., Sutula, M., Theroux, S., Beck, M., & Ode, P. R. 2022. Eutrophication thresholds associated with protection of biological integrity in California wadeable streams. *Ecological Indicators*, *142*, 109180.

Muggeo, V. M. 2003. Estimating regression models with unknown break-points. *Statistics in Medicine*, *22*, 3055-3071.

Poikane, S., Várbíró, G., Kelly, M. G., Birk, S., & Phillips, G. 2021. Estimating river nutrient concentrations consistent with good ecological condition: More stringent nutrient thresholds needed. *Ecological Indicators*, *121*, 107017.

R Core Team, A., & R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012.

Teply, M., & Bahls, L. 2007. Statistical evaluation of periphyton samples from Montana reference streams. Larix Systems Inc. and Hannaea. Helena, MT: Montana Department of Environmental Quality.

van Sickle, J. 2010. Correlated metrics yield multimetric indices with inferior performance. *Transactions of the American Fisheries Society*, *139*(6), 1802-1817.

Woods, A. J., Omernik, J. M., Nesser, J. A., Shelden, J., Comstock, J. A., & Azevedo, S. H. 2002. Ecoregions of Montana, 2nd edition (color poster with map, descriptive text, summary tables, and photographs).

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*, 3-36.

7.0 Appendix

Supplementary tables and figures are available as separate files. Below are the descriptions of each. All tables are in the file “supplementaryTables.xlsx”.

Table S1. Complete dataset of DEQ metadata, macroinvertebrate metric values, and water quality measurements.

Table S2. Spearman rank correlations among water quality variables, long format.

Table S3. Spearman rank correlations among macroinvertebrate metrics, long format.

Table S4. Spearman rank correlations between water quality variables and macroinvertebrate metrics, long format.

Table S5. Threshold analysis model results for all metrics-by-eutrophication indicators for each region.

Table S6. Residual analysis model results for top metrics and all non-TN water quality variables for each region.

Table S7. Multimetric index (MMI) analysis model results for all MMIs with logistic regression R2 within 10% of the top model for each region.

Figure S1. Histograms of untransformed and log10-transformed eutrophication indicators and water quality variables for each region. Available at “figS1_histograms.png”.

Figure S2. Heatmaps of macroinvertebrate metric Spearman correlations for each region. Available as 3 separate files in the folder “figS2_invertCorrelations”.

Figure S3. Heatmaps of macroinvertebrate metric-water quality variable Spearman correlations for each region. Available as 3 separate files in the folder “figS3_wqInvertCorrelations”.

Figure S4. Biplots with logistic regression curves between each of the top metrics and total nitrogen for each region. Available as multiple files in the folder “figS4_logisticPlots”.

Figure S5. Scatter plots with generalized additive model (GAM) curves between each water quality covariate and total nitrogen for each region. Available as multiple files in the folder “figS5_wqBiplots”.

Figure S6. Biplots with generalized additive model (GAM) curves between each water quality covariate and the residuals of all top metrics (from logistic models with total nitrogen) for each region. Available as multiple files in the folder “figS6_residualPlots”.