ECOLOGICAL APPLICATIONS

A model-based assessment of anthropogenic disturbance on lotic macroinvertebrate assemblages

Journal:	*Ecological Applications*
Manuscript ID	Draft
Wiley - Manuscript type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kopp, Darin; Oak Ridge Institute for Science and Education; United States Environmental Protection Agency, Office of Research and Development Stoddard, John; U.S. Environmental Protection Agency Kaufmann, Philip; Environmental Protection Agency, Office of Research and Development; Oregon State University, Fisheries & Wildlife Herlihy, Alan; Oregon State University, Department of Fisheries, Wildlife and Conservation Science Hill, Ryan; United States Environmental Protection Agency, Office of Research and Development Brehob, Meredith; Oak Ridge Institute for Science and Education Sabo, Robert; US EPA Office of Research and Development
Substantive Area:	Community Ecology < Substantive Area, Disturbance < Community Ecology < Substantive Area, Statistics and Modeling < Theory < Substantive Area, Conservation < Landscape < Substantive Area
Organism:	Invertebrates < Animals, Insects < Arthropods < Invertebrates < Animals
Habitat:	Freshwater < Aquatic Habitat < Habitat
Geographic Area:	United States < North America < Geographic Area
Key words/phrases:	Reference Condition, Biological Assessment, Biodiversity, Macroinvertebrates, Joint Species Distribution Modeling
Abstract:	Traditionally, the effects of anthropogenic disturbance on biological assemblages are elucidated by comparing an assemblage observed at a site to one that represents a minimally disturbed state. Unfortunately, defining a minimally disturbed state is extremely challenging because of the extent of human disturbance. We use a national scale dataset and a two-stage model-based approach to assess how benthic macroinvertebrate assemblages at 1,748 sites would change if common anthropogenic disturbances were removed from in-stream physiochemical variables. First, we used random forest models and current landscape data to predict physiochemical conditions and then infer abiotic condition in the absence of disturbance. Second, we combined these estimates with joint species distribution models to predict the assemblage that would

occur in these undisturbed conditions. Random forest models explained 48 – 75% of the variation in total nitrogen, phosphorous, sulfate, chloride, and substrate diameter. Generally, nutrient and salinity concentrations were higher, and substrates were finer than predicted to be without disturbances. Using this physiochemical data, joint species distribution models accurately explained genus richness (R2 = 0.73 – 0.85) and composition (Jaccard similarity index = 0.48 – 0.55). Depending on the ecoregion, we found that genus richness could change at 26 – 61% of sites if disturbance was removed. Different responses were observed for insect and non-insect taxa. For example, under anthropogenic disturbance, occurrence probabilities for Ephemeroptera, Plecoptera and Trichoptera tended to decrease at 5 – 26% of sites while occurrence probabilities for Mollusca and other non-insect, non-arthropod taxa increased at 5 – 33% and 11 – 24% of sites, respectively. Importantly, our framework provides an avenue for evaluating the effects of anthropogenic disturbance on macroinvertebrate assemblages without identifying reference sites.

A model-based assessment of anthropogenic disturbance on lotic macroinvertebrate assemblages

Darin A. Kopp[1*], John L. Stoddard[1], Philip R. Kaufmann[1,2], Alan T. Herlihy[2], Ryan A. Hill[1], Meredith M. Brehob[3], Robert D Sabo[4]

[1]US EPA, Office of Research and Development, Corvallis, Oregon, USA

[2]Department of Fisheries, Wildlife, & Conservation Sciences, Oregon State University, Corvallis, Oregon, USA

[3]Oak Ridge Institute for Science and Education Participant c/o US EPA, ORD, Washington, D.C., USA

[4]US EPA, Office of Research and Development, Center for Public Health and Environmental Assessment, Washington, DC, USA

*Correspondence: kopp.darin@epa.gov

Keywords: Joint Species Distribution Modeling, Richness, National Rivers and Streams Assessment, Benthic Macroinvertebrate, Reference Sites, National Aquatic Resource Surveys.

Open research statement: Benthic macroinvertebrate occurrence data and physiochemical variables were retrieved from the 2018-2019 National Rivers and Stream Assessment (https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys) Benthic macroinvertebrate data can be downloaded as .csv file from NRSA 1819 Benthic Macroinvertebrate Count - Data (CSV) (csv). Physiochemical data can be downloaded as .csv files from NRSA 1819 Water Chemistry_CHLA - Data (CSV) (csv) and NRSA 1819 Physical Habitat Larger Set of Metrics - Data (CSV) (csv) Landscape variables were retrieved from https://www.epa.gov/national-aquatic-resource-surveys/streamcat-metrics-and-definitions and PRISM Climate Group https://prism.oregonstate.edu/

Abstract:

Traditionally, the effects of anthropogenic disturbance on biological assemblages are elucidated by comparing an assemblage observed at a site to one that represents a minimally disturbed state. Unfortunately, defining a minimally disturbed state is extremely challenging because of the extent of human disturbance. We use a national scale dataset and a two-stage model-based approach to assess how benthic macroinvertebrate assemblages at 1,748 sites would change if common anthropogenic disturbances were removed from in-stream physiochemical variables. First, we used random forest models and current landscape data to predict physiochemical conditions and then infer abiotic condition in the absence of disturbance. Second, we combined these estimates with joint species distribution models to predict the assemblage that would occur in these undisturbed conditions. Random forest models explained 48 – 75% of the variation in total nitrogen, phosphorous, sulfate, chloride, and substrate diameter. Generally, nutrient and salinity concentrations were higher, and substrates were finer than predicted to be without disturbances. Using this physiochemical data, joint species distribution models accurately explained genus richness (R^2 = 0.73 – 0.85) and composition (Jaccard similarity index = 0.48 – 0.55). Depending on the ecoregion, we found that genus richness could change at 26 – 61% of sites if disturbance was removed. Different responses were observed for insect and non-insect taxa. For example, under anthropogenic disturbance, occurrence probabilities for Ephemeroptera, Plecoptera and Trichoptera tended to decrease at 5 – 26% of sites while occurrence probabilities for Mollusca and other non-insect, non-arthropod taxa increased at 5 – 33% and 11 – 24% of sites, respectively. Importantly, our framework provides an avenue for evaluating the effects of anthropogenic disturbance on macroinvertebrate assemblages without identifying reference sites.

Introduction:

Quantifying the extent and magnitude of anthropogenic disturbance on ecosystems requires a benchmark that represents a desired, expected, or previous condition (Stoddard et al. 2006, Hawkins et al. 2010, McNellie et al. 2020). Ideally, benchmarks should represent chemical, physical, and biological characteristics with minimal anthropogenic disturbance (Reynoldson et al. 1997, Stoddard et al. 2006). However, for many ecosystems the minimally disturbed condition is difficult to quantify or may no longer exist (Dudgeon et al. 2006, Vörösmarty et al. 2010, McNellie et al. 2020). The concept of "least disturbed" describes conditions that are derived from a collection of sites that represent the best available, given today's landscape (Stoddard et al. 2006). Unfortunately, because anthropogenic disturbance is not evenly distributed across the landscape, sites in least disturbed condition can be spatially biased and/or vary in quality (Herlihy et al. 2008, McNellie et al. 2020, Yuan et al. 2024). If least disturbed sites are spatially biased, their ecological, chemical, and physical conditions may be unattainable for some locations that need to be assessed because of differences in natural setting or biogeographic history. Alternatively, if reference site quality is lower for some ecosystems because of a relatively long legacy of human intensification, some sites will be judged against a lower benchmark and erroneously appear to be in better condition relative to others. Although least disturbed reference sites provides reasonable and defensible benchmarks for assessing biological condition (Stoddard et al. 2006, Herlihy et al. 2008, Mitchell et al. 2025), they are typically not undisturbed, nor evenly distributed among all ecosystem types (McNellie et al. 2020, Yuan et al. 2024). Novel approaches and concepts are urgently needed to overcome these limitations.

Quantifying taxon-environment relationships for all taxa within a region with empirical models may provide an alternative to relying on reference sites because they elucidate how individual taxa and entire assemblages change along physiochemical gradients (Chessman and Royal 2004, Kilgour and Stanfield 2005, Elias et al. 2016, Kopp et al. 2023). Because anthropogenic disturbance typically manifests as altered physiochemical conditions (Tang et al. 2020, Fergus et al. 2023), quantifying potential environmental divers across space, could be relevant for understanding how organisms may change with disturbance across time (Blüthgen et al. 2022). For example, taxa that have a negative relationship along an environmental gradient could be expected to become less prevalent if human disturbance contributes to more extreme environmental conditions and more prevalent if reducing human disturbance contributes to less extreme conditions (Kopp et al. 2023). From these relationships it is possible to predict which taxa from the regional species pool could occur at a site if it were minimally disturbed and compare them to the taxa that were actually observed (Chessman and Royal 2004, Elias et al. 2016). If the minimally disturbed assemblage is similar to the observed assemblage, then the site may be considered unaffected by anthropogenic disturbances.

Advances in species distribution modeling enhance our ability to quantify taxon-environment relationships (Franklin 2010, Guisan et al. 2017, Ovaskainen and Abrego 2020). Species distribution modeling consists of a diverse set of methods that typically focus on quantifying taxon-environment relationships for individual species (Elith and Leathwick 2009). Joint species distribution modeling (JSDM) is a recent multivariate extension of single species distribution modeling, capable of estimating taxon-environment relationships for all members of an assemblage simultaneously. Because of their complex structure, JSDMs are often fitted using Bayesian inference. As such, the resulting posterior distribution can be used for hypothesis testing (Johnson et al. 2022). For example, given a posterior distribution of values representing a biological assemblage (e.g. taxonomic richness) in the absence of human disturbance, it is possible to evaluate whether an observed assemblage is consistent with this distribution. In addition, these models use latent factors to account for unmeasured environmental variables and potential biotic interactions (Warton et al. 2015, Ovaskainen et al. 2017, Ovaskainen and Abrego 2020). This feature is particularly important for biological assessments because it can strengthen the inference gained from evaluating a specific suite of environmental variables that are typically altered by human disturbance. Thus, the ability to model all taxa in an assemblage and probabilistically evaluate whether an observation is consistent with a posterior distribution makes JSDMs well suited for understanding how anthropogenic factors affect biological assemblages.

Although taxon-environment relationships are necessary to circumvent the need for reference sites, they require a concurrent expectation of undisturbed physiochemical conditions to fully understand the effects of anthropogenic activities. Traditionally, undisturbed physiochemical conditions are obtained, or modeled, from least-disturbed reference sites (Olson and Hawkins 2012, Olson and Cormier 2019). However, empirical models have also linked changes in the physiochemical environment to anthropogenic activities and been used to estimate conditions if anthropogenic disturbance is reduced or eliminated without reference sites (Dodds and Oakes 2004, Herlihy and Sifneos 2008, Soranno et al. 2011). Random forest models are a machine learning algorithm that have superior predictive performance compared to other modeling techniques (Prasad et al. 2006, Peters et al. 2007) and are often used to model water quality parameters and physical habitat characteristics (Olson and Hawkins 2013, Olson and Cormier 2019, Sabo et al. 2023, Yuan et al. 2024). Further these models are more robust to nonlinearities, multicollinearity, and overfitting (Breiman 2001) and can evaluate the relative importance of different predictor variables (Lin et al. 2021, Sabo et al. 2023). Because these algorithms can accommodate complex relationships and a relatively high number of covariates, they are well suited for evaluating the relative importance of natural and anthropogenic variables on physiochemical conditions and, in turn, predicting the abiotic conditions if anthropogenic disturbances were removed (Yuan et al. 2024). Indeed, efforts that combined estimates of undisturbed physiochemical conditions with biological models have made progress towards addressing the limitations of reference site-based approaches (Chessman and Royal 2004, Kilgour and Stanfield 2005, Elias et al. 2016, Yuan et al. 2024).

In rivers and streams, assessments of biological integrity often rely on surveys of benthic macroinvertebrate assemblages because they are diverse, relatively easy to sample, and respond to changes in the physiochemical environment associated with anthropogenic disturbance (Hughes and Peck 2008, Buss et al. 2014). To elucidate the effects of anthropogenic activities on biotic integrity, benthic macroinvertebrate assemblages observed at a site are typically compared to reference assemblages collected from least disturbed sites (Hawkins 2006, Stoddard et al. 2008). Here, we develop a model-based assessment as a complementary method to evaluate how benthic macroinvertebrate assemblages at 1,748 stream sites may change if the effects of major anthropogenic disturbances were removed from the physiochemical environment. First, we use random forest models to relate in-stream physiochemical conditions to a suite of geoclimatic and anthropogenic factors and use these models to infer abiotic conditions if anthropogenic disturbance was eliminated. Second, we combined these estimates with JSDMs to predict how biological assemblages could change if physiochemical conditions were not altered by anthropogenic disturbance. Because JSDMs can leverage latent factors to account for unmeasured environmental factors, including potential biotic interactions (Ovaskainen et al. 2016), we focus exclusively on environmental variables that are commonly associated with anthropogenic disturbances. Specifically, we evaluate the effects of changing nutrient and salinity concentrations, physical habitat, and climate on benthic macroinvertebrate assemblages because these gradients are commonly used as abiotic screens used to identify least disturbed reference sites (Herlihy et al. 2008, Paulsen et al. 2008, USEPA 2023, Mitchell et al. 2025) and affect benthic macroinvertebrate distributions (Kopp et al. 2023).

Methods:

For our analysis, we focused on biological and environmental data collected at sites distributed among 9 ecoregions (Figure 1) and surveyed during the 2018/2019 National Rivers and Streams Assessment (NRSA). NRSA is a collaboration between the United States Environmental Protection Agency (USEPA) and state, tribal, and federal partners to assess the chemical, physical, and biological condition of streams and rivers of the United States (https://www.epa.gov/national-aquatic-resource-surveys/nrsa). Every five years, beginning in 2008/2009, NRSA surveys ~2,000 stream/river locations that are selected using a probabilistic survey design and handpicked by state and tribal partners. These sites are surveyed for biological, chemical, and physical characteristics using standardized field protocols (USEPA 2023). Only sites included in the probabilistic sample are selected to be representative of the entire population of streams and rivers in the contiguous US and used for assessment (Olsen and Peck 2008). In addition, approximately 10% of the probabilistic sites are revisited to assess within-year variability and approximately 30% of these sites are resampled in the next survey cycle to evaluate among-year variability.

Benthic macroinvertebrate assemblages and physiochemical variables

Procedures for collecting benthic macroinvertebrates are described in detail elsewhere (Hughes and Peck 2008, USEPA 2017a). Briefly, each survey location was defined as a length of stream or river equal to a multiple of its channel width. In wadeable and boatable sites, the reach length sampled was equal to 40 channel widths, a minimum of 150 m, or a maximum of 4km. Macroinvertebrate samples were collected along 11 equally spaced cross-section transects along the reach. In wadeable sites, samples were collected in an alternating left, center, right order along the transects using a D-frame kick net (500-um mesh, 0.09 m^2 area). In boatable sites, samples were collected along the left or right wadeable margin from a 1m linear sweep of the dominant habitat using a D-frame kick net (500-um mesh, 0.3048 m^2 area). Samples from each survey location were combined into a single composite sample, preserved in ethanol, and sent to a taxonomist for subsampling and identification.

To assess the effects of anthropogenic disturbance on in-stream physiochemical conditions and in turn macroinvertebrate assemblages, we focused on in situ nutrient concentrations, salinity, and physical habitat variables. These environmental gradients are commonly altered by anthropogenic activities in the surrounding watershed and frequently used as abiotic screens to identify least-disturbed reference sites (Herlihy et al. 2008, Paulsen et al. 2008, USEPA 2023). Specifically, we selected total nitrogen (NTL) and total phosphorus (PTL) as indicators of excess nutrients supplied from agriculture and/or urbanization (Herlihy et al. 1998) and chloride (CL) and sulfate (SO4) as indicators of salinity. Elevated CL is a general indicator of human disturbance in the catchment (Herlihy et al. 1998) and SO4 can indicate sites that are affected by mine drainage (Herlihy et al. 1991). Mean substrate diameter (SUBD) and riparian disturbance index (RPDI) were selected as indicators of physical habitat because human activities can increase fine sediment inputs or directly modify the riparian area (Kaufmann 1999, Kaufmann et al. 2022b). In addition, we included mean summer air temperature (Mean July and August temperature, MSAT) and total annual precipitation (TPRCP) to assess potential changes associated with differences from a 1900-1950 baseline. Although other factors are affected by human disturbances (e.g., metals and pesticides and hydrologic alteration), these variables provide a reasonably broad characterization of the physiochemical environment at a site (Herlihy et al. 2008, Paulsen et al. 2008)

During the survey, a single water sample was collected at each site and shipped to a central analytical laboratory. PTL and NTL were measured by persulfate digestion and colorimetry and CL and SO4 were measured by ion chromatography (USEPA 2017b). SUBD is the geometric mean of the numeric value assigned to substrate size classes measured in the field (Kaufmann 1999). RPDI summarizes the presence/absence of 11 categories of human disturbance, including buildings, landfill/ trash, logging, mining, developed parks or lawns, pavement or cleared lots, pipes (withdrawal or wastewater), roads, row crops, pastures or hayfields, and walls or revetments in the riparian area, adjacent to each to the 11 transects where macroinvertebrates were collected (Kaufmann 1999, USEPA 2017a, Kaufmann et al. 2022b). In addition, we included mean summer air temperature (Mean July and August temperature, MSAT) and total annual precipitation (TPRCP) for either 2018 or 2019 (depending on sample year) obtained from the PRISM Climate Group (https://prism.oregonstate.edu/).

Modeling genus-environment relationships

We quantified relationships between macroinvertebrate assemblages and eight physiochemical variables (i.e. NTL, PTL, CL, SO4, SUBD, RPDI, MSAT, and TPRCP) using joint species distribution models (JSDMs) fitted with the Hierarchical Modeling of Species Communities R package (Ovaskainen and Abrego 2020, Tikhonov et al. 2020). JSDMs are a multivariate hierarchal generalized linear mixed model fitted with Bayesian inference. They are uniquely suited to evaluate relationships between anthropogenic disturbance and biological assemblages because they are multi-species models that quantify taxon-environment relationships for all members in an assemblage simultaneously and account for unmeasured variables, including abiotic factors and species associations, using random effects specified at the sample-level (Warton et al. 2015, Ovaskainen and Abrego 2020, Deflem et al. 2021). Indeed, sample-level random effects are meaningful for multivariate models because they are not confounded by residual variation as with univariate models. More specifically, this attribute allows models to account for nonindependence among residuals for each site and improves estimates of the fixed effects (i.e. taxon-environment relationships) (Ovaskainen et al. 2016). In addition, JSDMs also allow for the inclusion of phylogenetic relatedness as a hierarchical term that can lend additional improvements to estimated taxon-environment relationships.

We provide a detailed description of the modeling framework and application elsewhere (Kopp et al. 2023). In brief, we used presence/absence data from 1,891 benthic macroinvertebrate assemblages surveyed as part of the probabilistic and handpicked sites and focused on taxa that were collected from a single site visit, identified to genus, and occurred at ≥10% of sites within an ecoregion (Table 1). Separate models were fit for each ecoregion boundaries to define regional species pools and thus assume that environmental conditions are the primary factor driving genus occurrence (Chessman and Royal 2004). All physiochemical variables were measured in the field during the survey, except for MSAT and TPRCP, which were obtained from PRISM Climate and matched to the appropriate survey year (Table 1). Physiochemical variables were used as linear fixed effects and sample-level random effects were used to statistically control for unmeasured variables. Previously, we found few genus-environment relationships were unimodal (Kopp et al. 2023) and therefore assumed linear relationships were appropriate for this study. Our analysis focused on relatively small number of environmental variables because these are commonly altered by human activities. We also use taxonomy as a surrogate for phylogenetic relatedness as a hierarchical term in the model. All models were fitted with the default prior distributions (Ovaskainen and Abrego 2020), using three independent chains (3,000 posterior samples). Convergence was determined to be satisfactory by potential scale reduction factor < 1.1 for fixed effects and phylogenic parameters.

Our primary motivation for using these models was to measure how macroinvertebrate assemblages may change if the influence of anthropogenic disturbances were removed from in stream physiochemical variables and all else remained unchanged (Figure 2). Since we do not use these models to predict to new locations, model performance was primarily evaluated with respect to its explanatory power, i.e. the fitted models ability to reproduce the observed genus richness and composition (Wilkinson et al. 2021, Abrego and Ovaskainen 2023). Predicted, taxon-specific occurrence probabilities were summed to obtain predicted richness and regressed against the observed richness (Calabrese et al. 2014). We determined model acceptability based on three criteria: $R^2 \geq 0.2$, $-1.5 \leq \text{intercept} \leq 1.5$, and $0.85 \leq \text{slope} \leq 1.15$ (Linke et al. 2005). We calculated model performance metrics using each of 3,000 posterior samples to obtain a distribution of plausible estimates of performance metrics and report the mean and 5th and 95th quantiles as measures of uncertainty. In addition, we assessed compositional similarity between predicted and observed assemblages using a probabilistic adaptation of Jaccard similarity (Scherrer et al. 2020) to avoid introducing error associated with converting predicted occurrence probabilities into binary outcomes (Calabrese et al. 2014). We calculated similarity for each site using the mean predicted occurrence probabilities and report the mean and 5th and 95th quantiles across all sites.

Predictive power for models that use random effects can only be assessed for cases where at least some sampling units were included in the calibration data (Abrego and Ovaskainen 2023). As part of NRSA, a subsample of sites are revisited with the intent of assessing temporal variability in metrics and indices (Table 1). Since we fit the JSDMs using sample-level random effects, these data were used to as a second type of validation. Specifically, we evaluated whether the richness observed at a revisited site was within the posterior distribution of the fitted models. Although this metric of validation may seem less restrictive compared to those used to evaluate explanatory power, we expected variation between samples to be rather large because of stochastic events (e.g. high or low stream flows), ecological processes (e.g. emergence and dispersal), and sampling procedures (e.g. field collection and laboratory subsampling) that our models were not designed to capture. Given this level of potential uncertainty, this metric was intended to address whether an observation is could have come from the same process that our model was intended to capture. We also compared the assemblages collected during the revisit to the predicted assemblages using the probabilistic adaptation of Jaccard similarity (Scherrer et al. 2020) and report the mean and 5th and 95th quantiles across sites.

Modeling physiochemical gradients

We used random forest models to relate anthropogenic and landscape geoclimatic factors to total nitrogen (NTL), total phosphorous (PTL), chloride (CL), sulfate (SO4), and substrate diameter (SUBD) and make predictions if anthropogenic disturbances were removed. The suite of predictor variables were selected based on their hypothesized relationship with stream water chemistry and bed particle size (Lin et al. 2021, Zak et al. 2021, Kaufmann et al. 2022b, Kaushal et al. 2023, Sabo et al. 2023). Geoclimatic factors included watershed morphology (e.g., basin area, elevation, and slope) and lithological characteristics (e.g., lithological phosphate, nitrogen, and sulfur, and soil erodibility). Anthropogenic factors included percent agriculture, road density, and presence of mines and impoundments (Appendix S1: Table S1). Predictor variables were obtained or derived from the StreamCat Database (Hill et al. 2016), National Atmospheric Deposition Program (nadp.slh.wisc.edu), and EPA's National Nutrient Inventory (Sabo et al. 2019, Lin et al. 2021, Sabo et al. 2021, Sabo et al. 2023).

Separate random forest models were fit for NTL, PTL, CL, SO4, or SUBD using the entire national dataset to maximize the range of the predictor variables included in each model. Prior to model fitting the dataset was randomly split into training and testing portions (80% and 20%, respectively). Model fit was assessed by the coefficient of determination (R^2) and root mean squared error (RMSE) on training and testing portions of the dataset. The relative importance of the covariates was evaluated by the change in mean squared error after permutating each variable (%IncMSE). Because variable importance can be influenced by correlated variables, we confirmed that all variables used in the models had pairwise Pearson's correlation coefficients < 0.7 and the variance inflation factors (VIF) were between 2.6 and 3.9. (Appendix S1: Table S3) In general, VIF > 5 indicates a potential problem with multicollinearity (O'brien 2007). Partial dependance plots were used to visualize the relationship between the most important anthropogenic factors and each physiochemical variable. Random forest analysis was performed using the quantregForest R Package (Meinshausen 2017).

Evaluating effects of human disturbance on physiochemical gradients

We used the fitted random forest models to predict values for NTL, PTL, CL, SO4, and SUBD if anthropogenic disturbance was removed by setting all anthropogenic factors to the lowest value observed in our dataset (often zero, Appendix S1: Table S2) and leaving geoclimatic factors unchanged. Hereafter "hindcast" refers to the removal of anthropogenic disturbance from the physiochemical variables. Although we included the entire range of predictor values, it is possible that hindcast data are not sufficiently similar to the data used to train the model and thus susceptible to extrapolation (Meyer and Pebesma 2021, Yuan et al. 2024). To test whether the hindcast dataset was sufficiently similar to the data used to train the models, we first mean-centered and scaled all predictor variables to equivalent units (i.e. standard deviations) and weighted them according to their importance in the model. We then calculated the minimum Euclidean distance between each site in the hindcasted data, and each site used in the training dataset using the same center and scale. The minimum value was then divided by the mean Euclidean distance among all training data. Following Yuan et al. (2024), we flagged hindcast predictions for sites that exceeded 0.5 as susceptible to extrapolation.

We evaluated the effects of anthropogenic disturbance on each physiochemical variable based on deviations from the hindcast estimates using standardized anomalies (i.e. z-scores), or the difference between the observed and hindcast value scaled by standard deviation. For variables we modeled using random forests, we standardized the difference between observed and hindcasted conditions using the RMSE of each model (Kilgour and Stanfield 2005). For climate variables, we obtained baseline historical climatic values (MSAT and TPRCP) as the mean summer air temperature for 1900-1950 from PRISM climate data (https://prism.oregonstate.edu/historical/). We then standardized the difference between the present-day values (i.e. 2018 or 2019) and the baseline using the standard deviation of the 50yr dataset. Importantly, using standardized anomalies rather than absolute differences has advantages because it accounts for unexplained variation in the model or natural variability among sites (Kilgour and Stanfield 2005). Furthermore, because the values are in units of standard deviations, thresholds to evaluate whether the magnitude of difference is sufficient to support an effect of anthropogenic disturbance can be rather intuitive. Specifically, we expected anthropogenic disturbance to elevate NTL, PTL, CL, and SO4 concentrations and increase or decrease SUBD, MSAT, and TPRCP. Accordingly, we identified sites with observed concentrations >2SD from the hindcast values for NTL, PTL, CL, and SO4 and > |2SD| for SUBD, MSAT, and TPRCP as having evidence of anthropogenic disturbance.

Since RPDI is a direct measure of anthropogenic disturbance, we initially set this variable to zero for all locations but found that this value may be too strict because of the number and diversity of factors included in the index and did not provide much insight into potential regional variation in disturbance. Instead, we used 0.33 as a threshold which is interpreted as one type of human disturbance observed within 10 m of the streambanks at no more than one third of the 22 riparian plots sampled, on average (Kaufmann et al. 2022b, USEPA 2023). Although this does mean that not all sites were completely free of riparian disturbance, it is consistent with other studies that have used this index to evaluate human disturbance in the riparian area (Kaufmann et al. 2022a, Kaufmann et al. 2022b, USEPA 2023).

For each physiochemical variable we estimated the total percent of streams that were potentially affected by anthropogenic disturbance (i.e. standardized anomaly > 2SD) using only the probabilistic samples and their weights reflective of the entire population of streams and rivers assigned to them by NRSA (USEPA 2023). Estimates for each ecoregion were generated using the cat_analysis() function from the spsurvey R package (Dumelle et al. 2023). In addition, we tallied the number of physiochemical variables that were potentially affected at each site to elucidate instances where human disturbance affects multiple environmental variables simultaneously.

Evaluating effects of human disturbance on macroinvertebrate assemblages

We evaluated the effects of removing anthropogenic disturbance from the physiochemical environment on macroinvertebrate assemblages using the fitted JSDMs. For the sites that had evidence of human disturbance (i.e. standardized anomaly > 2SD), we substituted the hindcasted value (either predicted from random forest model, 1900-1950 averages for MSAT and TPRCP or 0.33 for RPDI) for the present day value in the dataset and used these data to predict macroinvertebrate assemblages that could occur if disturbance was removed or reduced (Chessman and Royal 2004). To evaluate the relative effects of hindcasting each physiochemical variable, we created 4 scenarios changing either NTL and PTL (Nutrient Scenario), CL and SO4 (Salinity Scenario), RPDI and SUBD (Habitat Scenario), MSAT and TPRCP (Climate Scenario) and leaving the others at their observed values. In addition, we predicted the macroinvertebrate assemblage after changing all physiochemical variables to their hindcast value (Hindcast Scenario). For each scenario we compared predicted genus richness from present-day conditions to hindcast genus richness.

JSDM predictions are three-dimensional data arrays that contain 3,000 posterior samples of site-specific occurrence probabilities for each genus (i.e., site x genera x posterior samples). For each posterior sample we summed predicted occurrence probabilities for all genera to obtain 3,000 plausible estimates of hindcast genus richness for each site. We then compared the mean present-day genus richness (i.e., predicted from JSDM using present-day values for the physiochemical variables) to the hindcast posterior distribution. We evaluated whether present-day richness was below the 10^{th} or 25^{th} quantiles, indicating a reduction in present-day genus richness relative to hindcast predictions, or above the 75^{th} or 90^{th} quantiles, indicating an increase in genus richness relative to hindcasted (Figure 2). Hereafter, present-day values that are located in the extreme ends of the distribution are described as having either >0.75 or >0.90 support for a difference from hindcasted predictions, respectively. We assessed the effects of removing anthropogenic disturbance from each of the 4 combinations of environmental variables (i.e. Nutrient Scenario, Salinity Scenario, Habitat Scenario and Climate Scenario) and the effects of removing anthropogenic disturbance from all environmental variables simultaneously (Pristine Scenario) with >0.75 and >0.90 support as evidence for an effect of disturbance on physiochemical variables.

Identifying genera having site-specific occurrence probabilities that differ between present-day and hindcast conditions helps interpret assemblage-level changes in response to anthropogenic disturbance. We identified genera that had higher occurrence probability under present day conditions as “increasers” and genera with lower occurrence probability as “decreasers”. Differences were determined with >90% support. For each genus we calculated the proportion of sites where they increased or decreased. If a genus was identified as an increaser at a large proportion of sites it could become more prevalent in response to anthropogenic disturbance. Alternatively, if a genus was identified as a decreaser at a large proportion of sites, it could become less prevalent in response to human disturbances. For each ecoregion, we report the mean for major taxonomic groups (e.g. insects and non-insect genera) to further understand taxon-specific trends in the context of biodiversity loss (Jähnig et al. 2021, Rumschlag et al. 2023).

Identifying increaser and decreaser genera could also enhance our ability to measure the effects of removing anthropogenic disturbance because assemblage composition could be affected by anthropogenic disturbance without a subsequent change in richness (Van Sickle 2008). To assess compositional differences between present-day and hindcast assemblages we used increaser/decreaser assignments to create two community matrices, one representing present-day assemblage and the other representing hindcast assemblage. Increasers were assigned a value of 1 in the present-day matrix and 0 in the hindcast matrix because they had a higher occurrence probability under present-day conditions. Conversely, decreasers were assigned 0 in the present-day matrix and 1 in the hindcast matrix because they had a significantly higher occurrence probability under hindcasted conditions. For genera that did not have sufficient support (< 0.90) for changing occurrence probabilities were considered not to be affected by anthropogenic disturbance and were assigned 1 in both matrices. We compared the two matrices (i.e. site x genera) using Jaccard similarity index and identified sites with a similarity score of < 0.9 as changing compositionally.

We assessed the consequences of removing anthropogenic disturbance from in-stream physiochemical environment for macroinvertebrate assemblages by identifying sites within each region that had evidence for either a change in richness (>0.75 support) or composition (Jaccard Similarity < 0.9) from present-day conditions. Then, using only the probabilistic samples that were not flagged for extrapolation and had complete data (n = 1748) and their weights, estimated the total percent of streams within each ecoregion that were potentially affected by anthropogenic disturbance. Estimates for each ecoregion were generated using the cat_analysis() function from the spsurvey R package (Dumelle et al. 2023). In addition, we report results with different levels of support (> 0.75 and > 0.90) for richness change and separate compositional change to convey uncertainty and methodological differences.

Results:

Random Forest Modeling

The random forest models explained 46 to 77% of the variation in the training data (n=1502) and 51 to 78% of the variation in the test/validation data (n=375). The model of substrate diameter (SUBD) had the highest RMSE in both training and testing datasets compared to models of total nitrogen (NTL), total phosphorous (PTL), chloride (CL), and sulfate (SO4) (Table 2). Variable importance for each physiochemical model, measured by the percent change in MSE after permutation, revealed that runoff was among the most important geoclimatic variables for each model and that agricultural landcover in the watershed or riparian area ranked among the most important anthropogenic variables (Appendix S1: Table S2). Percent agricultural landcover in the watershed was the most important variable predicting NTL and PTL, the second most important for predicting SO4 and third most important for predicting CL. Other anthropogenic variables were also ranked relatively high in importance. For example, road density in the watershed was the second most important variable predicting CL and fifth most important variable for SO4, while percent natural vegetation cover in the riparian area was the third most important variable predicting SUBD.

We visualized the effects of the anthropogenic factors using partial dependance plots (Figure 3). In general, the anthropogenic factors were associated with increased nutrients and salinity and decreased substrate diameter. For example, percent crops and anthropogenic nutrient inputs of nitrogen and phosphorous were positively associated with NTL and PTL. Road density and mean summer air temperature had a positive association with CL while total annual precipitation had a weakly negative relationship. Density of coal mines in the watershed had a relatively strong positive effect on SO4 and SUBD had a strong positive relationship with natural vegetation cover and strong negative relationship with agricultural activities in the riparian area.

JSDM assemblage level performance

We fitted JSDMs for each region using presence/absence data for 59 – 127 genera from 152 – 266 sites surveyed during the 2018 – 2019 NRSA cycle (Table 1). The total number of sites used to fit JSDMs were higher (1,891) than the sites we ultimately assessed because we included probabilistic and handpicked sites (USEPA 2023). We found that summing predicted occurrence probabilities for all genera modeled could potentially overestimate observed genus richness (i.e. observed vs predicted richness intercepts < -1.5). Upon further inspection we found that this overprediction was due to many genera having exceptionally low predicted occurrence probabilities but nonetheless present in the regional species pool. Thus, we used a threshold to excluded genera with predicted occurrence probabilities < 0.05 or < 0.10 prior to summation to correct this bias (Table 3).

Because site level random effects accounted for unmeasured factors in the models, we did not expect our models to have strong predictive power when projected to new locations (Abrego and Ovaskainen 2023, Kopp et al. 2023). However, we could assess whether values obtained at revisited sites were within the predicted posterior distribution of the fitted models as a secondary form of validation. Although the number of revisited sites available for model testing was low (Table 1), we found that >90% of the genus richness values observed during revisit sampling were within the posterior distribution of the models. Importantly, this lends plausibility that the posterior distribution reflects processes characterizing macroinvertebrate assemblages (Table 3 and Appendix S1: Figure S2).

The probabilistic adaptation of Jaccard similarity revealed that the predicted and observed assemblages were on average >50 similar (Table 3). Sites located the WMT tended to have the most similar assemblages (mean = 0.6, 5th quantile = 0.36, 95th quantile= 0.75) while sites in the XER tended to have the least similarity (mean = 0.23, 5th quantile = 0.23, 95th quantile= 0.71). Although potentially vulnerable to sample size constraints, mean similarity between model predictions and assemblages observed at revisited sites were also relatively high, with mean values ranging between 0.35 and 0.68 (Table 3).

Disturbance effects on physiochemical conditions

We tested whether using the random forest models to predict physiochemical conditions after removing the effects of anthropogenic activities were susceptible to extrapolation. We found predictions at 77 of the sites (~4%) were potentially susceptible to extrapolation. This indicates that removing disturbance would cause some sites to be sufficiently different from the data used to train the models and suggest that these sites do not have a natural analog. Indeed, most of the sites flagged for potential extrapolation were located in the Coastal Plains Ecoregions (18%) and were associated with models used to predict NTL concentrations in the lower Mississippi and SUBD along the southern coast (Appendix S1: Figure S1). To avoid the potential for bias associated with extrapolation, we designated them as “Not Assessed” in our analysis.

For sites that were not potentially susceptible to extrapolation, removing anthropogenic disturbance from the physiochemical environment generally decreased nutrient concentrations, salinity, and riparian disturbance and increased substrate diameter (Figure 4). Change in NTL and PTL concentrations were most evident in Temperate Plains and Southern Plains and change in CL was most evident in the Northern Appalachians. Substrate diameter was noticeably coarser after hindcasting for all regions except Xeric, Southern Appalachians, and Northern Appalachians. Observed and hindcasted values for sulfate, and mean summer air temperature were generally similar for all ecoregions.

Anthropogenic disturbance affected physiochemical gradients differently depending on the ecoregion (Figure 5a). We found 60.8% (95% Confidence Interval, hereafter CI = 52.8-68.8%) of streams in Temperate Plains and 38.4% (CI = 32.0-44.9%) of streams Upper Midwest had elevated NTL relative to hindcasted estimates. In the Northern Appalachians, 41.4% (CI = 34.5-48.2%) of streams were found to have potentially elevated CL concentrations. Streams with excess fine substrates were most prevalent in the Temperate Plains (29.2%, CI = 21.2-37.1%) and Northern Plains (21.4%. CI = 14.1-28.9%). We also found similarities among regions. RPDI and TPRCP were the most prevalent disturbance among all ecoregions while disturbed SO4 and MSAT values were the least prevalent. Finally, anthropogenic disturbance could alter multiple physiochemical variables simultaneously (Figure 5b). In general sites with >2 physiochemical variables disturbed by anthropogenic activities were located in Northern Appalachians, Southern Plains, Temperate Plains, and Upper Midwest.

Macroinvertebrate assemblage response to undisturbed physiochemical environment

Macroinvertebrate models were fit for each region separately and we assessed whether removing anthropogenic disturbance would produce estimates that exceed the range of values used to fit each model. Hindcast predictions exceeded the range of present-day conditions at 1 location in CPL, SAP and XER, 6 locations in TPL, and 12 locations in WMT. In these instances, we reset the hindcasted values to be within the range (minimum) of data used to fit the mode to avoid potential extrapolation.

Hindcast predictions for each category of physiochemical variables generated different outcomes for macroinvertebrate assemblages (Figure 6a). For example, removing anthropogenic disturbance from NTL and PTL concentrations, produced macroinvertebrate assemblages with higher genus richness in the Temperate Plains and lower richness at some sites in the Southern Plains and Southern Appalachians ecoregions (Figure 6a). On the other hand, removing disturbance from CL and SO4 tended to increase genus richness in the Southern Plains but decreased richness at sites in the Northern Appalachians. Hindcasted physical habitat variables, mostly affected macroinvertebrate assemblages in the Northern Plains by decreasing richness but also increased genus richness in Coastal Plains, and Northern Appalachians and Upper Midwest (Figure 6a). TPRCP generally contributed to changes in macroinvertebrate assemblages because relatively few sites had MSAT temperature anomalies exceed 2 standard deviations from 1900-1950 mean. In the Southern Appalachians genus richness decreased from present day conditions while in the Northern Plains genus richness generally increased from present day conditions. Removing hindcasting all physiochemical variables typically increased genus richness in all regions except for the Northern Plains and Southern Appalachians (Figure 6b).

Identifying genera that had higher present-day occurrence probabilities (increasers), or higher hindcast occurrence probabilities (decreasers) indicated that insects tended to be decreasers while non-insects tended to be increasers, but this was not consistent across all regions (Table 4). For example, in the Southern Plains and Western Mountains, insects were almost equally likely to be increasers (0.1 and 0.6, respectively) or decreasers (0.09 and 0.03, respectively) and in the Northern Plains both insects and non-insects tended to be increasers at a larger proportion of sites (0.19 and 0.22). Among insects, members of chironomidae were generally most likely to be increasers. Among non-insects, in contrast, members of mollusca were most likely increasers (Table 4). Thus, higher present-day genus richness compared to hindcast estimates could be due to genera that are typically considered to be tolerant of human activities.

Using the probabilistic sites, we aggregated our results to reflect the total population of streams and rivers in each ecoregion and detected that 14.3 – 75.5% of streams have present-day assemblages that differ from assemblages expected from the hindcasted physiochemical environment (Figure 7). With respect to changes in richness with support >0.75, the Northern Plains (58.3%), Temperate Plains (54.8%), and Southern Plains (54.2%) the most streams with potentially altered macroinvertebrate assemblages while the Western Mountains (9.2%), Upper Midwest (19.6%), and Northern Appalachians (29.7%) have the fewest. Compositional change without a corresponding change in genus richness could increase the percentage of streams with altered assemblages. Evidence for compositional change was most pronounced in Temperate Plains (20.7%), Southern Appalachians (20.7 %) and Upper Midwest (18.8%). All regions had a percentage of streams that could not be assessed because of potential extrapolation of hindcasted physiochemical variables, incomplete data or macroinvertebrate assemblages consisting of only rare genera (i.e. prevalence < 10%). Among them, the Coastal Plains (12.9%) and Xeric (11.1%) had the largest percentage of streams that were not assessed, thus the total number of sites assessed was less than the total number of probabilistic sites surveyed (n = 1748).

Discussion

Assess the effects of anthropogenic disturbance on biological assemblages is challenging because few minimally disturbed sites remain (Stoddard et al. 2006). Sites in least-disturbed are often adopted as an alternative but are difficult to consistently define, vary in quality, and can be spatially aggregated (Hawkins et al. 2010, McNellie et al. 2020). The model-based framework we developed contributes to a suite of other approaches intended to circumvent the need for reference sites (Chessman and Royal 2004, Elias et al. 2016, Yuan et al. 2024). Specifically, we evaluated the effects of human disturbance on several physiochemical variables and, in turn, addressed whether altered physiochemical conditions affect macroinvertebrate assemblages. Using this approach, we found that anthropogenic disturbance can affect multiple physiochemical variables simultaneously and that the effects on any single factor can vary among ecoregions. We also found that removing or reducing disturbance could change genus richness at >50% of the streams in some ecoregions and up to 75% if compositional change is considered. Collectively, our framework offers a promising alternative to evaluating the effects of specific disturbances on macroinvertebrate assemblages that does not rely on reference sites.

Hindcasting physiochemical conditions

Random forest models have been used by others to predict physiochemical variables if human disturbance was reduced or removed (Yuan et al. 2024). In general, using models to infer physiochemical conditions eliminates the need to identify reference sites or collect excessive data from undisturbed locations (Herlihy and Sifneos 2008, Soranno et al. 2011). In this study, random forest models improve on earlier regression-based hindcasting approaches (Dodds and Oakes 2004, Herlihy and Sifneos 2008) because they incorporate a relatively large number of natural and anthropogenic predictor variables and accommodate their complex relationships with the response variables (Yuan et al. 2024). Importantly, because reference sites may not be representative of all sites that need to be assessed including natural and anthropogenic variables enabled us to make site-specific predictions rather than relying on an estimated mean value (i.e. the intercept from a multiple regression as a function of only disturbance) for an entire region (Dodds and Oakes 2004, Herlihy and Sifneos 2008). Thus, our hindcasting approach provides a first-order approximation for what conditions at a site could be if they are presently unknown.

Furthermore, random forest models permitted us to evaluate the relative importance of specific anthropogenic activities and characterize potential site and regional scale differences after reducing or removing them. For example, higher than expected total nitrogen in Temperate Plains and Upper Midwest and the relatively high importance of percent agriculture landcover our model suggests that nutrient inputs associated with agricultural activities may be elevating in-stream nutrient concentrations (Lin et al. 2021, Sabo et al. 2023). Similarly, elevated chloride concentrations in the Northern Appalachians and the relative high importance of road density in the model, could reflect the contribution of road salts to freshwater salinization (Kaushal et al. 2023).

Certainly, inferring physiochemical conditions for different levels of human disturbance depends on the quality and structure of the model. Although our models were evaluated using an independent validation dataset, hindcasting required a dataset that reduced or eliminated the effect of predictor variables associated with human disturbance. As consequence, the dataset used for hindcasting could be dissimilar from the data used to calibrate the model and potentially generate errors associated with extrapolation (Meyer and Pebesma 2021). For most regions, the hindcasting dataset was sufficiently similar to the training dataset such that we could assess the alteration of physiochemical variables for majority of sites. Nonetheless 18% of sites surveyed the Coastal Plains were flagged as a result of testing for extrapolation. Considering that many of the sites were located in the lower Mississippi River Basin, a unique and heavily modified system, it is perhaps unsurprising that removing human disturbances from this system is beyond the domain of our model. Testing for potential extrapolation in the context of hindcasting is novel and, although the threshold to determine whether our results were prone to extrapolation has been used previously (Yuan et al. 2024), more research should be devoted to optimizing and interpreting this threshold. When locations or streams are potentially susceptible to extrapolation, it might be more reasonable to select values that represent management goals, best professional judgement or those that maximize societal benefits (Bouleau and Pont 2015).

To assess how human-related disturbance effects physiochemical environment, we used standardized anomalies (i.e. z-scores), to account for either unexplained variation in the model values or variation along a baseline. Thresholds are of central importance for the communication and evaluation of human disturbances communication (Wood 2008). Here, we used a ±2SD threshold for deciphering whether human disturbances alter physiochemical conditions (Kilgour and Stanfield 2005). Although this threshold is intuitive, it could potentially be too strict for some variables. For example, values for MSAT at many locations did not exceed the threshold in our analysis, suggesting that 2018 and 2019 temperature averages are consistent with 1900-1950 averages, given interannual variation. However, warming of 1SD has been implicated in reductions of insects in agricultural landscapes (Outhwaite et al. 2022) and there is evidence for increasing summer air temperatures in the United States of approximately 0.09°C per decade since 1901 (USEPA 2024a). On the other hand, to evaluate riparian disturbance, we used a threshold of 0.33 which suggests that human disturbance is not completely absent for all sites that were below this threshold. Indeed, nearly every site in our analysis has RPDI >0 such that this threshold did not informatively differentiate regional variation in disturbance. Further, this is the threshold used by others to identify least disturbed sites (Kaufmann et al. 2022b, USEPA 2023) and we sought to evaluate how this could affect microinvertebrate assemblages. Although thresholds were necessary to identify which regions may have relatively more disturbed physiochemical conditions that others, future applications may select different thresholds, report continuous scores, or vary the amount of disturbance to meet specific applications (Yuan et al. 2024).

Application for biological assessment without reference sites

Using hindcast physiochemical variables to predict the assemblage expected to occur if anthropogenic disturbance was removed allowed us to infer the consequences of altering a relatively small number of physiochemical variables. Central to our application is that multivariate models can implement meaningful site-level random effects that statistically control for unmeasured environmental variables and potential biotic interactions (Warton et al. 2015, Ovaskainen et al. 2016, Kopp et al. 2023). Although, this feature kept our model sufficiently tractable it limits our ability to make accurate predictions to new locations (Abrego and Ovaskainen 2023). However, predicting to new locations was not our objective. Rather, our models establish empirical relationships between macroinvertebrate occurrences and then evaluate how those assemblages might differ if select physiochemical conditions changed while all else remained constant. Furthermore, because the data we used for our analysis was collected as part of a probabilistic survey, designed to be representative of the population of stream and rivers (Olsen and Peck 2008), our site-specific inferences can be aggregated to elucidate regional and sub-continental patterns. Indeed, it may be undesirable to refit models every time an assessment is needed, and future efforts should focus on improving predictive abilities by including immutable factors in addition to stressor gradients (Yuan et al. 2024).

Importantly our approach differs significantly from traditional, reference site-based approaches (Hawkins et al. 2000, Herlihy et al. 2008, USEPA 2023). Foremost, our efforts focus explicitly on a relatively small number of potential stressors whereas reference-site approaches focus implicitly on a theoretically larger, but undefined number of stressors. For example, the difference between a test site and reference sites could be related to a number of other stressors that were not used as biotic screens but co-occur with them (Herlihy et al. 2008, Paulsen et al. 2008). Alternatively, in our model-based approach, differences between present day and hindcast assemblages are only related to changes in the physiochemical variables included as fixed effects. Although the former may elucidate general disturbance effects, it is difficult to attribute differences between assemblages to a specific environmental disturbance (Paulsen et al. 2008). On the other hand, our model-based approach enhances interpretations with respect to specific physiochemical variables but may omit important anthropogenic stressors that were not explicitly included in the model. Thus, the choice between reference site-based and model-based assessments may be contingent on whether the environmental gradients that are commonly disturbed by human activities can be identified and appropriately modeled.

The JSDMs were also fitted using bayesian inference and, as such, yielded posterior distributions that can be used for hypothesis testing (Johnson et al. 2022). Specifically, we evaluate whether the effects of human disturbance on physiochemical variables were sufficient to alter benthic assemblages on a site-specific basis. In contrast, traditional reference site-based approaches evaluate each test site based on quantiles of the distribution obtained from reference sites (Herlihy et al. 2008, USEPA 2023). In this regard, JSDMs and bayesian inference may be advantageous in interpretating biological condition estimates because they account for uncertainty in predictions for each site. Furthermore, we also included all genera from the regional taxa pool while some typical reference site-based approaches include only taxa that occur at reference sites. This implies that in the absence of human disturbance those taxa should occur at all locations regardless of other taxa. This is problematic because genera that tolerate anthropogenic disturbance probably evolved under similar conditions that occurred naturally or have remarkable plasticity (Wiens et al. 2010, Heino et al. 2013). Finally, our model-based approach can reveal compositional changes. Although the probabilistic adaptation of Jaccard's similarity calculation suggests that predicting individual genera is more challenging than aggregated genus richness, this analysis provided additional information that is generally not easily available from other approaches (Hawkins 2006, Van Sickle 2008) and enhanced our ability to identify sites that are potentially affected by human disturbance.

Selecting the appropriate taxonomic level that specimens should be identified to is a critical decision in biological assessment (Chessman et al. 2007). We focused exclusively on taxa that were identified to the genus level and excluded those that could not be unambiguously identified. This possibly increases false absences rates in our study, but NRSA identifies most organisms to their genus such that these instances are relatively rare. Further, common taxonomy alleviates some ambiguity associated with aggregating unresolved taxa into operational taxonomic units (Yuan et al. 2008) and improves transferability of our genus-environment relationships to other studies. Fixed-count subsampling, performed as part of the standardized NRSA protocol, could also increase false absences in our study. Although, explicitly modeling taxon-specific detection probabilities from replicate subsamples may be an interesting avenue for future research (Doser et al. 2023, Doser et al. 2024), these data are presently unavailable. Finally, focusing on genera could also reduce interspecific variation, but species-level taxonomy was not available for this dataset. In general, finer taxonomic resolution would be substantially more expensive and perhaps only yield marginal benefits (Chessman et al. 2007). Nonetheless, metabarcoding approaches may have shown promise for bioassessments (Smucker et al. 2024) and could avoid some of the limitations associated with selecting the appropriate taxonomic resolution.

Although the model-based approaches may have some advantages over reference site-based approaches, they require further investigation before they can be fully assimilated into biomonitoring programs. Nevertheless, the National Rivers and Streams Assessment reports biological condition estimates at for the 2018-19 survey using a multi-metric index (USEPA 2024b) and it is worthwhile to compare our results to those. The results from NRSA are available at https://riverstreamassessment.epa.gov/dashboard. Based on a ranked comparison, the largest disagreements pertained to the percentage of stream miles in poor condition in the Plains ecoregions. Specifically, our model-based approach showed more streams to be altered in the Northern Plains, Southern Plains and Temperate Plains. Indeed, the US Great Plains (i.e. Temperate Plains, Northern Plains and Southern Plains) have undergone extensive conversion from grasslands to agriculture such that there may be few sites that are undisturbed (Samson and

Knopf 1994, Dodds et al. 2004, Olimb and Robinson 2019). Because of the large extent of anthropogenic activities in these regions, it is likely that the reference sites used to assess biological condition are potentially lower quality (Herlihy et al. 2008). Conversely, NRSA ranked the Coastal Plains as having a larger percentage of streams in poor condition than our model-based approach. Unfortunately, because of potential extrapolation associated with our model-based assessment, the Coastal Plains also had a relatively large number of streams that could not be assessed. The rankings for the remaining ecoregions tended be similar and both methods agreed that streams in the Western Mountains were the least-disturbed among the ecoregions. Indeed, about 75% of the land area of this ecoregion is in federal ownership, which could convey some protection to streams and rivers (Jenkins et al. 2015). Furthermore, Western Mountains have a relatively shorter history of anthropogenic disturbance than other ecoregions in the United States.

Macroinvertebrate response to anthropogenic disturbance

Beyond applications for bioassessment, our model-based approach contributes to a growing literature devoted to understanding of recent trends in macroinvertebrates (Crossley et al. 2020, Jähnig et al. 2021, Gebert et al. 2022, Spake et al. 2022, Rumschlag et al. 2023). In our study, hindcasting estimates tended to show increases in genus richness, suggesting that in many ecoregions fewer genera persist when physiochemical conditions are disturbed. Other studies have reported either modest increases in macroinvertebrate richness (Gebert et al. 2022, Rumschlag et al. 2023) or no clear evidence for a widespread decline of insects (Crossley et al. 2020). Because few datasets span more than several decades, an advantage of our space-for-time approach is that it could reflect a period before major anthropogenic disturbances (Blüthgen et al.

2022). In addition, the survey design and consistent methodology implemented by NRSA reduces the potential for confounding spatial and temporal dimensions that could be present in datasets that were compiled from multiple sources (Jähnig et al. 2021, Blüthgen et al. 2022, Boyd et al. 2023). Nonetheless, although our results suggest that sites typically support fewer macroinvertebrate genera under present-day conditions, we did detect increased in genus richness at several sites in the Northern Plains and Western Mountains which suggests that macroinvertebrate response to disturbance could be context dependent (Powell et al. 2023).

We compared genus-specific occurrence probabilities under present-day and hindcasted conditions to identify which genera contribute to assemblage level patterns. Insects in most ecoregions tended to decrease with disturbance while non-insects tended to increase. This suggests that insect genera may, in general, be less tolerant to anthropogenic disturbance (Jähnig et al. 2021) and aligns with other studies that have documented declines in insect richness (Rumschlag et al. 2023). Among insects, Ephemeroptera, Plecoptera and Trichoptera (EPT) are often used as indicators of anthropogenic disturbance (Stoddard et al. 2008) and may have already lost a considerable proportion of species (Sánchez-Bayo and Wyckhuys 2019). This is especially concerning because we found that EPT occurrence probabilities could continue to decrease in many regions of the US. Among non-insects, non-arthropods included members of Annelida, Nemertea, and Platyhelminthes tended to increase at a relatively large proportion of sites which is consistent with high pollution tolerance values typically assigned to these taxa (Carlisle et al. 2007, Griffith 2023).

Because our approach was based on quantifying genus-environment relationships along environmental gradients that are typically altered by human activity, we could also compare the relative effects of changing specific categories of physiochemical variables on macroinvertebrate assemblages. For example, higher genus richness under present-day conditions in the Northern Plains compared to hindcasted habitat variables suggests that disturbance may have increased genera richness. Interestingly, cattle in pasture and rangelands were among the most common types of disturbance measured by RPDI for this region (USEPA 2023) and could potentially increase genera richness by suppressing riparian forest cover and, in turn, elevating primary production (Mittelbach et al. 2001, Tonkin et al. 2013). Conversely, in the Northern Appalachians and Coastal Plains, where the presence of trash, landfills or buildings were the most common factors for RPDI, we found that genus richness was presently lower which is consistent with the negative effects of urbanization on stream macroinvertebrate assemblages (Morse et al. 2003). We also found different effects with respect to changes in precipitation whereby genus richness increased in the Northern Plains and decreased in Southern Appalachians. Collectively, these patterns reveal which environmental variables may be most important in structuring microinvertebrate assemblages and in what contexts – an interpretation that would not be possible with traditional reference site-based approaches.

Conclusions

We used a model-based approach to assess the potential effects of anthropogenic disturbance on physiochemical gradients and benthic macroinvertebrate assemblages. Our approach combines genus-environment relationships with estimates of a number of important dimensions in physiochemical condition after removing anthropogenic disturbance. The number of sites in minimally disturbed condition are progressively diminishing, so methods that circumvent the need for reference sites for biological assessments of streams and rivers are crucial to understanding the extent of anthropogenic impacts. Importantly, our framework could provide an avenue to conduct biological assessment without depending on least disturbed reference sites.

Acknowledgements

We thank S. Paulsen, L. Riato, E. Fergus, K. Blocksom, and D. Peck for valuable discussions throughout the development of this project. D. Benkendorf, S. Rumschlag, D. Carlisle and an anonymous reviewer provided comments on earlier versions of this manuscript. The information in this document has been funded entirely by the U.S. Environmental Protection Agency, in part through appointments to the Internship/Research Participation Program at the Office of Research and Development, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency. Any mention of trade names, products or services does not imply an endorsement by the U.S. Government or the U.S. Environmental Protection Agency. The EPA does not endorse and commercial products, services, or enterprise

References:

Abrego, N., and O. Ovaskainen. 2023. Evaluating the predictive performance of presence–absence models: Why can the same model appear excellent or poor? Ecology and Evolution **13**:e10784.

Blüthgen, N., M. Staab, R. Achury, and W. W. Weisser. 2022. Unravelling insect declines: can space replace time? Biology Letters **18**:20210666.

Bouleau, G., and D. Pont. 2015. Did you say reference conditions? Ecological and socio-economic perspectives on the European Water Framework Directive. Environmental Science & Policy **47**:32-41.

Boyd, R. J., G. D. Powney, and O. L. Pescott. 2023. We need to talk about nonprobability samples. Trends in Ecology & Evolution **38**:521-531.

Breiman, L. 2001. Random Forests. Machine Learning **45**:5-32.

Buss, D. F., D. M. Carlisle, T.-S. Chon, J. Culp, J. S. Harding, H. E. Keizer-Vlek, W. A. Robinson, S. Strachan, C. Thirion, and R. M. Hughes. 2014. Stream biomonitoring using macroinvertebrates around the globe: a comparison of large-scale programs. Environmental Monitoring and Assessment **187**:4132.

Calabrese, J. M., G. Certain, C. Kraan, and C. F. Dormann. 2014. Stacking species distribution models and adjusting bias by linking them to macroecological models. Global Ecology and Biogeography **23**:99-112.

Carlisle, D. M., M. R. Meador, S. R. Moulton, and P. M. Ruhl. 2007. Estimation and application of indicator values for common macroinvertebrate genera and families of the United States. Ecological Indicators **7**:22-33.

Chessman, B., S. Williams, and C. Besley. 2007. Bioassessment of streams with macroinvertebrates: effect of sampled habitat and taxonomic resolution. Journal of the North American Benthological Society **26**:546-565.

Chessman, B. C., and M. J. Royal. 2004. Bioassessment without reference sites: use of environmental filters to predict natural assemblages of river macroinvertebrates. Journal of the North American Benthological Society **23**:599-615.

Crossley, M. S., A. R. Meier, E. M. Baldwin, L. L. Berry, L. C. Crenshaw, G. L. Hartman, D. Lagos-Kutz, D. H. Nichols, K. Patel, S. Varriano, W. E. Snyder, and M. D. Moran. 2020. No net insect abundance and diversity declines across US Long Term Ecological Research sites. Nature Ecology & Evolution **4**:1368-1376.

Deflem, I. S., E. Bennetsen, Ø. H. Opedal, F. C. F. Calboli, O. Ovaskainen, G. Van Thuyne, F. A. M. Volckaert, and J. A. M. Raeymaekers. 2021. Predicting fish community responses to environmental policy targets. Biodiversity and Conservation **30**:1457-1478.

Dodds, W. K., K. Gido, M. R. Whiles, K. M. Fritz, and W. J. Matthews. 2004. Life on the Edge: The Ecology of Great Plains Prairie Streams. BioScience **54**:205-216.

Dodds, W. K., and R. M. Oakes. 2004. A technique for establishing reference nutrient concentrations across watersheds affected by humans. Limnology and Oceanography: Methods **2**:333-341.

Doser, J. W., A. O. Finley, and S. Banerjee. 2023. Joint species distribution models with imperfect detection for high-dimensional spatial data. Ecology **104**:e4137.

Doser, J. W., A. O. Finley, M. Kéry, and E. F. Zipkin. 2024. spAbundance: An R package for single-species and multi-species spatially explicit abundance models. Methods in Ecology and Evolution **n/a**.

Dudgeon, D., A. H. Arthington, M. O. Gessner, Z. Kawabata, D. J. Knowler, C. Lévêque, R. J. Naiman, A. H. Prieur-Richard, D. Soto, M. L. Stiassny, and C. A. Sullivan. 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. Biol Rev Camb Philos Soc **81**:163-182.

Dumelle, M., T. Kincaid, A. R. Olsen, and M. Weber. 2023. spsurvey: Spatial Sampling Design and Analysis in R. Journal of Statistical Software **105**:1 - 29.

Elias, C. L., A. R. Calapez, S. F. P. Almeida, B. Chessman, N. Simões, and M. J. Feio. 2016. Predicting reference conditions for river bioassessment by incorporating boosted trees in the environmental filters method. Ecological Indicators **69**:239-251.

Elith, J., and J. R. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics **40**:677-697.

Fergus, C. E., J. R. Brooks, P. R. Kaufmann, A. T. Herlihy, R. A. Hill, R. M. Mitchell, and P. Ringold. 2023. Disentangling natural and anthropogenic effects on benthic macroinvertebrate assemblages in western US streams. Ecosphere **14**:e4688.

Franklin, J. 2010. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press, Cambridge.

Gebert, F., M. K. Obrist, R. Siber, F. Altermatt, K. Bollmann, and N. Schuwirth. 2022. Recent trends in stream macroinvertebrates: warm-adapted and pesticide-tolerant taxa increase in richness. Biology Letters **18**:20210513.

Griffith, M. B. 2023. Development of national stressor-specific genus sensitivity values. Science of The Total Environment **895**:165121.

Guisan, A., W. Thuiller, and N. E. Zimmermann. 2017. Habitat Suitability and Distribution Models: With Applications in R. Cambridge University Press, Cambridge.

Hawkins, C. P. 2006. QUANTIFYING BIOLOGICAL INTEGRITY BY TAXONOMIC COMPLETENESS: ITS UTILITY IN REGIONAL AND GLOBAL ASSESSMENTS. Ecological Applications **16**:1277-1294.

Hawkins, C. P., R. H. Norris, J. N. Hogue, and J. W. Feminella. 2000. DEVELOPMENT AND EVALUATION OF PREDICTIVE MODELS FOR MEASURING THE BIOLOGICAL INTEGRITY OF STREAMS. Ecological Applications **10**:1456-1477.

Hawkins, C. P., J. R. Olson, and R. A. Hill. 2010. The reference condition: predicting benchmarks for ecological and water-quality assessments. Journal of the North American Benthological Society **29**:312-343.

Heino, J., D. Schmera, and T. Erős. 2013. A macroecological perspective of trait patterns in stream communities. Freshwater Biology **58**:1539-1555.

Herlihy, A. T., P. R. Kaufmann, and M. E. Mitch. 1991. Stream chemistry in the eastern United States: 2. Current sources of acidity in acidic and low acid-neutralizing capacity streams. Water Resources Research **27**:629-642.

Herlihy, A. T., S. G. Paulsen, J. V. Sickle, J. L. Stoddard, C. P. Hawkins, and L. L. Yuan. 2008. Striving for consistency in a national assessment: the challenges of applying a reference-condition approach at a continental scale. Journal of the North American Benthological Society **27**:860-877.

Herlihy, A. T., and J. C. Sifneos. 2008. Developing nutrient criteria and classification schemes for wadeable streams in the conterminous US. Journal of the North American Benthological Society **27**:932-948.

Herlihy, A. T., J. L. Stoddard, and C. B. Johnson. 1998. The Relationship Between Stream Chemistry and Watershed Land Cover Data in the Mid-Atlantic Region, U.S. Water, Air, and Soil Pollution **105**:377-386.

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. JAWRA Journal of the American Water Resources Association **52**:120-128.
Hughes, R. M., and D. V. Peck. 2008. Acquiring data for large aquatic resource surveys: the art of compromise among science, logistics, and reality. Journal of the North American Benthological Society **27**:837-859.
Jähnig, S. C., V. Baranov, F. Altermatt, P. Cranston, M. Friedrichs-Manthey, J. Geist, F. He, J. Heino, D. Hering, F. Hölker, J. Jourdan, G. Kalinkat, J. Kiesel, F. Leese, A. Maasri, M. T. Monaghan, R. B. Schäfer, K. Tockner, J. D. Tonkin, and S. Domisch. 2021. Revisiting global trends in freshwater insect biodiversity. WIREs Water **8**:e1506.
Jenkins, C. N., K. S. Van Houtan, S. L. Pimm, and J. O. Sexton. 2015. US protected lands mismatch biodiversity priorities. Proceedings of the National Academy of Sciences **112**:5081-5086.
Johnson, A. A., M. Q. Ott, and M. Dogucu. 2022. Bayes Rules!: An Introduction to Applied Bayesian Modeling Chapman and Hall/CRC.
Kaufmann, P. R., R. M. Hughes, S. G. Paulsen, D. V. Peck, C. W. Seeliger, T. Kincaid, and R. M. Mitchell. 2022a. Physical habitat in conterminous US streams and Rivers, part 2: A quantitative assessment of habitat condition. Ecological Indicators **141**:109047.
Kaufmann, P. R., R. M. Hughes, S. G. Paulsen, D. V. Peck, C. W. Seeliger, M. H. Weber, and R. M. Mitchell. 2022b. Physical habitat in conterminous US streams and rivers, Part 1: Geoclimatic controls and anthropogenic alteration. Ecological Indicators **141**:109046.
Kaufmann, P. R., P. Levine, E.G. Robison, C. Seeliger, and D.V. Peck. 1999. Quantifying Physical Habitat in Wadeable Streams. EPA/620/R-99/003, U.S. Environmental Protection Agency, Washington, D.C.

Kaushal, S. S., G. E. Likens, P. M. Mayer, R. R. Shatkay, S. A. Shelton, S. B. Grant, R. M. Utz, A. M. Yaculak, C. M. Maas, J. E. Reimer, S. V. Bhide, J. T. Malin, and M. A. Rippy. 2023. The anthropogenic salt cycle. Nature Reviews Earth & Environment **4**:770-784.

Kilgour, B., and L. Stanfield. 2005. A new approach for defining reference conditions in streams. AGU Spring Meeting Abstracts.

Kopp, D. A., J. L. Stoddard, R. A. Hill, J. M. Doyle, P. R. Kaufmann, A. T. Herlihy, and S. G. Paulsen. 2023. Joint species distribution models reveal taxon-specific sensitivities to potential anthropogenic alteration. Freshwater Science **42**:268-284.

Lin, J., J. E. Compton, R. A. Hill, A. T. Herlihy, R. D. Sabo, J. R. Brooks, M. Weber, B. Pickard, S. G. Paulsen, and J. L. Stoddard. 2021. Context is Everything: Interacting Inputs and Landscape Characteristics Control Stream Nitrogen. Environmental Science & Technology **55**:7890-7899.

Linke, S., R. H. Norris, D. P. Faith, and D. Stockwell. 2005. ANNA: A new prediction method for bioassessment programs. Freshwater Biology **50**:147-158.

McNellie, M. J., I. Oliver, J. Dorrough, S. Ferrier, G. Newell, and P. Gibbons. 2020. Reference state and benchmark concepts for better biodiversity conservation in contemporary ecosystems. Global Change Biology **26**:6702-6714.

Meinshausen, N. 2017. quantregForest: Quantile Regression Forests. R package version 1.3-7.

Meyer, H., and E. Pebesma. 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution **12**:1620-1633.

Mitchell, R. M., A. T. Herlihy, and R. M. Hughes. 2025. Striving for consistency in a national lake assessment: Defining reference status and littoral macroinvertebrate condition in lakes across the conterminous United States. Ecological Indicators **170**:112992.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. WHAT IS THE OBSERVED RELATIONSHIP BETWEEN SPECIES RICHNESS AND PRODUCTIVITY? Ecology **82**:2381-2396.

Morse, C. C., A. D. Huryn, and C. Cronan. 2003. Impervious Surface Area as a Predictor of the Effects of Urbanization on Stream Insect Communities in Maine, U.S.A. Environmental Monitoring and Assessment **89**:95-127.

O'brien, R. M. 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity **41**:673-690.

Olimb, S. K., and B. Robinson. 2019. Grass to grain: Probabilistic modeling of agricultural conversion in the North American Great Plains. Ecological Indicators **102**:237-245.

Olsen, A. R., and D. V. Peck. 2008. Survey design and extent estimates for the Wadeable Streams Assessment. Journal of the North American Benthological Society **27**:822-836.

Olson, J. R., and S. M. Cormier. 2019. Modeling Spatial and Temporal Variation in Natural Background Specific Conductivity. Environmental Science & Technology **53**:4316-4325.

Olson, J. R., and C. Hawkins, P. . 2013. Developing site-specific nutrient criteria from empirical models. Freshwater Science **32**:719-740.

Olson, J. R., and C. P. Hawkins. 2012. Predicting natural base-flow stream water chemistry in the western United States. Water Resources Research **48**.

Outhwaite, C. L., P. McCann, and T. Newbold. 2022. Agriculture and climate change are reshaping insect biodiversity worldwide. Nature **605**:97-102.

Ovaskainen, O., and N. Abrego. 2020. Joint Species Distribution Modelling: With Applications in R. Cambridge University Press, Cambridge.

Ovaskainen, O., D. B. Roy, R. Fox, and B. J. Anderson. 2016. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. Methods in Ecology and Evolution **7**:428-436.

Ovaskainen, O., G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters **20**:561-576.

Paulsen, S. G., A. Mayio, D. V. Peck, J. L. Stoddard, E. Tarquinio, S. M. Holdsworth, J. V. Sickle, L. L. Yuan, C. P. Hawkins, A. T. Herlihy, P. R. Kaufmann, M. T. Barbour, D. P. Larsen, and A. R. Olsen. 2008. Condition of stream ecosystems in the US: an overview of the first national assessment. Journal of the North American Benthological Society **27**:812-821.

Peters, J., B. D. Baets, N. E. C. Verhoest, R. Samson, S. Degroeve, P. D. Becker, and W. Huybrechts. 2007. Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling **207**:304-318.

Powell, K. E., T. H. Oliver, T. Johns, M. González-Suárez, J. England, and D. B. Roy. 2023. Abundance trends for river macroinvertebrates vary across taxa, trophic group and river typology. Glob Chang Biol **29**:1282-1295.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems **9**:181-199.

Reynoldson, T. B., R. H. Norris, V. H. Resh, K. E. Day, and D. M. Rosenberg. 1997. The Reference Condition: A Comparison of Multimetric and Multivariate Approaches to Assess Water-Quality Impairment Using Benthic Macroinvertebrates. Journal of the North American Benthological Society **16**:833-852.

Rumschlag, S. L., M. B. Mahon, D. K. Jones, W. Battaglin, J. Behrens, E. S. Bernhardt, P. Bradley, E. Brown, F. De Laender, R. Hill, S. Kunz, S. Lee, E. Rosi, R. Schäfer, T. S. Schmidt, M. Simonin, K. Smalling, K. Voss, and J. R. Rohr. 2023. Density declines, richness increases, and composition shifts in stream macroinvertebrates. Science Advances **9**:eadf4896.

Sabo, R. D., C. M. Clark, J. Bash, D. Sobota, E. Cooter, J. P. Dobrowolski, B. Z. Houlton, A. Rea, D. Schwede, S. L. Morford, and J. E. Compton. 2019. Decadal Shift in Nitrogen Inputs and Fluxes Across the Contiguous United States: 2002–2012. Journal of Geophysical Research: Biogeosciences **124**:3104-3124.

Sabo, R. D., C. M. Clark, D. A. Gibbs, G. S. Metson, M. J. Todd, S. D. LeDuc, D. Greiner, M. M. Fry, R. Polinsky, Q. Yang, H. Tian, and J. E. Compton. 2021. Phosphorus Inventory for the Conterminous United States (2002–2012). Journal of Geophysical Research: Biogeosciences **126**:e2020JG005684.

Sabo, R. D., B. Pickard, J. Lin, B. Washington, C. M. Clark, J. E. Compton, M. Pennino, B. Bierwagen, S. D. LeDuc, J. N. Carleton, M. Weber, M. Fry, R. Hill, S. Paulsen, A. Herlihy, and J. L. Stoddard. 2023. Comparing Drivers of Spatial Variability in U.S. Lake and Stream Phosphorus Concentrations. Journal of Geophysical Research: Biogeosciences **128**:e2022JG007227.

Samson, F., and F. Knopf. 1994. Prairie conservation in North America. BioScience **44**:418-421.

Sánchez-Bayo, F., and K. A. G. Wyckhuys. 2019. Worldwide decline of the entomofauna: A review of its drivers. Biological Conservation **232**:8-27.

Scherrer, D., H. K. Mod, and A. Guisan. 2020. How to evaluate community predictions without thresholding? Methods in Ecology and Evolution **11**:51-63.

Smucker, N. J., E. M. Pilgrim, C. T. Nietch, L. Gains-Germain, C. Carpenter, J. A. Darling, L. L. Yuan, R. M. Mitchell, and A. I. Pollard. 2024. Using DNA metabarcoding to characterize national scale diatom-environment relationships and to develop indicators in streams and rivers of the United States. Science of The Total Environment **939**:173502.

Soranno, P. A., T. Wagner, S. L. Martin, C. McLean, L. N. Novitski, C. D. Provence, and A. R. Rober. 2011. Quantifying regional reference conditions for freshwater ecosystem management: A comparison of approaches and future research needs. Lake and Reservoir Management **27**:138-148.

Spake, R., M. P. Barajas-Barbosa, S. A. Blowes, D. E. Bowler, C. T. Callaghan, M. Garbowski, S. D. Jurburg, R. van Klink, L. Korell, E. Ladouceur, R. Rozzi, D. S. Viana, W.-B. Xu, and J. M. Chase. 2022. Detecting Thresholds of Ecological Change in the Anthropocene. Annual Review of Environment and Resources **47**:797-821.

Stoddard, J. L., A. T. Herlihy, D. V. Peck, R. M. Hughes, T. R. Whittier, and E. Tarquinio. 2008. A process for creating multimetric indices for large-scale aquatic surveys. Journal of the North American Benthological Society **27**:878-891.

Stoddard, J. L., D. P. Larsen, C. P. Hawkins, R. K. Johnson, and R. H. Norris. 2006. SETTING EXPECTATIONS FOR THE ECOLOGICAL CONDITION OF STREAMS: THE CONCEPT OF REFERENCE CONDITION. Ecological Applications **16**:1267-1276.

Tang, T., R. J. Stevenson, and J. B. Grace. 2020. The importance of natural versus human factors for ecological conditions of streams and rivers. Science of The Total Environment **704**:135268.

Tikhonov, G., Ø. H. Opedal, N. Abrego, A. Lehikoinen, M. M. J. de Jonge, J. Oksanen, and O. Ovaskainen. 2020. Joint species distribution modelling with the r-package Hmsc. Methods in Ecology and Evolution **11**:442-447.

Tonkin, J. D., R. G. Death, and K. J. Collier. 2013. Do productivity and disturbance interact to modulate macroinvertebrate diversity in streams? Hydrobiologia **701**:159-172.

USEPA. 2017a. National Rivers and Streams Assessment 2018-19: Field Operations Manual–Wadeable. EPA-841-B-17-003a. U.S. Environmental Protection Agency, Office of Water, Washington, DC. .

USEPA. 2017b. National Rivers and Streams Assessment 2018-19: Laboratory Operations Manual. EPA841-B-17-004. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA. 2023. National Rivers and Streams Assessment 2018-2019 Technical Support Document. EPA 841-R-22-005. U.S. Environmental Protection Agency**,** Office of Water and Office of Research and Development, Washington, D.C. .

USEPA. 2024a. Climate change indicators in the Unites States. EPA 430-R-24-003.

USEPA. 2024b. National Rivers and Streams Assessment: The Third Collaborative Survey. U.S. Environmental Protection Agency, Office of Water and Office of Research and Development.

Van Sickle, J. 2008. An index of compositional dissimilarity between observed and expected assemblages. Journal of the North American Benthological Society **27**:227-235.

Vörösmarty, C. J., P. B. McIntyre, M. O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, S. E. Bunn, C. A. Sullivan, C. R. Liermann, and P. M. Davies. 2010. Global threats to human water security and river biodiversity. Nature **467**:555-561.

Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. 2015. So Many Variables: Joint Modeling in Community Ecology. Trends in Ecology & Evolution **30**:766-779.

Wiens, J. J., D. D. Ackerly, A. P. Allen, B. L. Anacker, L. B. Buckley, H. V. Cornell, E. I. Damschen, T. Jonathan Davies, J.-A. Grytnes, S. P. Harrison, B. A. Hawkins, R. D. Holt, C. M. McCain, and P. R. Stephens. 2010. Niche conservatism as an emerging principle in ecology and conservation biology. Ecology Letters **13**:1310-1324.

Wilkinson, D. P., N. Golding, G. Guillera-Arroita, R. Tingley, and M. A. McCarthy. 2021. Defining and evaluating predictions of joint species distribution models. Methods in Ecology and Evolution **12**:394-404.

Wood, G. 2008. Thresholds and criteria for evaluating and communicating impact significance in environmental statements: 'See no evil, hear no evil, speak no evil'? Environmental Impact Assessment Review **28**:22-38.

Yuan, L. L., D. M. Carlisle, R. M. Mitchell, and A. I. Pollard. 2024. Using models of local environmental conditions for biological assessment. Freshwater Science:000-000.

Yuan, L. L., C. P. Hawkins, and J. V. Sickle. 2008. Effects of regionalization decisions on an O/E index for the US national assessment. Journal of the North American Benthological Society **27**:892-905.

Zak, D., M. Hupfer, A. Cabezas, G. Jurasinski, J. Audet, A. Kleeberg, R. McInnes, S. M. Kristiansen, R. J. Petersen, H. Liu, and T. Goldhammer. 2021. Sulphate in freshwater ecosystems: A review of sources, biogeochemical cycles, ecotoxicological effects and bioremediation. Earth-Science Reviews **212**:103446.

Tables

Table 1: Sites, genera, and environmental variables included in the analysis. Revisit sites are locations that were revisited during the survey to assess within year variability. Values for the environmental variables are median values for each region. Range is given in parentheses. NTL = Total Nitrogen, PTL = Total Phosphorous, CL = Chloride, SO4 = Sulfate, RPDI = Riparian Disturbance Index, SUBD = Substrate Diameter, TPRCP = Total Precipitation and MSAT = Mean Summer Air Temperature. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric.

Region	Site (#)	Genera (#)	Revisit Sites (#)	NTL (ug/L)	PTL (ug/L)	CL (mg/L)	SO4 (mg/L)	RPDI	SUBD (mm)	TPRCP (mm)	MSAT (ºC)
CPL	226	71	36	786 (91-7713)	84.81 (5.2-3922.25)	10.06 (0.52-4797.9)	7.83 (0.11-1900.89)	0.5 (0-5.39)	0.35 (0.01-341.65)	1613.83 (388.66-2397.1)	27.46 (23.75-31.68)
NAP	228	127	29	431 (81-6413)	23.76 (3.12-587.2)	16.73 (0.07-668.16)	5.58 (0.06-467.93)	0.65 (0-5.24)	11.66 (0.01-864.9)	1295.41 (888.91-1883.17)	21.49 (17.24-24.72)
NPL	152	76	7	881.5 (71-15675)	85.61 (3.53-9248.31)	12.09 (0.11-1251.58)	486.24 (3.5-4079.78)	1.27 (0-4.49)	0.51 (0.01-560.5)	437.42 (157.84-1055.11)	20.52 (15.62-23.46)
SAP	266	111	32	557.5 (36-18700)	29.3 (3.79-4050)	5.89 (0.36-197.39)	7.46 (0.58-397.69)	0.76 (0-4.56)	16.28 (0.01-5656.85)	1475.78 (889.64-2451.81)	24.76 (19.19-27.73)
SPL	174	59	5	1165 (145-21175)	147.12 (5.21-4351.7)	30.36 (0.43-5220)	88.89 (1.96-3716.6)	1.04 (0-5.88)	0.35 (0.01-5656.85)	616.17 (240.1-1404.61)	25.88 (11.96-32.75)
TPL	223	74	29	1806 (236-16219)	165.08 (11.45-1066.87)	18.9 (1.44-736.62)	34.4 (5.79-1386.77)	0.83 (0-5.47)	0.35 (0.01-5656.85)	1004.22 (376.7-1801.08)	23.33 (18.28-27.09)
UMW	201	104	10	1168 (195-17675)	57.4 (8.16-659)	9.97 (0.01-306.69)	9.28 (0.04-160.4)	0.62 (0-6.43)	0.35 (0.01-812.79)	960.16 (496.44-1718.2)	20.84 (17.05-23.43)
WMT	225	94	19	133 (22-4719)	21.84 (2.71-569.8)	0.92 (0.04-521.64)	2.97 (0.07-1682.05)	0.48 (0-3.95)	38.14 (0.01-1288.61)	669.31 (175.89-3946.33)	18.08 (11.71-29.4)
XER	196	75	16	344.5 (46-8000)	54.62 (4.29-4667.41)	5.35 (0.1-1867.57)	27.8 (0.02-3286.24)	1.16 (0-4.36)	2.02 (0.01-368.11)	287.58 (100.46-1176.56)	22.76 (14.69-36.11)

Table 2: Random Forest model performance metrics for testing and out-of-bag training datasets. NTL = Total Nitrogen, PTL = Total Phosphorus, CL = Chloride, SO4 = Sulfate, SUBD = Substrate Diameter. RMSE = Root mean squared error of random forest models fitted with ln(x + 1) (NTL, PTL, and CL) or ln(x) (SO4) or Log10(SUBD) transformations.

Variable	R^2_{train}	$RMSE_{train}$	R^2_{test}	$RMSE_{test}$
ln(NTL+1) (ug/L)	0.73	0.61	0.72	0.60
ln(PTL+1) (ug/L)	0.60	0.78	0.69	0.70
ln(CL+1) (mg/L)	0.73	0.75	0.68	0.85
ln(SO4) (mg/L)	0.77	0.96	0.78	0.97
log_{10}(SUBD) (mm)	0.46	0.98	0.51	1.00

Table 3. Regression coefficients between predicted and observed richness and compositional similarity. Occurrence probabilities thresholds (Pr) were used to exclude genera with low predicted occurrence probabilities. We considered $R^2 \geq 0.2$, $-1.5 \leq$ intercept ≤ 1.5, and $0.85 \leq$ slope ≤ 1.15 are indicative of adequate model performance. Values in parentheses are the 5th and 95th percentile of estimates from the 3000 posterior samples. Compositional similarity was measured using a probabilistic adaptation of Jaccard index measured for each site. Parentheses are the 5th and 95th percentile of values for all sites.

		Genus Richness				Composition	
	Pr	Intercept	Slope	R^2	Revisit Sites within Posterior Distribution (%)	Mean Jaccard Similarity	Mean Jaccard Similarity at revisit sites
CPL	0.05	0.37 (-0.37 - 1.02)	1.00 (0.96 - 1.05)	0.85 (0.82 - 0.87)	97	0.52 (0.26 - 0.7)	0.68 (0.24-0.95)
NAP	0.05	-0.45 (-2.02 - 1.02)	1.03 (0.99 - 1.08)	0.74 (0.71 - 0.78)	90	0.52 (0.31 - 0.64)	0.35 (0.17-0.65)
NPL	0.05	-0.07 (-1.21 - 1.04)	1.02 (0.96 - 1.09)	0.76 (0.71 - 0.81)	100	0.58 (0.34 - 0.77)	0.39 (0.12-0.71)
SAP	0.05	0.40 (-0.66 - 1.41)	1.00 (0.97 - 1.04)	0.85 (0.82 - 0.87)	97	0.52 (0.26 - 0.67)	0.56 (0.21-0.87)
SPL	0.10	0.48 (-0.45 - 1.29)	1.02 (0.97 - 1.09)	0.79 (0.75 - 0.83)	100	0.52 (0.25 - 0.72)	0.61 (0.52-0.98)
TPL	0.05	0.06 (-0.91 - 0.94)	1.01 (0.97 - 1.06)	0.82 (0.79 - 0.84)	100	0.55 (0.31 - 0.77)	0.59 (0.18-1)
UMW	0.05	-0.74 (-2.52 - 0.99)	1.04 (0.98 - 1.10)	0.73 (0.66 - 0.78)	100	0.55 (0.34 - 0.69)	0.53 (0.27-0.9)
WMT	0.10	0.24 (-1.24 - 1.59)	1.04 (0.99 - 1.10)	0.74 (0.70 - 0.78)	100	0.60 (0.36 - 0.75)	0.45 (0.16-0.79)
XER	0.05	-0.12 (-1.11 - 0.74)	1.03 (0.98 - 1.08)	0.83 (0.80 - 0.86)	94	0.50 (0.23 - 0.71)	0.48 (0.07-0.83)

Table 4: Mean proportion of sites in each taxonomic group that was identified as an increaser or decreaser. Increasers (I) are genera that have a significantly higher probability of occurrence under present day conditions compared to hindcasted conditions; decreasers (D) are genera that have a significantly lower probability of occurrence under present-day conditions compared to hindcasted conditions. The values in the table are the mean proportion of sites for genera within the major taxonomic group. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = xeric. EPT = Ephemeroptera, Plecoptera and Trichoptera.

	CPL		NAP		NPL		SAP		SPL		TPL		UMW		WMT		XER	
	D	I	D	I	D	I	D	I	D	I	D	I	D	I	D	I	D	I
Insects	0.16	0.08	0.08	0.05	0.1	0.17	0.1	0.11	0.09	0.12	0.16	0.11	0.11	0.12	0.03	0.06	0.09	0.05
EPT	0.24	0.01	0.07	0.04	0.17	0.11	0.13	0.08	0.09	0.04	0.23	0.04	0.14	0.03	0.04	0.06	0.1	0.03
CHIRONOMIDAE	0.11	0.09	0.09	0.05	0.04	0.18	0.04	0.12	0.02	0.17	0.12	0.14	0.06	0.16	0.02	0.07	0.06	0.1
Other Insects	0.21	0.07	0.08	0.08	0.07	0.21	0.13	0.12	0.23	0.02	0.18	0.11	0.09	0.04	0.02	0.03	0.11	0.01
Non-Insects	0.14	0.16	0.06	0.18	0.06	0.19	0.11	0.13	0.06	0.13	0.15	0.15	0.08	0.14	0.03	0.08	0.04	0.14
ARTHROPODA	0.21	0.11	0.07	0.16	0.11	0.02	0.16	0.09	0.05	0.06	0.23	0.35	0.1	0.04	0.04	0.03	0.05	0.04
MOLLUSCA	0.08	0.11	0.03	0.28	0.02	0.3	0.1	0.15	0.09	0.06	0.21	0.05	0.14	0.07	0.01	0.06	0.01	0.26
Other Non-Insects	0.15	0.22	--	0.14	0.05	0.13	0.04	0.16	0.02	0.19	0.05	0.15	0.01	0.25	0.04	0.14	0.04	0.14

Figure Captions

Figure 1: Ecoregions and survey locations for the National Rivers and Streams Assessment 2018-2019 survey. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric

Figure 2: Conceptual diagram of posterior distributions generated from bayesian inference. The grey is a posterior distribution of genus richness generated after removing anthropogenic disturbances (Hindcasted) and the white is a posterior distribution generated from present-day conditions. The dotted line is the expected value (i.e. mean) of the present day posterior distribution and dark shading represents the lower 10% or the upper 90% of the HC posterior distribution. At site A, the PD mean is within the <10% of the HC distribution and indicates that PD richness is likely lower than HC richness. Alternatively, at site B, PD richness is >90% of the HC distribution and indicates that PD richness is likely higher. At site C, the two posterior distributions are similar such that these is likely no difference between PD and HC.

Figure 3: Partial dependence plots showing the effects of anthropogenic variables on in-stream physiochemical factors. For visualization, each anthropogenic factor was rescaled between 0-1 and labeled low, medium, and high. CoalMineDen = coal mine density, MineDen = gravel mine density, MSAT = mean summer air temperature, N_dep = atmospheric nitrogen deposition, N_input = anthropogenic nitrogen inputs, P_input = anthropogenic phosphorous inputs, PctCrop = percent crop in the watershed, PctCropRP = percent crop in the riparian area, PctNatRP = percent natural vegetation in riparian area, RdDen = road density, S_dep = atmospheric sulfur deposition, TPRCP = total precipitation, and W1_HAG = agricultural disturbance adjacent to stream reach. NTL = Total Nitrogen (ug/L), PTL = Total Phosphorus (ug/L), CL = Chloride (mg/L), SO4 = Sulfate (mg/L), SUBD = Substrate Diameter Log10(mm).

Figure 4: Observed versus hindcasted values for each environmental gradient. Points are regional means and vertical and horizontal bars represent the 10^{th} and 90^{th} quantiles of observed values or hindcasted values within each region, respectively. For RPDI, all hindcast values were < 0.33 and the ecoregions were plotted separately for visualization. The dashed line is the 1:1 relationship for all plots except for RPDI where it represents the 0.33 threshold applied to all ecoregions. NTL = Total Nitrogen (ug/L), PTL = Total Phosphorus (ug/L), CL = Chloride (mg/L), SO4 = Sulfate (mg/L), SUBD = Substrate Diameter (mm), RPDI = Riparian Disturbance Index, MSAT = mean summer air temperature (ºC) and TPRCP = total precipitation (mm). CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric. See Appendix S1: Table S4 for observed and hindcasted mean values and quantiles for each region.

Figure 5: A) Proportion of sites where hindcasted abiotic conditions were >2 standard deviations from the observed value for Total Nitrogen (NTL), Total Phosphorus (PTL), Chloride (CL), Sulfate (SO4), Substrate Diameter (SUBD), mean summer air temperature (MSAT), and Total Precipitation (TPRCP) and > 0.33 for Riparian Disturbance Index (RPDI). B) The number of environmental variables affected by human disturbance. Locations where the effects were not detected are plotted separately. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric.

Figure 6: Difference in macroinvertebrate genus richness after removing anthropogenic disturbance from each category of physiochemical variables (A) and all variables simultaneously (B). The points in each panel are sites that had a change in genus richness with >75% support after removing disturbance.

Figure 7: Percentage of the population of streams where genus richness or composition could change given a hindcasted physiochemical environment. Shaded bars indicate that the probability of mean present-day richness differs from hindcast with > 0.90 (Black) or 0.75-0.90 (Gray) support. Hatched bars indicate that Jaccard similarity was < 0.9 but support for a difference in richness were < 0.75. White bars indicate the proportion of sites that had < 0.75 support for change in richness and >0.9 compositional similarity. Yellow bars indicate the proportion of sites that could not be assessed because of insufficient data or potential extrapolation from predicting hindcast physiochemical conditions. Regions are arranged according to richness. NPL = Northern Plains, TPL = Temperate Plains, SPL = Southern Plains, XER = Xeric, SAP = Southern Appalachians, CPL = Coastal Plains, NAP = Northern Appalachians, UMW = Upper Midwest and WMT = Western Mountains.

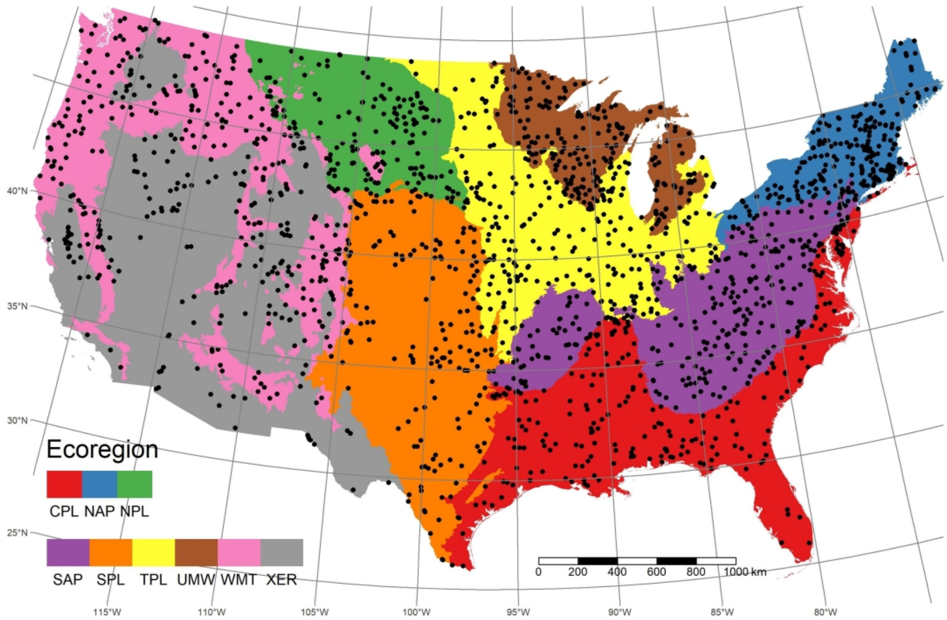


Figure 1: Ecoregions and survey locations for the National Rivers and Streams Assessment 2018-2019 survey. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric

279x177mm (300 x 300 DPI)

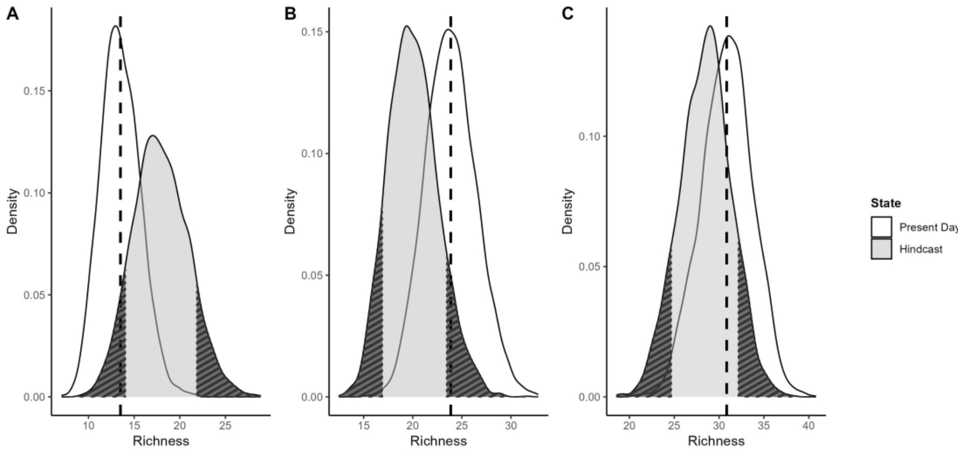


Figure 2: Conceptual diagram of posterior distributions generated from bayesian inference. The grey is a posterior distribution of genus richness generated after removing anthropogenic disturbances (Hindcasted) and the white is a posterior distribution generated from present-day conditions. The dotted line is the expected value (i.e. mean) of the present day posterior distribution and dark shading represents the lower 10% or the upper 90% of the HC posterior distribution. At site A, the PD mean is within the <10% of the HC distribution and indicates that PD richness is likely lower than HC richness. Alternatively, at site B, PD richness is >90% of the HC distribution and indicates that PD richness is likely higher. At site C, the two posterior distributions are similar such that these is likely no difference between PD and HC.

1164x529mm (72 x 72 DPI)

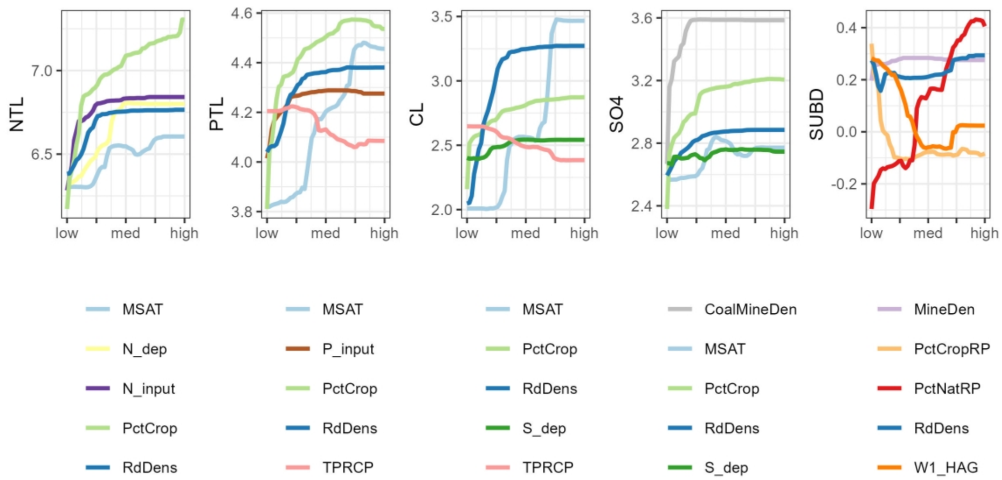


Figure 3: Partial dependence plots showing the effects of anthropogenic variables on in-stream physiochemical factors. For visualization, each anthropogenic factor was rescaled between 0-1 and labeled low, medium, and high. CoalMineDen = coal mine density, MineDen = gravel mine density, MSAT = mean summer air temperature, N_dep = atmospheric nitrogen deposition, N_input = anthropogenic nitrogen inputs, P_input = anthropogenic phosphorous inputs, PctCrop = percent crop in the watershed, PctCropRP = percent crop in the riparian area, PctNatRP = percent natural vegetation in riparian area, RdDen = road density, S_dep = atmospheric sulfur deposition, TPRCP = total precipitation, and W1_HAG = agricultural disturbance adjacent to stream reach. NTL = Total Nitrogen (ug/L), PTL = Total Phosphorus (ug/L), CL = Chloride (mg/L), SO4 = Sulfate (mg/L), SUBD = Substrate Diameter Log10(mm).

846x423mm (72 x 72 DPI)

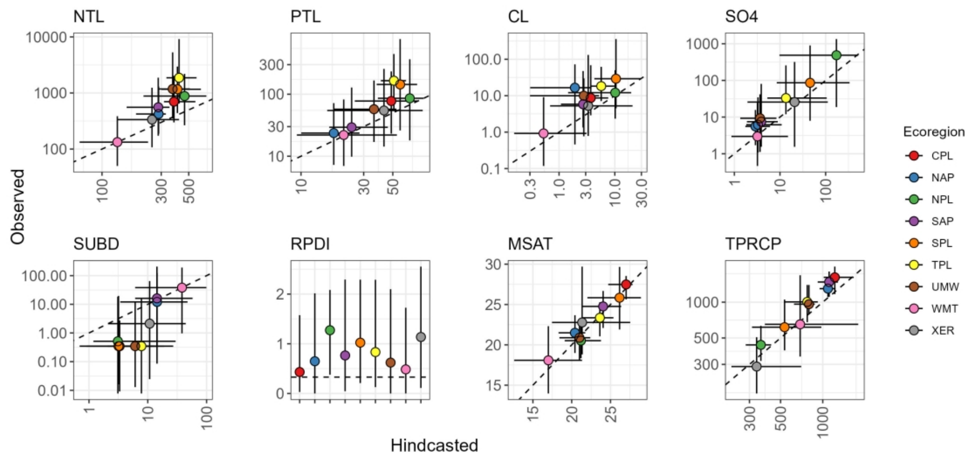


Figure 4: Observed versus hindcasted values for each environmental gradient. Points are regional means and vertical and horizontal bars represent the 10th and 90th quantiles of observed values or hindcasted values within each region, respectively. For RPDI, all hindcast values were < 0.33 and the ecoregions were plotted separately for visualization. The dashed line is the 1:1 relationship for all plots except for RPDI where it represents the 0.33 threshold applied to all ecoregions. NTL = Total Nitrogen (ug/L), PTL = Total Phosphorus (ug/L), CL = Chloride (mg/L), SO4 = Sulfate (mg/L), SUBD = Substrate Diameter (mm), RPDI = Riparian Disturbance Index, MSAT = mean summer air temperature (ºC) and TPRCP = total precipitation (mm). CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric. See Appendix S1: Table S4 for observed and hindcasted mean values and quantiles for each region.

1164x529mm (72 x 72 DPI)

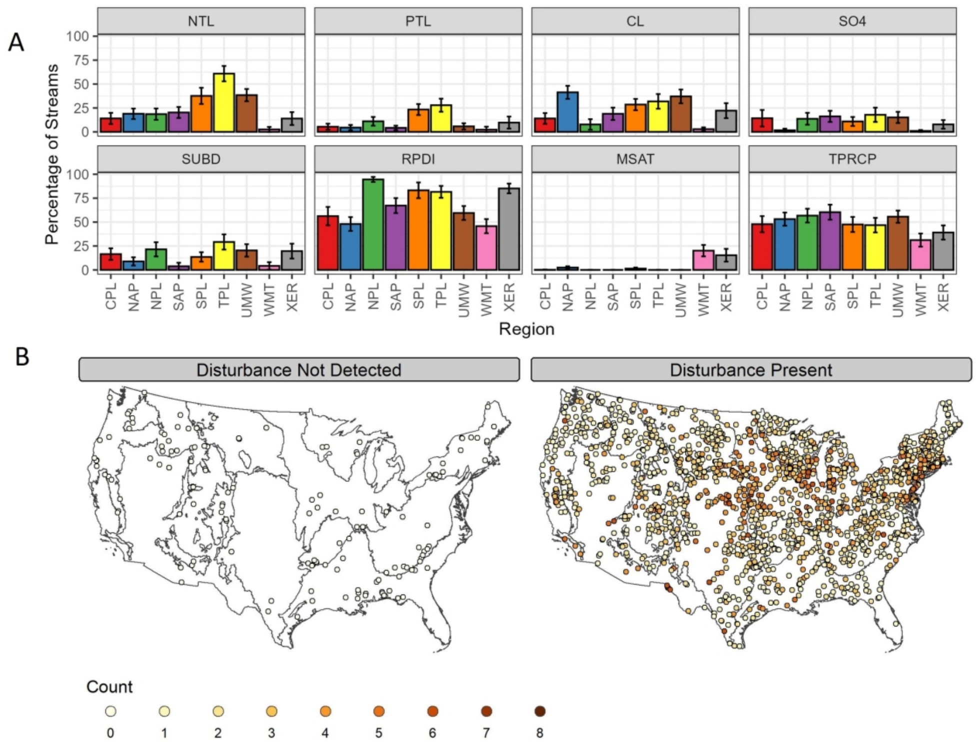


Figure 5: A) Proportion of sites where hindcasted abiotic conditions were >2 standard deviations from the observed value for Total Nitrogen (NTL), Total Phosphorus (PTL), Chloride (CL), Sulfate (SO4), Substrate Diameter (SUBD), mean summer air temperature (MSAT), and Total Precipitation (TPRCP) and > 0.33 for Riparian Disturbance Index (RPDI). B) The number of environmental variables affected by human disturbance. Locations where the effects were not detected are plotted separately. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric.

218x163mm (300 x 300 DPI)

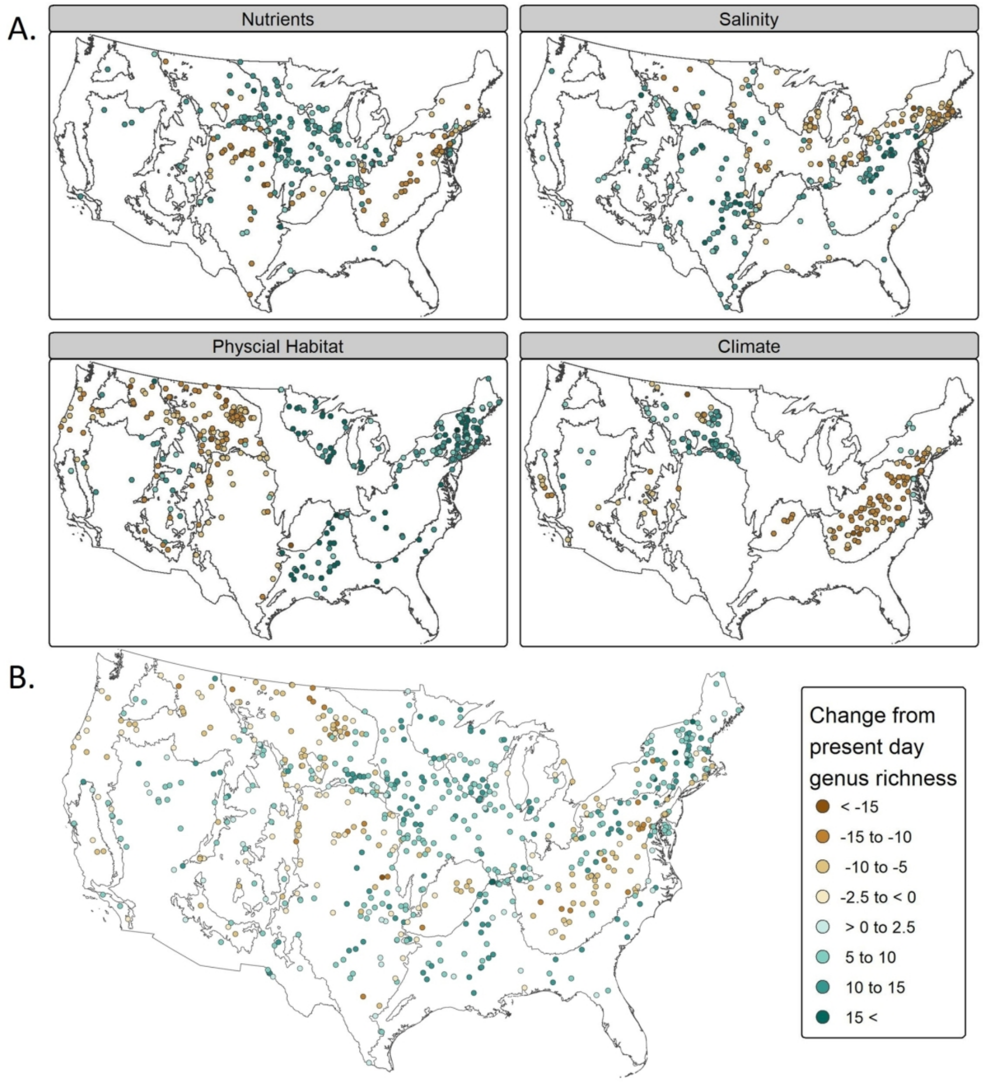


Figure 6: Difference in macroinvertebrate genus richness after removing anthropogenic disturbance from each category of physiochemical variables (A) and all variables simultaneously (B). The points in each panel are sites that had a change in genus richness with >75% support after removing disturbance.

175x191mm (300 x 300 DPI)

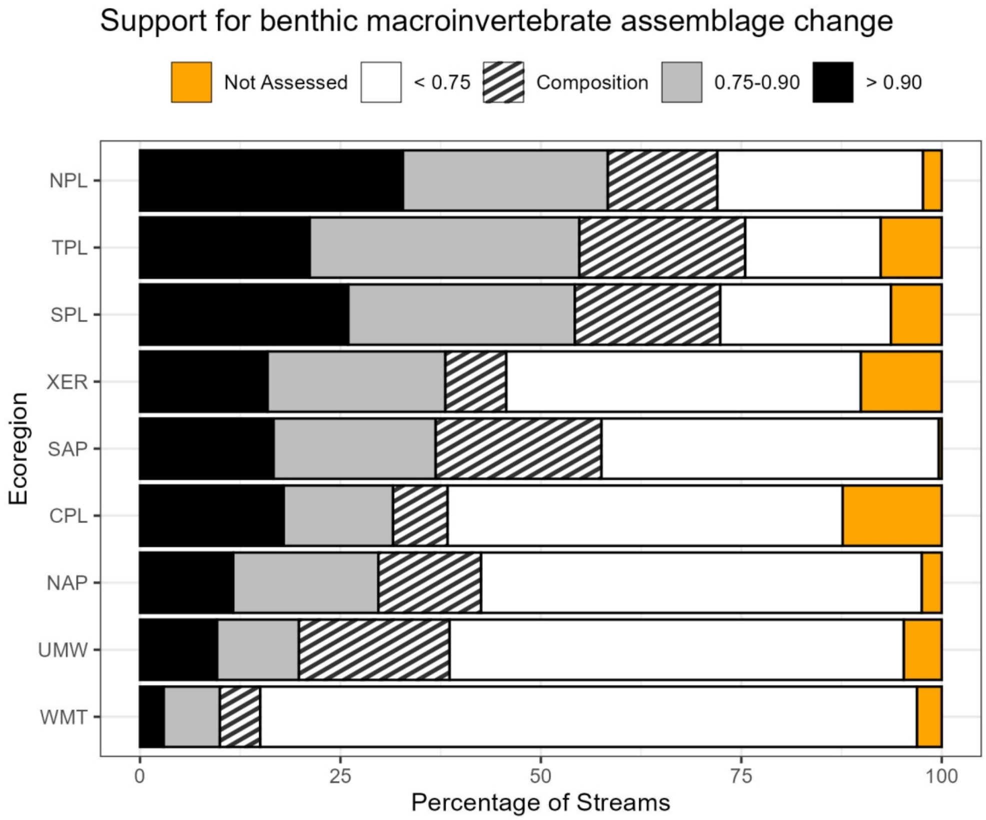


Figure 7: Percentage of the population of streams where genus richness or composition could change given a hindcasted physiochemical environment. Shaded bars indicate that the probability of mean present-day richness differs from hindcast with > 0.90 (Black) or 0.75-0.90 (Gray) support. Hatched bars indicate that Jaccard similarity was < 0.9 but support for a difference in richness were < 0.75. White bars indicate the proportion of sites that had < 0.75 support for change in richness and >0.9 compositional similarity. Yellow bars indicate the proportion of sites that could not be assessed because of insufficient data or potential extrapolation from predicting hindcast physiochemical conditions. Regions are arranged according to richness. NPL = Northern Plains, TPL = Temperate Plains, SPL = Southern Plains, XER = Xeric, SAP = Southern Appalachians, CPL = Coastal Plains, NAP = Northern Appalachians, UMW = Upper Midwest and WMT = Western Mountains.

635x529mm (72 x 72 DPI)

Supplementary Information

A model-based assessment of anthropogenic disturbance on lotic macroinvertebrate assemblages

Darin A. Kopp[1*], John L. Stoddard[1], Philip R. Kaufmann[1,2], Alan T. Herlihy[2], Ryan A. Hill[1], Meredith M. Brehob[3], Robert D Sabo[4]

[1]US EPA, Office of Research and Development, Corvallis, Oregon, USA

[2]Department of Fisheries, Wildlife, & Conservation Sciences, Oregon State University, Corvallis, Oregon, USA

[3]Oak Ridge Institute for Science and Education Participant c/o US EPA, ORD, Washington, D.C., USA

[4]US EPA, Office of Research and Development, Center for Public Health and Environmental Assessment, Washington, DC, USA

*Correspondence: kopp.darin@epa.gov

Keywords: Joint Species Distribution Modeling, Richness, National Rivers and Streams Assessment, Benthic Macroinvertebrate, Reference Sites, National Aquatic Resource Surveys.

Table S1: Geoclimatic and anthropogenic variables used in random forest models. Variables were selected based on their hypothesized relationship with total nitrogen, total phosphorus, chloride, sulfate and substrate diameter. StreamCat data is publicly accessible at https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset (Hill et al. 2016) and National Atmospheric Deposition Program Data is publicly accessible at https://nadp.slh.wisc.edu/maps-data/ntn-gradient-maps/

Variable	Description	Source
BFIWs	The component of streamflow that can be attributed to ground-water discharge	StreamCat, Hill et al. 2016
ClayWs	Mean % clay content of soils in watershed	StreamCat
ElevWs	Mean elevation of watershed (m)	StreamCat
KffactWs	Mean of STATSGO Kffactor	StreamCat
DepClWs	Mean atmospheric chloride deposition 2018	National Atmospheric Deposition Program
DepPWs	Atmospheric phosphorous deposition 2013	Sabo et al 2023
NWs	Rock derived nitrogen in watershed	StreamCat
P2O5Ws	Mean % of lithological phosphorous oxide (P2O5) content in surface or near surface geology	StreamCat
PctAlluvCoastWs	% of watershed area classified with lithology type as alluvium and fine-textured coastal zone sediment	StreamCat
PctHbWetWs	% of watershed area classified as herbaceous wetland land cover	StreamCat
PermWs	Mean permeability (cm/hour) of soils (STATSGO) within watershed	StreamCat
RunoffWs	Mean runoff (mm) within watershed	StreamCat
SandWs	Mean % sand content of soils (STATSGO) within watershed	StreamCat
StreamPower	An index reflecting the amount of energy the water exerts on the sides and bottom of a stream	(((TPRCP - ET)/1000)*(0.032*WSArea)^0.5) * slope
SWs	Mean % of lithological sulfur (S) content in surface or near surface geology within watershed	StreamCat
WsAreaSqKm	Area of watershed (square km)	StreamCat
CoalMineDensWs	Density of coal mines sites within watershed (mines/square km)	StreamCat
DepNWs	Mean atmospheric nitrogen deposition in watershed 2018	National Atmospheric Deposition Program
DepSWs	Mean atmospheric sulfur deposition in watershed 2018	National Atmospheric Deposition Program
MineDensWs	Density of coal mines sites within watershed (mines/square km)	StreamCat
N_inputs	Sum of anthropogenic nitrogen inputs: N_Fert_FarmWs, N_Fert_UrbanWs, N_Human_WasteWs, N_Livestock_WasteWs	Sabo et al. 2023
P_inputs	Sum of anthropogenic phosphorus inputs: P_f_fertilizerWs, P_human_wasteWs, P_livestock_WasteWs, P_nf_fertilizerWs	Sabo et al. 2023
PctCropWs	% of watershed area classified as crop land use	StreamCat
PctCropWsRp100	% of watershed riparian area classified as crop land use	StreamCat
PctNatWs	% of watershed with natural vegetation cover	StreamCat
TPRCP	Ln(Mean total annual precipitation, 2018 or 2019)	PRISM
RdDensWs	Density of roads (2010 Census Tiger Lines) within watershed (km/square km)	StreamCat
MSAT	Mean summer (July/August) air temperature (C°), 2018 or 2019	PRISM
W1_HAG	Agricultural disturbance adjacent to a stream	NRSA Field Data
NABD_NrmStorWs	Volume all reservoirs per unit area of watershed (cubic meters/square km)	StreamCat

Table S2: Variable importance rankings for random forest models and values used to estimate physiochemical conditions in the absence of anthropogenic disturbance. The 5 most important variables for each model are bold.

Predictor Variable	NTL	PTL	CL	SO4	SUBD	Range of Predictor Present-day	Value Without Disturbance
Geoclimatic							
BFIWs	6	13	7	8	-	2.6 - 88	--
ClayWs	-	10	8	12	-	2.35 - 60.6	--
ElevWs	10	**4**	6	9	-	2.02 - 3587.98	--
KffactWs	13	11	15	14	9	0.01 - 0.48	--
DepClWs	-	-	9	-	-	0.13-43.16	--
DepPWs	-	**5**	-	-	-	2.03-15.9	--
NWs	9	-	11	**4**	-	0-1.59	--
P2O5Ws	16	9	18	15	-	0.02-2.42	--
PctAlluvCoastWs	-	-	-	-	6	0-100	--
PctHbWetWs	7	8	10	16	-	0-40.35	--
PermWs	14	14	17	13	-	0.32-45.23	--
RunoffWs	**2**	**2**	**4**	**1**	**2**	0.05-3731.74	--
SandWs	-	-	-	-	8	2.8-92.04	--
StreamPower	-	-	-	-	**1**	0-0.44	--
SWs	12	15	**5**	**3**	-	0.01-23.66	--
WsAreaSqKm	15	12	14	11	**5**	0.21-2,874,021.46	--
Anthropogenic							
CoalMineDensWs	-	-	19	7	13	0 - 5.74	0
DepNWs	**5**	-	-	-	-	1.12 - 26.15	5 kgN/ha/yr*
DepSWs	-	-	12	10	-	0.22 - 5.88	1.65 kgS/ha/yr*
MineDensWs	-	-	16	18	12	0 - 0.13	0
N_inputs	**3**	-	-	-	-	0 - 6.22	0
P_inputs	-	7	-	-	-	0 - 2.19	0
PctCropWs	**1**	**1**	**3**	**2**	-	0 - 96.7	0
PctCropWsRp100	-	-	-	-	**4**	0 - 95.52	0
PctNatWs	-	-	-	-	**3**	0 - 100	100
TPRCP	11	16	13	17	-	4.61 - 8.28	Average 1900-1950
RdDensWs	**4**	6	**2**	5	10	0 - 15.15	0
MSAT	8	**3**	**1**	6	-	11.71 - 36.11	Average 1900-1950
W1_HAG	-	-	-	-	7	0 - 2.18	0
NABD_NrmStorWs	-	-	-	-	11	0 – 1,886,390.68	0

* Estimates for atmospheric nitrogen and sulfur deposition were obtained from Clark et al. 2018. Although they used 0.4 kgN/ha and 0.1kgS/ha no sites had deposition values below this level. Instead, we selected the higher values of 3-5KgN/ha as reasonable estimates for background deposition values before 1900. For S deposition, other efforts have estimated pre-industrial S deposition at 0.32–2.98 kgS/ha (Granat et al. 1976, Fakhraei et al. 2016) and we used the used middle number of this interval as a reasonable estimate for background.

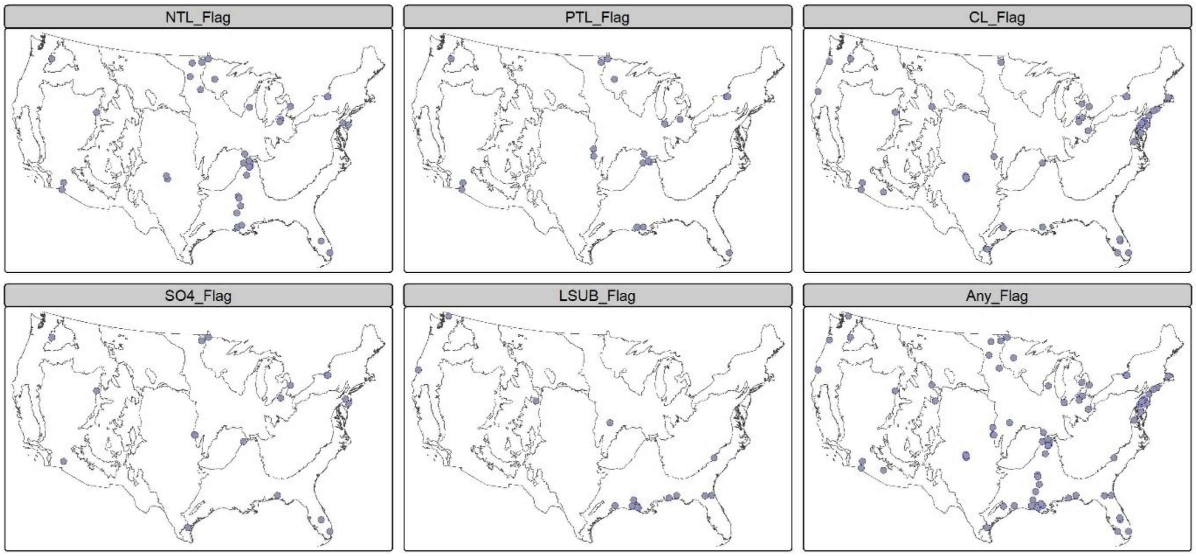


Figure S1: Sites flagged for potential extrapolation after removing anthropogenic effects from the random forest models. NTL = total nitrogen; PTL is total phosphorus; CL is chloride; SO4 is sulfate; LSUB = substrate diameter, Any = sites that were flagged for extrapolation for any model and removed from the analysis. The number of sites removed is as follows (CPL = 40 (18%); NAP = 2 (1%); NPL = 1 (1%); SAP = 1 (<1%); SPL = 6 (3%); TPL = 13 (6%); UMW=3 (1%); WMT = 2(1%) and XER = 5 (3%))

Table S3. Variance inflation factors (VIF) for predictor variables used for joint species distribution models. In general, VIF >5 indicates a strong correlation between variables while values VIF < 3 indicate low correlation between values. Bold font = VIF >3; NTL = Total Nitrogen, PTL = Total Phosphorus, CL = Chloride, SO4 = Sulfate, SUBD = Substrate Diameter, RPDI = Riparian Disturbance index, MSAT = Mean Summer Air Temperature, PRCP = Total Annual Precipitation.

	NTL (ug/L)	PTL (ug/L)	CL (mg/L)	SO4 (mg/L)	SUB (mm)	RPDI	MSAT (DegC)	TPRCP (mm)
CPL	2.17	1.87	2.83	2.5	1.09	1.3	1.25	1.14
NAP	**3.08**	2.6	**3.34**	1.67	1.1	1.28	2.47	1.07
NPL	2.8	2.2	2.82	2.94	1.39	1.11	1.84	1.5
SAP	2.42	1.85	2.6	1.49	1.13	1.3	1.21	1.04
SPL	2.49	2.57	2.65	2.12	1.22	1.17	1.63	1.28
TPL	1.57	1.35	1.22	1.45	1.11	1.11	1.52	1.5
UMW	2.37	1.92	**3.13**	2.57	1.09	1.46	**4.25**	1.15
WMT	1.79	1.6	2.22	1.86	1.25	1.34	1.47	1.29
XER	2.28	1.77	**4.1**	**3.23**	1.29	1.19	1.69	1.44

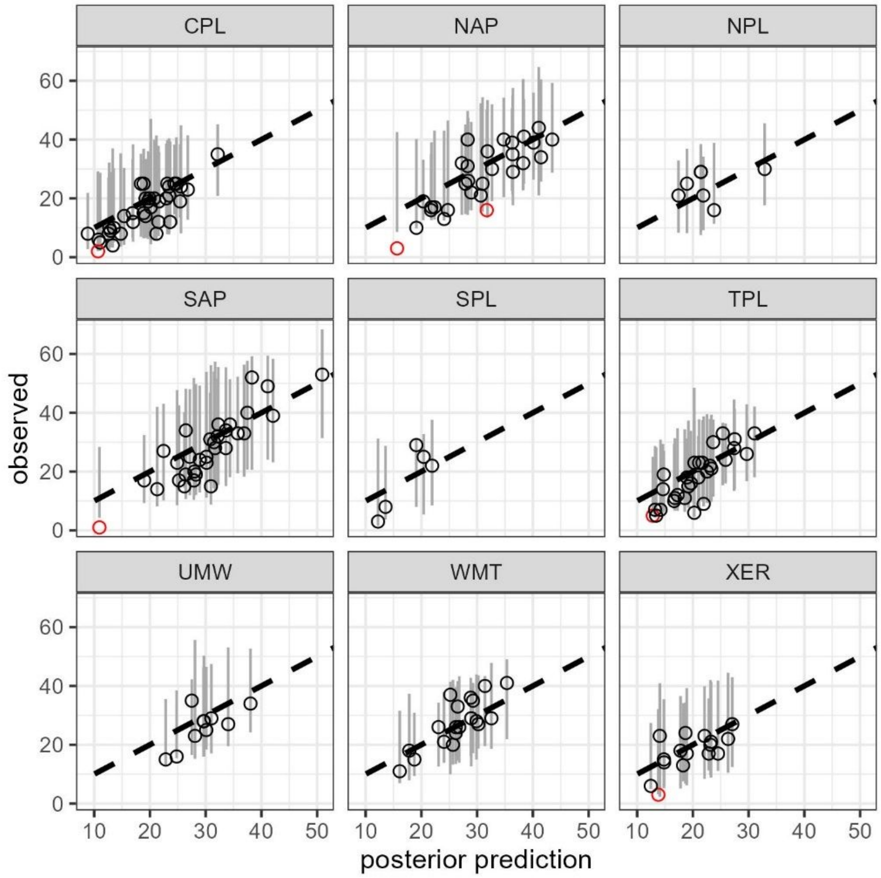


Figure S2. Validation of JSDM using revisit surveys. As part of NRSA, a subsample of sites are resurveyed. We used these data to provide additional validation of the predictive capabilities of JSDMs. We found that when using condition prediction (Ovaskainen and Abrego 2020), the posterior distribution of all JSDM typically contained all observed genus richness values. This result suggests that the observed genus richness could be a random sample drawn from our modeled posterior distribution and suggests that our models can accurately predict the occurrence of taxa during the initial visit. Black points indicate that the observed genus richness is within the predicted posterior distribution; Red points indicate that the observed genus richness is outside the predicted posterior distribution; Dotted line is 1:1 between observed and posterior mean richness and grey lines represent the distribution of 3,000 samples from the posterior distribution.

Table S4. Regional means of present-day (PD) and hindcasted (HC) values for each environmental gradient. Parentheses are 10th and 90th quantiles of observed hindcasted values. NTL = Total Nitrogen, PTL = Total Phosphorus, CL = Chloride, SO4 = Sulfate, SUBD = Substrate Diameter, RPDI = Riparian Disturbance index, MSAT = Mean Summer Air Temperature, PRCP = Total Annual Precipitation.

	NTL (ug/L)		PTL (ug/L)		CL (mg/L)		SO4 (mg/L)		SUB (mm)		RPDI		MSAT (DegC)		TPRCP (mm)	
	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC
CPL	702.5 (299.2-2027.2)	380 (315.4-566.4)	77.82 (23.07-234.27)	48.67 (25.67-77.47)	8.7 (2.89-68.94)	3.7 (2.92-7.65)	6.3 (1.08-34.55)	3.29 (1.82-11.02)	0.35 (0.01-2.73)	3.32 (0.7-14.32)	0.44 (0.03-1.58)	0.33 (0.03-0.33)	27.49 (25.63-28.54)	27.02 (24.69-27.84)	1623.25 (1178.37-1990.68)	1219.49 (1015.53-1664.18)
NAP	422 (187-1301)	286 (188-327)	23.62 (7.25-71.81)	17.76 (10.04-27.89)	16.58 (0.46-72.61)	1.92 (0.3-3.82)	5.57 (1.71-15.63)	2.9 (1.74-6.07)	12.09 (0.09-122.43)	14.36 (3.4-47.45)	0.65 (0-2.02)	0.33 (0-0.33)	21.49 (18.99-23.67)	20.41 (18.31-21.75)	1299.05 (1106.82-1629.65)	1089.44 (949.5-1234.19)
NPL	881.5 (265.9-2193.5)	462.75 (216-694)	85.89 (18.01-366.21)	67.33 (38.47-85.9)	12.09 (2.34-50.02)	10.16 (1.45-19.62)	486.24 (18.73-1340.1)	172.31 (9.75-517.88)	0.51 (0.01-18.45)	3.1 (1.19-29.75)	1.27 (0.38-2.08)	0.33 (0.33-0.33)	20.52 (18.34-22.49)	21.19 (19.1-23.77)	437.42 (320.65-632.12)	361.68 (282.17-428.71)
SAP	556 (173-1876)	284 (158-350)	29.26 (9.83-128.25)	24.3 (16.41-45.85)	5.82 (1.34-47.53)	2.7 (1.08-3.82)	7.47 (1.55-80.36)	3.88 (1.88-8.95)	16.13 (0.35-206.69)	14.32 (6.1-57.76)	0.77 (0.05-2.29)	0.33 (0.05-0.33)	24.76 (22.09-26.68)	24.05 (21.37-26.37)	1476.55 (1108.36-1809.43)	1110.8 (920.87-1283.9)
SPL	1165 (438.5-2961)	405.25 (301.75-627)	143.62 (26.05-774.69)	56.88 (37.68-76.58)	29.21 (4.51-358.38)	10.47 (4.29-29.42)	86.1 (7.94-898.01)	45.6 (8.5-334.85)	0.35 (0.02-19.61)	3.18 (0.92-12.68)	1.02 (0.21-2.29)	0.33 (0.21-0.33)	25.84 (21.9-29.66)	26.17 (22.03-28.92)	616.17 (393.76-1051.75)	534.19 (312.86-977.73)
TPL	1868 (647.6-9147.8)	418 (330.6-578)	165.08 (54.39-444.97)	50.96 (40.55-63)	18.29 (7.1-62.09)	5.7 (3.7-10.38)	32.99 (11.93-256.93)	13.48 (5.28-105.55)	0.35 (0.01-16.28)	7.79 (2.51-18.96)	0.83 (0.13-2.29)	0.33 (0.13-0.33)	23.36 (21.05-25.28)	23.63 (21.3-25.38)	1004.99 (678.76-1395.29)	775.15 (537.81-922.38)
UMW	1188 (352.2-5320)	371 (289.6-511.49)	57.4 (16.93-166.88)	36.03 (17.9-53.25)	9.98 (0.76-29.9)	2.72 (0.3-5.21)	9.28 (1.12-40.68)	3.65 (1.34-8.36)	0.35 (0.01-12.6)	6.1 (2.19-27.43)	0.62 (0-2.1)	0.33 (0-0.33)	20.84 (18.13-22.56)	20.98 (18.42-22.38)	961.61 (801.29-1395.23)	799.34 (623.5-877.62)
WMT	133 (50-385.6)	133 (66-235.1)	22.12 (6.97-82.29)	21.14 (9.3-53.81)	0.92 (0.12-9.14)	0.52 (0.16-3.13)	2.97 (0.47-24.57)	3.2 (0.86-15.04)	38.14 (0.97-192.14)	38.17 (6.1-99.8)	0.48 (0-1.73)	0.33 (0-0.33)	18.08 (13.97-22.33)	16.96 (12.57-21.33)	648.86 (349.11-1670.25)	688.27 (386.2-1778.26)

XER	333 (107.2 - 1245.6)	253 (133.8-418)	53.87 (14.3-240.71)	42.76 (15.83-84.65)	5.23 (0.79-132.36)	3.17 (0.68-20.77)	25.08 (1.54-324.3)	20.85 (1.93-112.43)	2.18 (0.02-66.09)	10.78 (2.1-44.24)	1.14 (0.11-2.54)	0.33 (0.11-0.33)	22.76 (18.81 - 29.64)	21.33 (17.46 - 26.95)	287.58 (170.89-511.45)	337.66 (223.14-702.18)

A model-based assessment of anthropogenic disturbance on lotic macroinvertebrate assemblages

Darin A. Kopp[1,~~2~~]*, John L. ~~Stoddard[2]~~Stoddard[1], Philip R. ~~Kaufmann[2,3]~~Kaufmann[1,2], Alan T. ~~Herlihy[3]~~Herlihy[2], Ryan ~~Hill[2]~~A. Hill[1], Meredith ~~Brehob[5]~~M. Brehob[3], Robert ~~Sabo[6]~~D Sabo[4]

~~[1]Oak Ridge Institute for Science and Education Participant c/o US EPA, ORD, Corvallis, Oregon~~

~~[2]US~~[1]US EPA, Office of Research and Development, Corvallis, Oregon, USA

~~[3]Department~~[2]Department of Fisheries, Wildlife, & Conservation Sciences, Oregon State University,

Corvallis, Oregon~~-~~, USA

~~[4]Current affiliation: Oregon Department of Fish and Wildlife, Salem, Oregon~~

~~[5]Oak~~[3]Oak Ridge Institute for Science and Education Participant c/o US EPA, ORD, Washington, D.C~~.~~., USA

~~[6]US~~[4]US EPA, Office of Research and Development, Center for Public Health and Environmental Assessment, ~~DC,~~ Washington, DC, USA

*Correspondence: kopp.darin@epa.gov

Keywords: Joint Species Distribution Modeling, Richness, National Rivers and Streams Assessment, Benthic Macroinvertebrate, Reference Sites, National Aquatic Resource Surveys.

Open research statement: Benthic macroinvertebrate occurrence data and physiochemical variables were retrieved from the 2018-2019 National Rivers and Stream Assessment (https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys) Benthic macroinvertebrate data can be downloaded as .csv file from NRSA 1819 Benthic Macroinvertebrate Count - Data (CSV) (csv). Physiochemical data can be downloaded as .csv files from NRSA 1819 Water Chemistry_CHLA - Data (CSV) (csv) and NRSA 1819 Physical Habitat Larger Set of Metrics - Data (CSV) (csv) Landscape variables were retrieved from https://www.epa.gov/national-aquatic-resource-surveys/streamcat-metrics-and-definitions and PRISM Climate Group https://prism.oregonstate.edu/

Abstract:

Traditionally, the effects of anthropogenic disturbance on biological assemblages are elucidated by comparing an assemblage observed at a site to one that represents a minimally disturbed state. Unfortunately, defining a minimally disturbed state is extremely challenging because of the extent of human disturbance. We use a national scale dataset and a two-stage ~~modeling~~model-based approach to assess how benthic macroinvertebrate assemblages at 1,~~824~~748 sites would change if common anthropogenic disturbances were removed from in-stream physiochemical variables. First, we used random forest models and current landscape data to predict physiochemical conditions and then infer abiotic condition in the absence of disturbance. Second, we combined these estimates with joint species distribution models to predict the assemblage that would occur in these undisturbed conditions. Random forest models explained 48 – 75% of the variation in total nitrogen, phosphorous, sulfate, chloride, and substrate diameter. Generally, nutrient and salinity concentrations were higher, and substrates were finer than predicted to be without disturbances. Using this physiochemical data, joint species distribution models accurately explained genus richness ($R^2 = 0.73 – 0.85$) and composition (Jaccard similarity index = 0.48 – 0.55). Depending on the ecoregion, we found that genus richness could change at 26 – 61% of sites if disturbance was removed. Different

responses were observed for insect and non-insect taxa. For example, under anthropogenic disturbance, occurrence probabilities for Ephemeroptera, Plecoptera and Trichoptera tended to decrease at 5 – 26% of sites while occurrence probabilities for Mollusca and other non-insect, non-arthropod taxa increased at 5 – 33% and 11 – 24% of sites, respectively. Importantly, our framework provides an avenue for evaluating the effects of anthropogenic disturbance on macroinvertebrate assemblages without ~~relying on defining~~identifying reference sites.

Introduction:

Quantifying the extent and magnitude of anthropogenic disturbance on ecosystems requires a benchmark that represents a desired, expected, or previous condition ~~(Stoddard et al. 2006, Hawkins et al. 2010, McNellie et al. 2020)~~(Stoddard et al. 2006, Hawkins et al. 2010, McNellie et al. 2020). Ideally, benchmarks should represent chemical, physical, and biological characteristics with minimal anthropogenic disturbance ~~(Reynoldson et al. 1997, Stoddard et al. 2006)~~(Reynoldson et al. 1997, Stoddard et al. 2006). However, for many ecosystems the minimally disturbed condition is difficult to quantify or may no longer exist ~~(Dudgeon et al. 2006, Vörösmarty et al. 2010, McNellie et al. 2020)~~(Dudgeon et al. 2006, Vörösmarty et al. 2010, McNellie et al. 2020). The concept of "least disturbed" describes conditions that are derived from a collection of sites that represent the best available, given today's landscape ~~(Stoddard et al. 2006)~~(Stoddard et al. 2006). Unfortunately, because anthropogenic disturbance is not evenly distributed across the landscape, sites in least disturbed condition can be spatially ~~aggregated~~ biased and/or vary in quality ~~among ecoregions (Herlihy et al. 2008, McNellie et al. 2020)~~(Herlihy et al. 2008, McNellie et al. 2020, Yuan et al. 2024)~~. This makes it challenging to elucidate the effects of anthropogenic disturbance on biological assemblages because the~~

~~condition represented by a suite of~~. If least disturbed ~~stites~~sites are spatially biased, their ecological, chemical, and physical conditions may be unattainable for some locations that need to be assessed because of differences in natural setting or ~~be~~ biogeographic history. Alternatively, if reference site quality is lower for ~~ecoregions that have had~~some ecosystems because of a relatively long legacy of human intensification ~~(McNellie et al. 2020).~~, some sites will be judged against a lower benchmark and erroneously appear to be in better condition relative to others. Although least disturbed reference sites provides reasonable and defensible benchmarks for assessing biological condition (Stoddard et al. 2006, Herlihy et al. 2008, Mitchell et al. 2025), they are typically not undisturbed, nor evenly distributed among all ecosystem types (McNellie et al. 2020, Yuan et al. 2024). Novel approaches and concepts are urgently needed to overcome these limitations.

Quantifying taxon-environment relationships for all taxa within a region with empirical models may provide an alternative to relying on reference sites because they elucidate how individual taxa and entire assemblages change along physiochemical gradients (Chessman and Royal 2004, Kilgour and Stanfield 2005, Elias et al. 2016, Kopp et al. 2023). Because anthropogenic disturbance typically manifests as altered physiochemical conditions (Tang et al. 2020, Fergus et al. 2023), quantifying potential environmental divers across space, could be relevant for understanding how organisms may change with disturbance across time (Blüthgen et al. 2022). For example, taxa that have a negative relationship along an environmental gradient could be expected to become less prevalent if human disturbance contributes to more extreme environmental conditions and more prevalent if reducing human disturbance contributes to less extreme conditions (Kopp et al. 2023). From these relationships it is possible to predict which taxa from the regional species pool could occur at a site if it were minimally disturbed and compare them to the taxa that were actually observed (Chessman and Royal 2004, Elias et al. 2016). If the minimally disturbed assemblage is similar to the observed assemblage, then the site may be considered unaffected by anthropogenic disturbances.

Advances in species distribution modeling enhance our ability to quantify taxon-environment relationships (Franklin 2010, Guisan et al. 2017, Ovaskainen and Abrego 2020). Species distribution modeling consists of a diverse set of methods that typically focus on quantifying taxon-environment relationships for individual species (Elith and Leathwick 2009). Joint species distribution modeling (JSDM) is a recent multivariate extension of single species distribution modeling, capable of estimating taxon-environment relationships for all members of an assemblage simultaneously. Because of their complex structure, JSDMs are often fitted using Bayesian inference. As such, the resulting posterior distribution can be used for hypothesis testing (Johnson et al. 2022). For example, given a posterior distribution of values representing a biological assemblage (e.g. taxonomic richness) in the absence of human disturbance, it is possible to evaluate whether an observed assemblage is consistent with this distribution. In addition, these models use latent factors to account for unmeasured environmental variables and potential biotic interactions (Warton et al. 2015, Ovaskainen et al. 2017, Ovaskainen and Abrego 2020). This feature is particularly important for biological assessments because it can strengthen the inference gained from evaluating a specific suite of environmental variables that are typically altered by human disturbance. Thus, the ability to model all taxa in an assemblage and probabilistically evaluate whether an observation is consistent with a posterior distribution makes JSDMs well suited for understanding how anthropogenic factors affect biological assemblages.

Although taxon-environment relationships are necessary to circumvent the need for reference sites, they require a concurrent expectation of undisturbed physiochemical conditions to fully understand the effects of anthropogenic activities. Traditionally, undisturbed physiochemical conditions are obtained, or modeled, from least-disturbed reference sites (Olson and Hawkins 2012, Olson and Cormier 2019). However, empirical models have also linked changes in the physiochemical environment to anthropogenic activities and been used to estimate conditions if anthropogenic disturbance is reduced or eliminated without reference sites (Dodds and Oakes 2004, Herlihy and Sifneos 2008, Soranno et al. 2011). Random forest models are a machine learning algorithm that have superior predictive performance compared to other modeling techniques (Prasad et al. 2006, Peters et al. 2007) and are often used to model water quality parameters and physical habitat characteristics (Olson and Hawkins 2013, Olson and Cormier 2019, Sabo et al. 2023, Yuan et al. 2024). Further these models are more robust to nonlinearities, multicollinearity, and overfitting (Breiman 2001) and can evaluate the relative importance of different predictor variables (Lin et al. 2021, Sabo et al. 2023). Because these algorithms can accommodate complex relationships and a relatively high number of covariates, they are well suited for evaluating the relative importance of natural and anthropogenic variables on physiochemical conditions and, in turn, predicting the abiotic conditions if anthropogenic disturbances were removed (Yuan et al. 2024). Indeed, efforts that combined estimates of undisturbed physiochemical conditions with biological models have made progress towards addressing the limitations of reference site-based approaches (Chessman and Royal 2004, Kilgour and Stanfield 2005, Elias et al. 2016, Yuan et al. 2024).

In rivers and streams, ~~preserving and maintaining biological integrity is a central management objective (Jackson and Davis 1994, Paulsen et al. 2008, Hering et al. 2010).~~

~~Benthic macroinvertebrate assemblages are often surveyed as part of assessing biological integrity~~assessments of biological integrity often rely on surveys of benthic macroinvertebrate assemblages because they are diverse, relatively easy to sample, and respond to changes in the physiochemical environment associated with anthropogenic disturbance ~~(Hughes and Peck 2008, Buss et al. 2014)~~(Hughes and Peck 2008, Buss et al. 2014). To elucidate the effects of anthropogenic activities on biotic integrity, benthic macroinvertebrate assemblages observed at a site are typically compared to reference assemblages collected from least disturbed sites ~~(Hawkins 2006, Stoddard et al. 2008)~~(Hawkins 2006, Stoddard et al. 2008)~~. Reference assemblages should be obtained from sites that are minimally disturbed (Reynoldson et al. 1997) but, given the pervasiveness of anthropogenic disturbance and its covariance with geoclimatic predictors, this is often unattainable. Instead, macroinvertebrate assemblages are typically collected from the best available sites such that benchmarks typically represent a least-disturbed condition (Yuan 2006, Herlihy et al. 2008, Chen et al. 2019). Although least-disturbed benchmarks provide a reasonable and defensible management target, they are not undisturbed. Thus, they can mask the true extent of anthropogenic disturbance.~~

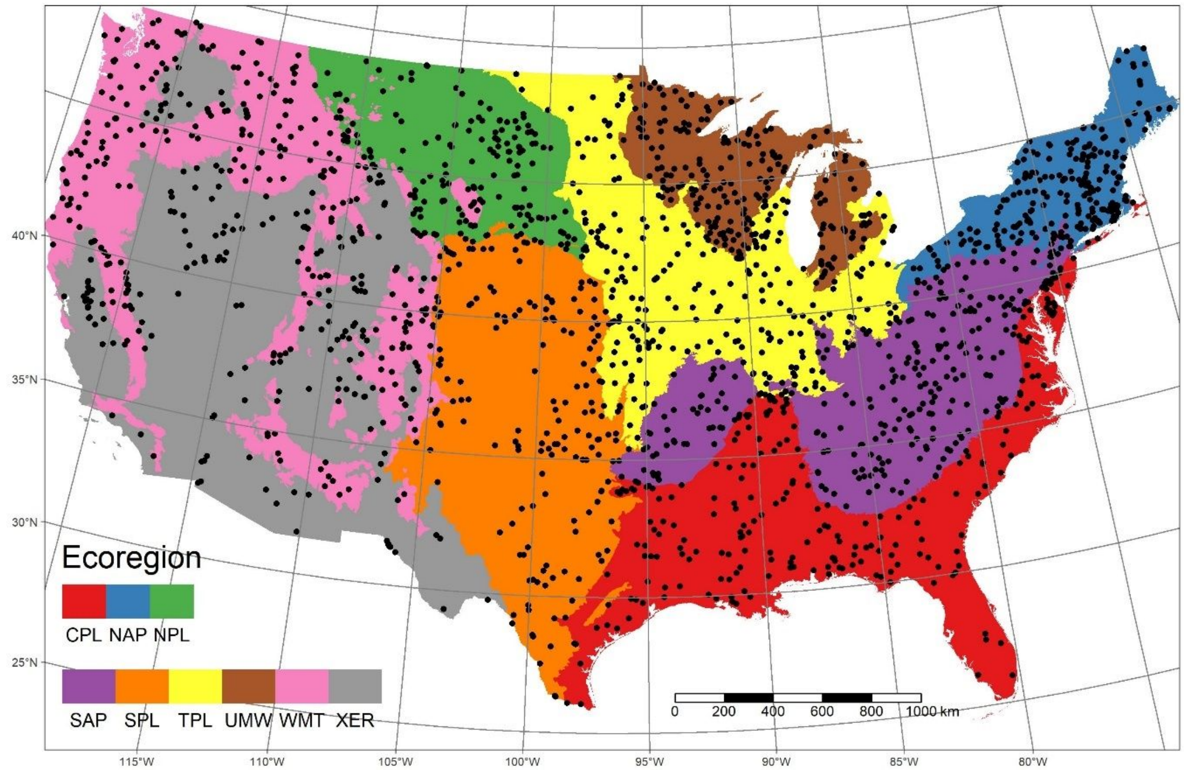
~~Aquatic organisms are directly affected by the physiochemical environment of the stream which, in turn, is determined by a combination of current and past geoclimatic and anthropogenic factors in the watershed (Tang et al. 2020, Fergus et al. 2023). Empirical models that link the physiochemical environment to anthropogenic disturbance in the watershed could be used to estimate physiochemical characteristics if anthropogenic disturbance is reduced (Dodds and Oakes 2004, Herlihy and Sifneos 2008, Soranno et al. 2011). These estimates can then be combined with taxon-environment relationships to elucidate how individual taxa and entire assemblages might respond to changing physiochemical conditions (Chessman and Royal 2004,~~

~~Kilgour and Stanfield 2005, Elias et al. 2016, Kopp et al. 2023). Because physiochemical characteristics are often used as management targets, this framework may offer a more direct link between management intervention and biotic responses compared to using reference sites (Hering et al. 2010, McNellie et al. 2020, Deflem et al. 2021).~~

~~Here, we use a two-stage modeling approach to assess how benthic macroinvertebrate assemblages at 1,824~~. Here, we develop a model-based assessment as a complementary method to evaluate how benthic macroinvertebrate assemblages at 1,748 stream sites may change if the effects of major anthropogenic disturbances were removed from the physiochemical environment. First, we use random forest models to relate in-stream physiochemical conditions to a suite of geoclimatic and anthropogenic factors and ~~used~~use these models to infer abiotic conditions if anthropogenic disturbance was eliminated. Second, we combined these estimates with ~~joint species distribution models (JSDMs)~~JSDMs to predict how biological assemblages could change if physiochemical conditions were not altered by anthropogenic disturbance. ~~We focus on~~Because JSDMs can leverage latent factors to account for unmeasured environmental factors, including potential biotic interactions (Ovaskainen et al. 2016), we focus exclusively on environmental variables that are commonly associated with anthropogenic disturbances. Specifically, we evaluate the effects of changing nutrient and salinity concentrations, physical habitat, and climate on benthic macroinvertebrate assemblages because these gradients are ~~known to be influenced by humans~~ commonly used as abiotic screens used to identify least disturbed reference sites ~~(Herlihy et al. 2008, Paulsen et al. 2008, USEPA 2023)~~(Herlihy et al. 2008, Paulsen et al. 2008, USEPA 2023, Mitchell et al. 2025) and affect benthic macroinvertebrate distributions ~~(Kopp et al. 2023)~~(Kopp et al. 2023).

Methods:

For our analysis, we focused on biological and environmental data collected at sites distributed among 9 ecoregions (Figure 1) and surveyed during the 2018/2019 National Rivers and Streams Assessment (NRSA). NRSA is a collaboration between the United States Environmental Protection Agency (USEPA) and state, tribal, and federal partners to assess the chemical, physical, and biological condition of streams and rivers of the United States (https://www.epa.gov/national-aquatic-resource-surveys/nrsa). Every five years, beginning in 2008/2009, NRSA ~~uses a statistical survey design to randomly select~~surveys ~2,000 stream/river locations that are ~~representative of the entire population of streams and rivers in the contiguous US (Olsen and Peck 2008).~~selected using a probabilistic survey design and handpicked by state and tribal partners. These sites are surveyed for biological, chemical, and physical characteristics using standardized field protocols (USEPA 2023). Only sites included in the probabilistic sample are selected to be representative of the entire population of streams and rivers in the contiguous US and used for assessment (Olsen and Peck 2008). In addition, approximately 10% of the probabilistic sites are revisited to assess within-year variability and approximately 30% of these sites are resampled in the next survey cycle to evaluate among-year variability.



~~Figure 1: Ecoregions and survey locations for the National Rivers and Streams Assessment 2018-2019 survey. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric~~

Benthic macroinvertebrate assemblages and physiochemical variables

Procedures for collecting benthic macroinvertebrates are described in detail elsewhere ~~(Hughes and Peck 2008, USEPA 2017a)~~(Hughes and Peck 2008, USEPA 2017a). Briefly, each survey location was defined as a length of stream or river equal to a multiple of its channel width. In wadeable and boatable sites~~<13 m wide~~, the reach length sampled was equal to 40 channel widths, ~~or~~ a minimum of 150 m~~. For wadeable sites >13 m wide, and boatable sites, the minimum reach length sampled was the longer~~, or a maximum of ~~500 m or 20 channel widths~~4km. Macroinvertebrate samples were collected along 11 equally spaced cross-section transects along the reach. In wadeable sites, samples were collected in an alternating left, center, right order along the transects using a D-frame kick net (500-um mesh, 0.09 m^2 area). In boatable sites, samples were collected along the left or right wadeable margin from a 1m linear sweep of the dominant habitat using a D-frame kick net (500-um mesh, 0.3048 m^2 area). Samples from each survey location were combined into a single composite sample, preserved in ethanol, and sent to a taxonomist for subsampling and identification. ~~We used presence/absence data from the ~2,000 benthic macroinvertebrate collections and focused on taxa that were collected from a single site visit, identified to genus, and occurred at ≥10% of sites within an ecoregion (Table 1).~~

~~We~~To assess the effects of anthropogenic disturbance on in-stream physiochemical conditions and in turn macroinvertebrate assemblages, we focused on in situ nutrient concentrations, salinity, and physical habitat variables~~, because they~~. These environmental gradients are commonly altered by anthropogenic activities in the surrounding watershed and frequently used as abiotic screens to identify least-disturbed reference sites ~~(Herlihy et al. 2008, Paulsen et al. 2008, USEPA 2023)~~(Herlihy et al. 2008, Paulsen et al. 2008, USEPA 2023)~~ and influence macroinvertebrate distributions (Kopp et al. 2023). We selected total nitrogen (NTL) and total phosphorus (PTL) as indicators of nutrients; chloride (CL) and sulfate (SO4) as indicators of salinity; and mean substrate diameter (SUBD) and riparian disturbance index (RPDI) as indicators of physical habitat. A~~. Specifically, we selected total nitrogen (NTL) and total phosphorus (PTL) as indicators of excess nutrients supplied from agriculture and/or urbanization (Herlihy et al. 1998) and chloride (CL) and sulfate (SO4) as indicators of salinity. Elevated CL is a general indicator of human disturbance in the catchment (Herlihy et al. 1998) and SO4 can indicate sites that are affected by mine drainage (Herlihy et al. 1991). Mean substrate diameter (SUBD) and riparian disturbance index (RPDI) were selected as indicators of physical habitat because human activities can increase fine sediment inputs or directly modify the riparian area (Kaufmann 1999, Kaufmann et al. 2022b). In addition, we included mean summer air temperature (Mean July and August temperature, MSAT) and total annual precipitation (TPRCP) to assess potential changes associated with differences from a 1900-1950 baseline. Although other factors are affected by human disturbances (e.g., metals and pesticides and hydrologic alteration), these variables provide a reasonably broad characterization of the physiochemical environment at a site (Herlihy et al. 2008, Paulsen et al. 2008)

During the survey, a single water sample was collected at each site and shipped to a central analytical laboratory. PTL and NTL were measured by persulfate digestion and colorimetry and CL and SO4 were measured by ion chromatography ~~(USEPA 2017b)~~(USEPA 2017b). SUBD is the geometric mean of the numeric value assigned to substrate size classes measured in the field ~~(Kaufmann 1999).~~(Kaufmann 1999). RPDI summarizes the presence/absence of 11 categories of human disturbance, including buildings, landfill/ trash, logging, mining, developed parks or lawns, pavement or cleared lots, pipes (withdrawal or wastewater), roads, row crops, pastures or hayfields, and walls or revetments in the riparian area, adjacent to each to the 11 transects where macroinvertebrates were collected ~~(Kaufmann 1999, USEPA 2017a, Kaufmann et al. 2022)~~(Kaufmann 1999, USEPA 2017a, Kaufmann et al. 2022b). In addition, we included mean summer air temperature (Mean July and August temperature, MSAT) and total annual precipitation (TPRCP) for either 2018 or 2019 (depending on sample year) obtained from the PRISM ~~dataset~~Climate Group (https://prism.oregonstate.edu/).

~~Table 1: Sites, genera, and environmental variables included in the analysis. Values for the environmental variables are median values for each region. Range is given in parentheses. NTL = Total Nitrogen, PTL = Total Phosphorous, CL = Chloride, SO4 = Sulfate, RPDI = Riparian Disturbance Index, SUBD = Substrate Diameter, TPRCP = Total Precipitation and MSAT = Mean Summer Air Temperature. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric.~~

Region	Site (#)	Genera (#)	NTL (ug/L)	PTL (ug/L)	CL (mg/L)	SO4 (mg/L)	RPDI	SUBD (mm)	TPRCP (mm)	MSAT (°C)
CPL	226	71	786 (91-7713)	84.81 (5.2-3922.25)	10.06 (0.52-4797.9)	7.83 (0.11-1900.89)	0.5 (0-5.39)	0.35 (0.01-341.65)	1613.83 (388.66-2397.1)	27.46 (23.75-31.68)
NAP	228	127	431 (81-6413)	23.76 (3.12-587.2)	16.73 (0.07-668.16)	5.58 (0.06-467.93)	0.65 (0-5.24)	11.66 (0.01-864.9)	1295.41 (888.91-1883.17)	21.49 (17.24-24.72)
NPL	152	76	881.5 (71-15675)	85.61 (3.53-9248.31)	12.09 (0.11-1251.58)	486.24 (3.5-4079.78)	1.27 (0-4.49)	0.51 (0.01-560.5)	437.42 (157.84-1055.11)	20.52 (15.62-23.46)
SAP	266	111	557.5 (36-18700)	29.3 (3.79-4050)	5.89 (0.36-197.39)	7.46 (0.58-397.69)	0.76 (0-4.56)	16.28 (0.01-5656.85)	1475.78 (889.64-2451.81)	24.76 (19.19-27.73)
SPL	174	59	1165 (145-21175)	147.12 (5.21-4351.7)	30.36 (0.43-5220)	88.89 (1.96-3716.6)	1.04 (0-5.88)	0.35 (0.01-5656.85)	616.17 (240.1-1404.61)	25.88 (11.96-32.75)
TPL	223	74	1806 (236-16219)	165.08 (11.45-1066.87)	18.9 (1.44-736.62)	34.4 (5.79-1386.77)	0.83 (0-5.47)	0.35 (0.01-5656.85)	1004.22 (376.7-1801.08)	23.33 (18.28-27.09)
UMW	201	104	1168 (195-17675)	57.4 (8.16-659)	9.97 (0.01-306.69)	9.28 (0.04-160.4)	0.62 (0-6.43)	0.35 (0.01-812.79)	960.16 (496.44-1718.2)	20.84 (17.05-23.43)
WMT	225	94	133 (22-4719)	21.84 (2.71-569.8)	0.92 (0.04-521.64)	2.97 (0.07-1682.05)	0.48 (0-3.95)	38.14 (0.01-1288.61)	669.31 (175.89-3946.33)	18.08 (11.71-29.4)
XER	196	75	344.5 (46-8000)	54.62 (4.29-4667.41)	5.35 (0.1-1867.57)	27.8 (0.02-3286.24)	1.16 (0-4.36)	2.02 (0.01-368.11)	287.58 (100.46-1176.56)	22.76 (14.69-36.11)

Modeling ~~taxon~~genus-environment relationships

We quantified relationships between macroinvertebrate assemblages and eight physiochemical variables (i.e. NTL, PTL, CL, SO4, SUBD, RPDI, MSAT, and TPRCP) using joint species distribution models (JSDMs) fitted with the Hierarchical Modeling of Species Communities R package ~~(Ovaskainen and Abrego 2020, Tikhonov et al. 2020)~~(Ovaskainen and Abrego 2020, Tikhonov et al. 2020)~~. JSDMs~~. JSDMs are a multivariate hierarchal generalized linear mixed model fitted with Bayesian inference. They are uniquely suited to evaluate relationships between anthropogenic disturbance and biological assemblages because they are multi-species models that quantify taxon-environment relationships for all members in an assemblage simultaneously and account for unmeasured variables, including abiotic factors and species associations, using random effects specified at the sample-level ~~(Warton et al. 2015, Ovaskainen and Abrego 2020, Deflem et al. 2021)~~(Warton et al. 2015, Ovaskainen and Abrego 2020, Deflem et al. 2021)~~. They also allow for the inclusion of phylogenetic relatedness and species traits as hierarchical terms that can improve estimates of taxon-environment relationships.~~ . Indeed, sample-level random effects are meaningful for multivariate models because they are not confounded by residual variation as with univariate models. More specifically, this attribute allows models to account for nonindependence among residuals for each site and improves estimates of the fixed effects (i.e. taxon-environment relationships) (Ovaskainen et al. 2016). In addition, JSDMs also allow for the inclusion of phylogenetic relatedness as a hierarchical term that can lend additional improvements to estimated taxon-environment relationships.

We provide a detailed description of the modeling framework and application elsewhere ~~(Kopp et al. 2023)~~(Kopp et al. 2023)~~. In brief, the JSDM is a multivariate hierarchal generalized~~

~~linear mixed model fitted with Bayesian inference. We fit separate models for each ecoregion using genus occurrence (i.e., presence/absence) as the response variable, environmental variables as fixed effects, and site-level random effects. We also use taxonomy as a surrogate for phylogenetic relatedness but excluded traits to avoid complications with incomplete data and increase the number of genera included in our analysis.~~. In brief, we used presence/absence data from 1,891 benthic macroinvertebrate assemblages surveyed as part of the probabilistic and handpicked sites and focused on taxa that were collected from a single site visit, identified to genus, and occurred at ≥10% of sites within an ecoregion (Table 1). Separate models were fit for each ecoregion boundaries to define regional species pools and thus assume that environmental conditions are the primary factor driving genus occurrence (Chessman and Royal 2004). All physiochemical variables were measured in the field during the survey, except for MSAT and TPRCP, which were obtained from PRISM Climate and matched to the appropriate survey year (Table 1). Physiochemical variables were used as linear fixed effects and sample-level random effects were used to statistically control for unmeasured variables. Previously, we found few genus-environment relationships were unimodal (Kopp et al. 2023) and therefore assumed linear relationships were appropriate for this study. Our analysis focused on relatively small number of environmental variables because these are commonly altered by human activities. We also use taxonomy as a surrogate for phylogenetic relatedness as a hierarchical term in the model. All models were fitted with the default prior distributions ~~(Ovaskainen and Abrego 2020)~~(Ovaskainen and Abrego 2020), using three independent chains (~~3000~~3,000 posterior samples). Convergence was determined to be satisfactory by potential scale reduction factor < 1.1 for fixed effects and phylogenic parameters.

~~Model performance was evaluated with respect to the fitted models ability to reproduce the observed genus richness and composition (Wilkinson et al. 2021). Predicted, taxon-specific occurrence probabilities were summed to obtain predicted richness and regressed against the observed richness~~ Our primary motivation for using these models was to measure how macroinvertebrate assemblages may change if the influence of anthropogenic disturbances were removed from in stream physiochemical variables and all else remained unchanged (Figure 2). Since we do not use these models to predict to new locations, model performance was primarily evaluated with respect to its explanatory power, i.e. the fitted models ability to reproduce the observed genus richness and composition (Wilkinson et al. 2021, Abrego and Ovaskainen 2023). Predicted, taxon-specific occurrence probabilities were summed to obtain predicted richness and regressed against the observed richness ~~(Calabrese et al. 2014)~~(Calabrese et al. 2014). We determined model acceptability based on three criteria: $R^2 \geq 0.2$, $-1.5 \leq$ intercept ≤ 1.5, and $0.85 \leq$ slope ≤ 1.15 ~~(Linke et al. 2005)~~(Linke et al. 2005). ~~Compositional~~We calculated model performance metrics using each of 3,000 posterior samples to obtain a distribution of plausible estimates of performance metrics and report the mean and 5th and 95th quantiles as measures of uncertainty. In addition, we assessed compositional similarity between predicted and observed assemblages ~~was assessed~~ using a probabilistic adaptation of Jaccard similarity ~~(Scherrer et al. 2020)~~(Scherrer et al. 2020) to avoid introducing error associated with converting predicted occurrence probabilities into binary outcomes ~~(Calabrese et al. 2014)~~(Calabrese et al. 2014). -We calculated ~~model performance metrics~~ similarity for each site using ~~each of 3,000 posterior samples to obtain a distribution of plausible estimates for the performance metrics~~the mean predicted occurrence probabilities and report the mean and 5th and 95th quantiles ~~as measures~~across all sites.

Predictive power for models that use random effects can only be assessed for cases where at least some sampling units were included in the calibration data (Abrego and Ovaskainen 2023). As part of NRSA, a subsample of sites are revisited with the intent of assessing temporal variability in metrics and indices (Table 1). Since we fit the JSDMs using sample-level random effects, these data were used to as a second type of validation. Specifically, we evaluated whether the richness observed at a revisited site was within the posterior distribution of the fitted models. Although this metric of validation may seem less restrictive compared to those used to evaluate explanatory power, we expected variation between samples to be rather large because of stochastic events (e.g. high or low stream flows), ecological processes (e.g. emergence and dispersal), and sampling procedures (e.g. field collection and laboratory subsampling) that our models were not designed to capture. Given this level of potential uncertainty~~.~~, this metric was intended to address whether an observation is could have come from the same process that our model was intended to capture. We also compared the assemblages collected during the revisit to the predicted assemblages using the probabilistic adaptation of Jaccard similarity (Scherrer et al. 2020) and report the mean and 5th and 95th quantiles across sites.

Modeling ~~Physiochemical Gradients~~physiochemical gradients

We used random forest models to relate anthropogenic and landscape geoclimatic factors to total nitrogen (NTL), total phosphorous (PTL), chloride (CL), sulfate (SO4), and substrate diameter (SUBD) and make predictions if anthropogenic disturbances were removed. The suite of predictor variables were selected based on their hypothesized relationship with stream water chemistry and bed particle size ~~(Lin et al. 2021, Zak et al. 2021, Kaufmann et al. 2022, Kaushal et al. 2023, Sabo et al. 2023)~~(Lin et al. 2021, Zak et al. 2021, Kaufmann et al. 2022b, Kaushal et al. 2023, Sabo et al. 2023)~~ and bed particle size.~~. Geoclimatic factors included watershed morphology (e.g., basin area, elevation, and slope) and lithological characteristics (e.g., lithological phosphate, nitrogen, ~~or~~and sulfur, and soil erodibility). Anthropogenic factors included percent agriculture, road density, and presence of mines and impoundments (Appendix S1: Table S1). ~~Most predictor~~Predictor variables were obtained or derived from the StreamCat Database ~~(Hill et al. 2016)~~(Hill et al. 2016), National Atmospheric Deposition Program (nadp.slh.wisc.edu), and EPA's National Nutrient Inventory (Sabo et al. 2019, Lin et al. 2021, Sabo et al. 2021, Sabo et al. 2023).

~~For each gradient, the~~Separate random forest models were fit for NTL, PTL, CL, SO4, or SUBD using the entire national dataset to maximize the range of the predictor variables included in each model. Prior to model fitting the dataset was randomly split into training and testing portions (~~75/25,~~80% and 20%, respectively~~) and separate random forest models were fit using the complete list of candidate variables.~~). Model fit was assessed by ~~out-~~the coefficient of~~-bag~~ determination (R^2~~,~~) and root mean squared error (RMSE) on training and testing portions of the dataset. The relative importance of the covariates was evaluated by the change in mean squared error after permutating each variable (%IncMSE). ~~The presence of strong correlations between geoclimatic and anthropogenic predictor variables could generate misleading~~ Because variable importance ~~rankings because a~~ can be influenced by correlated ~~variable may become a substitute when the other is permutated (i.e., no change in %IncMSE). We did not detect strong correlations (i.e., r > 0.7) between the anthropogenic~~variables, we confirmed that all variables used in the models had pairwise Pearson's correlation coefficients < 0.7 and the variance inflation factors (VIF) were between 2.6 and ~~geoclimatic variables. We then used partial~~3.9. (Appendix S1: Table S3) In general, VIF > 5 indicates a potential problem with multicollinearity

(O'brien 2007). Partial dependance plots were used to visualize the relationship between the most important anthropogenic factors and each ~~environmental gradient.~~physiochemical variable. Random forest analysis was performed using the quantregForest R Package ~~(Meinshausen 2017)~~(Meinshausen 2017).

Evaluating effects of human disturbance on physiochemical gradients

We used the fitted random forest models to predict values for NTL, PTL, CL, SO4, and SUBD if anthropogenic disturbance was removed by setting all anthropogenic factors to the lowest value observed in our dataset (often zero, Appendix S1: Table S2) and leaving geoclimatic factors unchanged. Hereafter "hindcast" refers to the removal of anthropogenic disturbance from the physiochemical variables. ~~Because modeled relationships have unexplained residual variation, we expressed the differences between observed and hindcast values in units of standard deviations (i.e., SD = (Observed – Hindcast) / RMSE) to quantify the effects of anthropogenic disturbance relative to unexplained variation~~Although we included the entire range of predictor values, it is possible that hindcast data are not sufficiently similar to the data used to train the model and thus susceptible to extrapolation (Meyer and Pebesma 2021, Yuan et al. 2024). To test whether the hindcast dataset was sufficiently similar to the data used to train the models, we first mean-centered and scaled all predictor variables to equivalent units (i.e. standard deviations) and weighted them according to their importance in the model. We then calculated the minimum Euclidean distance between each site in the hindcasted data, and each site used in the training dataset using the same center and scale. The minimum value was then divided by the mean Euclidean distance among all training data. Following Yuan et al. (2024), we flagged hindcast predictions for sites that exceeded 0.5 as susceptible to extrapolation.

We evaluated the effects of anthropogenic disturbance on each physiochemical variable based on deviations from the hindcast estimates using standardized anomalies (i.e. z-scores), or the difference between the observed and hindcast value scaled by standard deviation. For variables we modeled using random forests, we standardized the difference between observed and hindcasted conditions using the RMSE of each model ~~(Kilgour and Stanfield 2005)~~(Kilgour and Stanfield 2005)~~. We expected anthropogenic disturbance to elevate NTL, PTL, CL, and SO4 concentrations and increase or decrease SUBD. Accordingly, we identified sites with observed concentrations greater than 2SD from the hindcast values for NTL, PTL, CL, and SO4 and either 2SD higher or lower for SUBD as disturbed.~~ . For climate variables, we obtained baseline historical climatic values (MSAT and TPRCP) as the mean summer air temperature for 1900-1950 from PRISM climate data (https://prism.oregonstate.edu/historical/). We then standardized the difference between the present-day values (i.e. 2018 or 2019) and the baseline using the standard deviation of the 50yr dataset. Importantly, using standardized anomalies rather than absolute differences has advantages because it accounts for unexplained variation in the model or natural variability among sites (Kilgour and Stanfield 2005). Furthermore, because the values are in units of standard deviations, thresholds to evaluate whether the magnitude of difference is sufficient to support an effect of anthropogenic disturbance can be rather intuitive. Specifically, we expected anthropogenic disturbance to elevate NTL, PTL, CL, and SO4 concentrations and increase or decrease SUBD, MSAT, and TPRCP. Accordingly, we identified sites with observed concentrations >2SD from the hindcast values for NTL, PTL, CL, and SO4 and > |2SD| for SUBD, MSAT, and TPRCP as having evidence of anthropogenic disturbance.

~~We also evaluated the effects of anthropogenic disturbance on climate variables and the riparian disturbance index. We obtained historical climatic values (MSAT and TPRCP) from~~

~~PRISM 1940-1950 climate data (https://prism.oregonstate.edu/historical/). We selected 1940-1950 climate because it predates the peak in atmospheric radiocarbon from nuclear fallout – an apparent geologic marker of the Anthropocene (Lewis and Maslin 2015). For each location we calculated the 10 yr mean and standard deviation and identified sites that could be disturbed when differences between present-day and historical values exceeded 2SD (either higher or lower).~~ Since RPDI is a direct measure of anthropogenic disturbance, we initially set this variable to zero for all locations but found ~~nearly all the sites have RPDI > 0.~~that this value may be too strict because of the number and diversity of factors included in the index and did not provide much insight into potential regional variation in disturbance. Instead, we used 0.33 as a threshold which ~~means that on average only~~ is interpreted as one type of human disturbance ~~should be~~ observed within 10 m of the streambanks at no ~~mor~~more than one third of the 22 riparian plots sampled, on average ~~(Kaufmann et al. 2022, USEPA 2023)~~(Kaufmann et al. 2022b, USEPA 2023). ~~For each environmental gradient we calculated the proportion of sites in each region that~~ Although this does mean that not all sites were completely free of riparian disturbance, it is consistent with other studies that have used this index to evaluate human disturbance in the riparian area (Kaufmann et al. 2022a, Kaufmann et al. 2022b, USEPA 2023).

For each physiochemical variable we estimated the total percent of streams that were potentially affected by anthropogenic disturbance~~.~~ (i.e. standardized anomaly > 2SD) using only the probabilistic samples and their weights reflective of the entire population of streams and rivers assigned to them by NRSA (USEPA 2023). Estimates for each ecoregion were generated using the cat_analysis() function from the spsurvey R package (Dumelle et al. 2023). In addition, we tallied the number of physiochemical variables that were potentially affected at each site to elucidate instances where human disturbance affects multiple environmental variables simultaneously.

Evaluating effects of human disturbance on macroinvertebrate assemblages

We evaluated the effects of removing anthropogenic disturbance from the physiochemical environment on macroinvertebrate assemblages using the fitted JSDMs. For the sites ~~with disturbed environmental variables,~~that had evidence of human disturbance (i.e. standardized anomaly > 2SD), we ~~substitued~~substituted the hindcasted value (either predicted from random forest model, ~~1940~~1900-1950 ~~climate average~~averages for MSAT and TPRCP or 0.33 for RPDI) for the present day value in the dataset and used these data to predict macroinvertebrate assemblages that could occur if disturbance was ~~absent~~removed or reduced ~~(Chessman and Royal 2004)~~(Chessman and Royal 2004)~~. We then compared these hindcasted assemblages to assemblages predicted under present-day conditions.~~. To evaluate the relative ~~and additive~~ effects of ~~removing anthropogenic disturbance from the~~hindcasting each physiochemical ~~environment~~variable, we created ~~5~~4 scenarios changing either NTL and PTL (Nutrient Scenario), CL and SO4 (Salinity Scenario), RPDI and SUBD (Habitat Scenario), MSAT and TPRCP (Climate Scenario) and leaving the others at their observed ~~value or~~values. In addition, we predicted the macroinvertebrate assemblage after changing all physiochemical variables to their hindcast value (~~Pristine~~Hindcast Scenario). For each scenario we compared predicted genus richness from present-day conditions to hindcast genus richness.

JSDM predictions are three-dimensional data arrays that contain 3,000 posterior samples of site-specific occurrence probabilities for each genus (i.e., site x genera x posterior samples). For each posterior sample we summed predicted occurrence probabilities for all ~~taxa~~genera to obtain 3,000 plausible estimates of hindcast genus richness for each site. We then compared the mean present-day genus richness (i.e., predicted from JSDM using present-day values for the physiochemical variables) to the hindcast posterior distribution~~ to evaluate~~. We evaluated whether present-day richness was below the 10th or 25th quantiles, indicating a reduction in present-day genus richness relative to hindcast predictions, or above the 75th or 90th~~, quantiles (i.e. hereafter termed~~ quantiles, indicating an increase in genus richness relative to hindcasted (Figure 2). Hereafter, present-day values that are located in the extreme ends of the distribution are described as having either >0.75 ~~support~~ or >0.90 support for a difference from hindcasted predictions, respectively~~ because either 75% or 90% of the hindcast posterior distribution is greater or less than the present-day mean richness).~~. We assessed the effects of removing anthropogenic disturbance from each of the 4 combinations of environmental variables (i.e. Nutrient Scenario, Salinity Scenario, Habitat Scenario and Climate Scenario) and the effects of removing anthropogenic disturbance from all environmental variables simultaneously (Pristine Scenario) with >0.75 and >0.90 support ~~for differences in genus richness~~as evidence for an effect of disturbance on physiochemical variables.

~~Because assemblage composition could be affected by anthropogenic disturbance without an subsequent change in richness (Van Sickle 2008), we also assessed compositional differences between present-day and hindcast assemblages. First, we determined whether the present-day mean occurrence probability of a taxon was higher or lower (>90% support) than the hindcast occurrence probability for each site. Taxa that had higher occurrence probability were increasers while taxa with lower occurrence probability were decreasers. Assessing compositional dissimilarity between two assemblages requires binary values to indicate whether a taxon is present or absent in each assemblage. We used the increaser/decreaser assignments to create two~~

~~community matrices, one representing present-day assemblage and the other representing the hindcast assemblage. Increasers were assigned a value of 1 in the present-day matrix and 0 in the hindcast matrix. Conversely, decreasers were assigned 0 in the present-day matrix and 1 in the hindcast matrix. Taxa with < 0.90 support (i.e. mean present-day occurrence probability was within the 10th and 90th quantile of the hindcast distribution) were considered not to be affected by anthropogenic disturbance and were assigned 1 in both matrices. We used Jaccard similarity index to compare the two matrices (i.e. site x taxa) and identified compositional differences by Jaccard similarity < 0.9.~~

~~Identifying increaser and decreaser taxa can also elucidate differences among major taxonomic groups in their response to anthropogenic disturbance and help identify which major taxonomic groups are contributing most to assemblage level patterns. Accordingly, for each genus we calculated the proportion of sites in which they either~~Identifying genera having site-specific occurrence probabilities that differ between present-day and hindcast conditions helps interpret assemblage-level changes in response to anthropogenic disturbance. We identified genera that had higher occurrence probability under present day conditions as “increasers” and genera with lower occurrence probability as “decreasers”. Differences were determined with >90% support. For each genus we calculated the proportion of sites where they increased or decreased. If a genus was identified as an increaser at a large proportion of sites it could ~~be considered tolerant~~become more prevalent in response to anthropogenic disturbance. Alternatively, if a genus was identified as a decreaser at a large proportion of sites, it could ~~be considered sensitive~~become less prevalent in response to ~~disturbance.~~human disturbances. For each ecoregion, we ~~calculated the proportion of sites wherein each genus increased or decreased and then calculated~~report the mean ~~proportion increasing and decreasing by~~for major taxonomic groups (~~i.~~e.g. insects and non-~~insects).~~insect genera) to further understand taxon-specific trends in the context of biodiversity loss (Jähnig et al. 2021, Rumschlag et al. 2023).

Identifying increaser and decreaser genera could also enhance our ability to measure the effects of removing anthropogenic disturbance because assemblage composition could be affected by anthropogenic disturbance without a subsequent change in richness (Van Sickle 2008). To assess compositional differences between present-day and hindcast assemblages we used increaser/decreaser assignments to create two community matrices, one representing present-day assemblage and the other representing hindcast assemblage. Increasers were assigned a value of 1 in the present-day matrix and 0 in the hindcast matrix because they had a higher occurrence probability under present-day conditions. Conversely, decreasers were assigned 0 in the present-day matrix and 1 in the hindcast matrix because they had a significantly higher occurrence probability under hindcasted conditions. For genera that did not have sufficient support (< 0.90) for changing occurrence probabilities were considered not to be affected by anthropogenic disturbance and were assigned 1 in both matrices. We compared the two matrices (i.e. site x genera) using Jaccard similarity index and identified sites with a similarity score of < 0.9 as changing compositionally.

We assessed the consequences of removing anthropogenic disturbance from in-stream physiochemical environment for macroinvertebrate assemblages by identifying sites within each region that had evidence for either a change in richness (>0.75 support) or composition (Jaccard Similarity < 0.9) from present-day conditions. Then, using only the probabilistic samples that were not flagged for extrapolation and had complete data (n = 1748) and their weights, estimated the total percent of streams within each ecoregion that were potentially affected by anthropogenic disturbance. Estimates for each ecoregion were generated using the cat_analysis()

function from the spsurvey R package (Dumelle et al. 2023). In addition, we report results with different levels of support (> 0.75 and > 0.90) for richness change and separate compositional change to convey uncertainty and methodological differences.

Results:

Random Forest Modeling

~~Results:~~

~~*Random Forest Modeling*~~

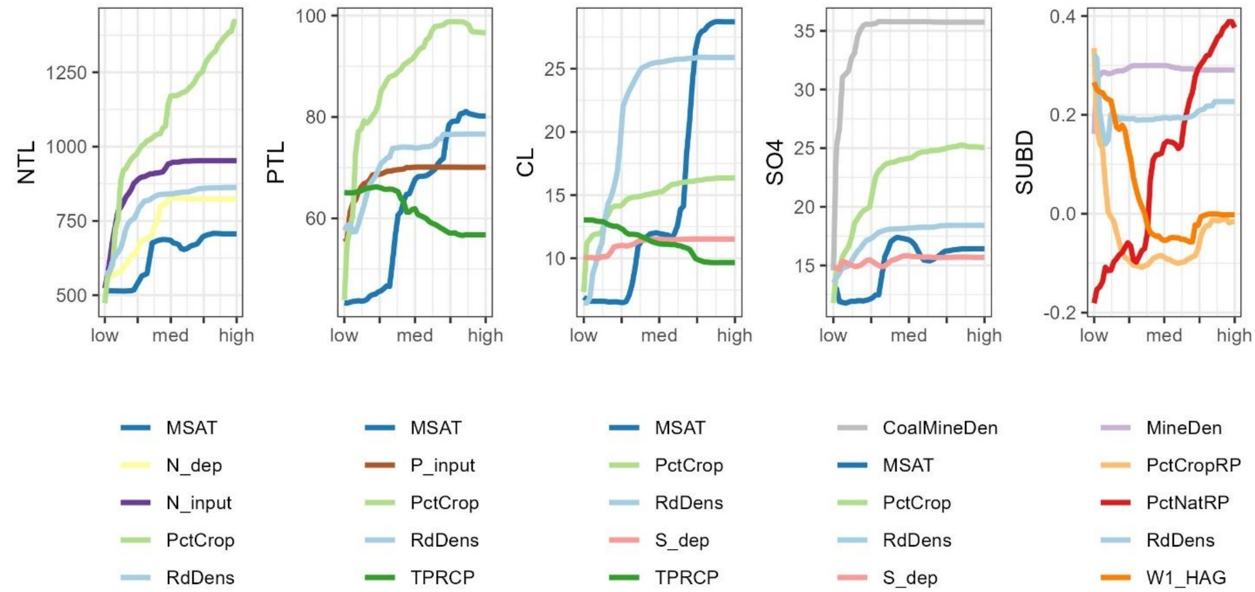
The random forest models explained ~~approximately 48~~46 to 77% of the variation in the training data (n=1502) and 51 to 78% of the variation in the test/validation data ~~for~~(n=375). The model of substrate diameter ~~and ≥59% of the variation in~~ (SUBD) had the highest RMSE in both training and testing datasets compared to models of total nitrogen~~,~~ (NTL), total phosphorous~~,~~ (PTL), chloride~~,~~ (CL), and sulfate (SO4) (Table 2). ~~The~~Variable importance for each physiochemical model, measured by the percent change in MSE after permutation, revealed that runoff was among the most important geoclimatic variables for each model and that ~~at least one anthropogenic factor~~ agricultural landcover in the watershed or riparian area ranked among the most important anthropogenic variables (Appendix S1: Table S2). ~~For example, percent~~Percent agricultural landcover in the watershed was the most important variable predicting NTL and PTL ~~and the~~, the second most important for predicting SO4 and third most important for predicting ~~SO4. Road~~CL. Other anthropogenic variables were also ranked relatively high in importance. For example, road density in the watershed was the second most important variable predicting

CL and fifth most important variable for SO4, while percent natural vegetation cover in the riparian area was the third most important variable predicting SUBD.

~~Table 2: Random Forest model performance metrics for testing and out-of-bag training datasets. NTL = Total Nitrogen, PTL = Total Phosphorus, CL = Chloride, SO4 = Sulfate, SUBD = Substrate Diameter. RMSE = Root mean squared error of random forest models fitted with ln(x + 1) (NTL, PTL, and CL) or ln(x) (SO4) or Log10(SUBD) transformations.~~

~~Variable~~	~~R^2_{train}~~	~~$RMSE_{train}$~~	~~R^2_{test}~~	~~$RMSE_{test}$~~
~~NTL (ug/L)~~	~~0.72~~	~~0.62~~	~~0.72~~	~~0.61~~
~~PTL (ug/L)~~	~~0.61~~	~~0.77~~	~~0.59~~	~~0.76~~
~~CL (mg/L)~~	~~0.70~~	~~0.77~~	~~0.71~~	~~0.82~~
~~SO4 (mg/L)~~	~~0.77~~	~~0.96~~	~~0.75~~	~~1.02~~
~~SUBD (mm)~~	~~0.47~~	~~0.98~~	~~0.48~~	~~1.00~~

We visualized the effects of the anthropogenic factors using partial dependance plots (Figure ~~2~~3). In general, the anthropogenic factors were associated with increased nutrients and salinity and decreased substrate diameter. For example, percent crops and anthropogenic nutrient inputs ~~(i.e.,~~of nitrogen and ~~phosphorus)~~phosphorous were positively associated with ~~total nitrogen~~NTL and ~~total phosphorous concentrations.~~PTL. Road density and mean summer air temperature had a positive association with ~~chloride~~CL while ~~density~~total annual precipitation had a weakly negative relationship. Density of coal mines in the watershed ~~was positively associated~~had a relatively strong positive effect on SO4 and SUBD had a strong positive relationship with ~~sulfate. Substrate diameter increased with percent~~ natural vegetation cover and strong negative relationship with agricultural activities in the riparian area ~~and decreased with percent agriculture and anthropogenic disturbance.~~.

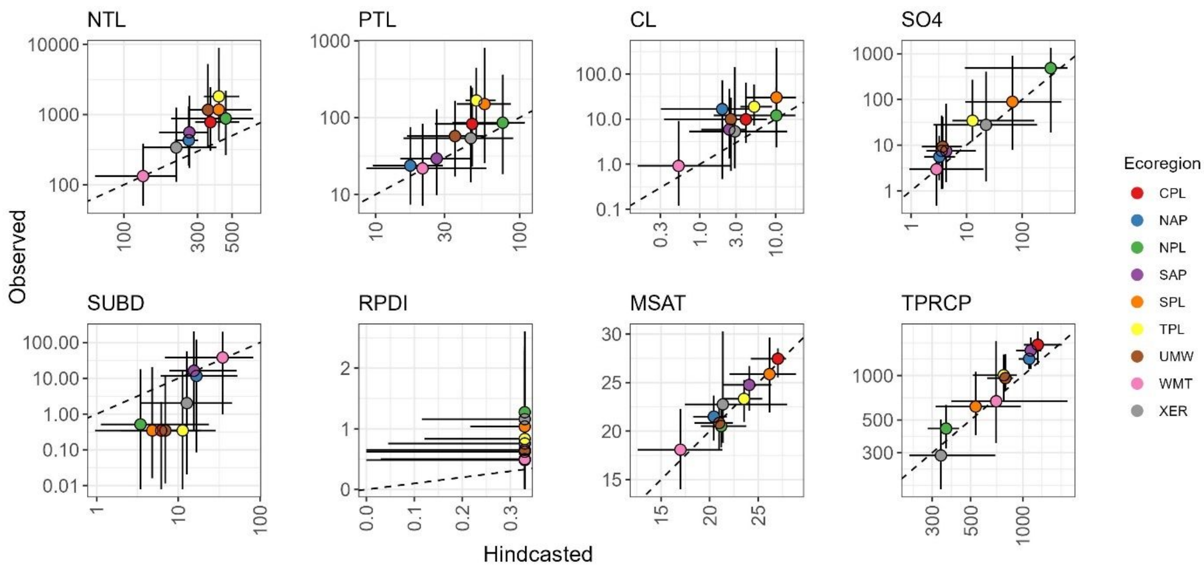


~~Figure 2: Partial dependence plots showing the effects of the anthropogenic variables on each environmental gradient. For visualization, each anthropogenic factor was rescaled between 0-1 and labeled low, medium, and high. CoalMineDen = coal mine density, MineDen = gravel mine density, MSAT = Mean summer air temperature, N_dep = atmospheric nitrogen deposition, N_input = Anthropogenic nitrogen inputs, P_input = Anthropogenic phosphorous inputs, PctCrop = Percent crop in the watershed, PctCropRP = percent crop in the riparian area, PctNatRP = percent natural vegetation in riparian area, RdDen = Road density, S_dep = atmospheric sulfur deposition, TPRCP = total precipitation, and W1_HAG = agricultural disturbance adjacent to stream reach. NTL = Total Nitrogen (ug/L), PTL = Total Phosphorus (ug/L), CL = Chloride (mg/L), SO4 = Sulfate (mg/L), SUBD = Substrate Diameter Log10(mm).~~

~~*Disturbance effects on physiochemical conditions*~~

~~Removing anthropogenic disturbance from the physiochemical environment generally decreased nutrient concentrations, salinity, and riparian disturbance and increased substrate diameter (Figure 3). Change in total nitrogen and phosphorus concentrations were most evident in Temperate Plains and Southern Plains and change in chloride was most evident in the Northern Appalachians. Substrate diameter was noticeably coarser after hindcasting for all regions except Xeric, Southern Appalachians, and Northern Appalachians. Observed and~~

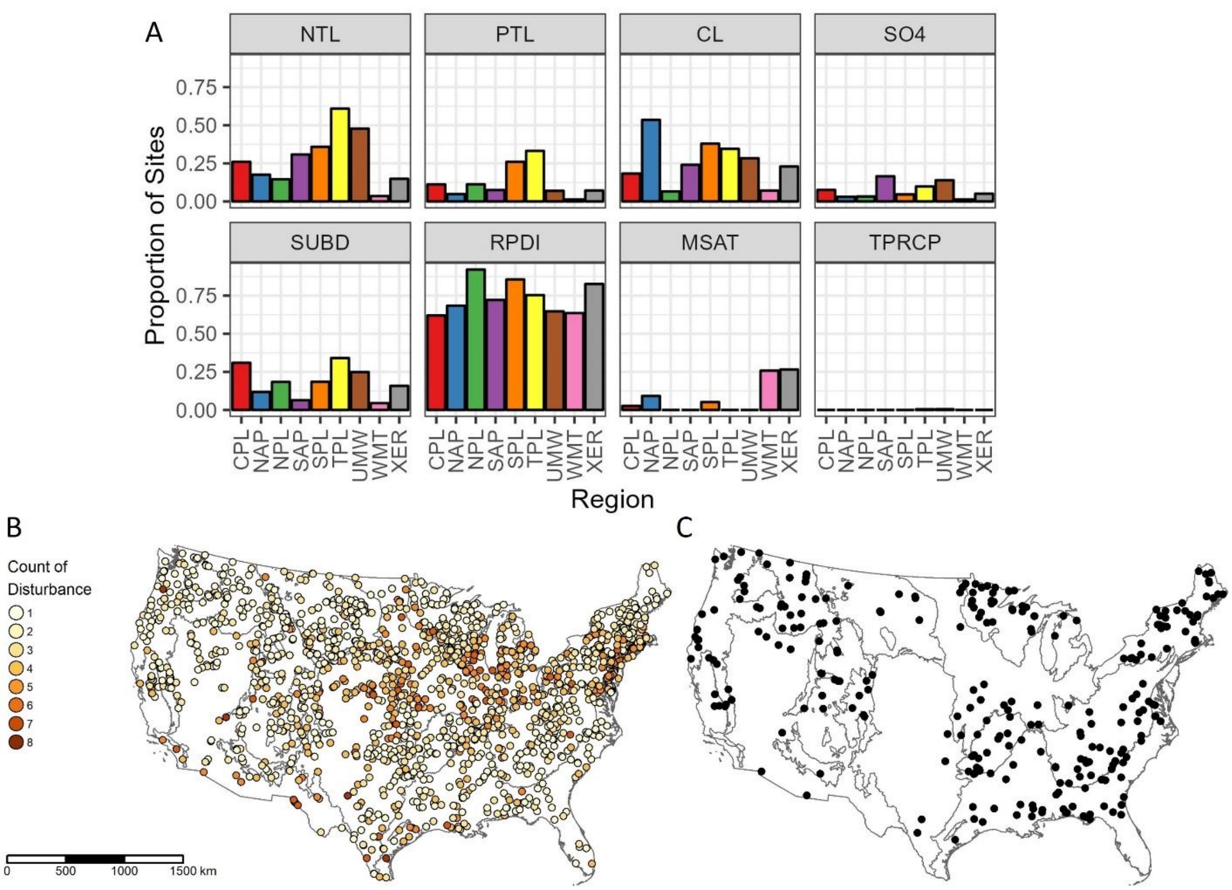
~~hindcasted values for sulfate, mean summer air temperature and total precipitation were generally similar for all ecoregions.~~



~~Figure 3: Ecoregional values of observed versus hindcasted values for each environmental gradient. Points are regional means and vertical and horizontal bars represent the 10^th^ and 90^th^ quantiles of observed values or hindcasted values within each region, respectively. The bars demarcate the typical range of values for each environmental variable within a region under each scenario. The dashed line is the 1:1 relationship, such that points above the line indicate that, on average, present-day values for the environmental variables within an ecoregion are greater than the average hindcasted value. NTL = Total Nitrogen (ug/L), PTL = Total Phosphorus (ug/L), CL = Chloride (mg/L), SO4 = Sulfate (mg/L), SUBD = Substrate Diameter (mm), RPDI = Riparian Disturbance Index, MSAT = mean summer air temperature (ºC) and TPRCP = total precipitation (mm). CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric. See Table S3 for observed and hindcasted mean values and quantiles for each region.~~

~~Anthropogenic disturbance affected physiochemical gradients differently depending on the ecoregion (Figure 4a). For example, the difference between present-day and hindcasted total nitrogen concentrations exceeded 2 standard deviations in 60% of the sites in Temperate Plains and 47% of sites in Upper Midwest. Similarly, 53% of the sites in the Northern Appalachians were found to have potentially elevated CL concentrations and >30% of sites in Coastal Plains and Temperate Plains had finer substrates. Most sites exceeded the 0.33 threshold for RPDI~~

~~based on the observed and hindcasted values while relatively fewer sites exceeded thresholds for sulfate concentrations, mean summer air temperature, or total precipitation. Among the ecoregions, Southern Appalachians had the largest proportion of sites with elevated sulfate concentrations relative to hindcasted values and several sites in Xeric and Western Mountains had historic MSAT higher than present-day. Anthropogenic disturbance could alter multiple physiochemical variables simultaneously (Figure 4b). For example, >50% of sites in Northern Appalachians, Southern Plains, Temperate Plains, and Upper Midwest had >2 physiochemical gradients change as a result of removing anthropogenic disturbance. Conversely, Western Mountains and Southern Appalachians had the greatest proportion of sites (>20%) without disturbance.~~



~~Figure 4: A) Proportion of sites where hindcasted abiotic conditions were >2 standard deviations from the observed value for Total Nitrogen (NTL), Total Phosphorus (PTL), Chloride (CL), Sulfate (SO4), Substrate Diameter (SUBD), mean summer air temperature (MSAT), and Total Precipitation (TPRCP) and > 0.33 for Riparian Disturbance Index (RPDI). B) The number of environmental variables where differences between hindcasted and observed abiotic conditions exceeded the threshold. C) Locations where differences between hindcasted and observed abiotic conditions did not exceeded the threshold for any variable. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric.~~

JSDM assemblage level performance ———

~~Depending on the ecoregion, we~~We fitted JSDMs for ~~63~~each region using presence/absence data for 59 – 127 genera ~~using 127 – 255~~from 152 – 266 sites surveyed during the 2018 – 2019 NRSA cycle (Table 1). The total number of sites used to fit JSDMs were higher (1,891) than the sites we ultimately assessed ~~(1,824)~~ because we included ~~all available information to maximize the number of occurrences for each genus~~probabilistic and handpicked sites ~~(USEPA 2023)~~(USEPA 2023). We found that summing predicted occurrence probabilities for all ~~taxa~~genera modeled ~~overestimated~~could potentially overestimate observed genus richness (i.e. observed vs predicted richness intercepts < -1.5). Upon further inspection we found that this overprediction was due to many ~~taxa~~genera having exceptionally low predicted occurrence probabilities ~~potentially owing to the large spatial extent of our ecoregions and~~but nonetheless present in the regional ~~taxa~~species pool. Thus, we used a threshold to excluded ~~taxa~~genera with predicted occurrence probabilities < 0.05 or < 0.10 prior to summation to correct this bias (Table 3).

Because site level random effects accounted for unmeasured factors in the models, we did not expect our models to have strong predictive power when projected to new locations (Abrego and Ovaskainen 2023, Kopp et al. 2023). However, we could assess whether values obtained at revisited sites were within the predicted posterior distribution of the fitted models as a secondary form of validation. Although the number of revisited sites available for model testing was low (Table 1), we found that >90% of the genus richness values observed during revisit sampling were within the posterior distribution of the models. Importantly, this lends plausibility that the posterior distribution reflects processes characterizing macroinvertebrate assemblages (Table 3 and Appendix S1: Figure S2).

The probabilistic adaptation of Jaccard similarity revealed that the predicted and observed assemblages were on average ~~48-56% similar (Table 3).~~>50 similar (Table 3). Sites located the WMT tended to have the most similar assemblages (mean = 0.6, 5th quantile = 0.36, 95th quantile= 0.75) while sites in the XER tended to have the least similarity (mean = 0.23, 5th quantile = 0.23, 95th quantile= 0.71). Although potentially vulnerable to sample size constraints, mean similarity between model predictions and assemblages observed at revisited sites were also relatively high, with mean values ranging between 0.35 and 0.68 (Table 3).

Disturbance effects on physiochemical conditions

We tested whether using the random forest models to predict physiochemical conditions after removing the effects of anthropogenic activities were susceptible to extrapolation. We found predictions at 77 of the sites (~4%) were potentially susceptible to extrapolation. This indicates that removing disturbance would cause some sites to be sufficiently different from the data used to train the models and suggest that these sites do not have a natural analog. Indeed, most of the sites flagged for potential extrapolation were located in the Coastal Plains Ecoregions (18%) and were associated with models used to predict NTL concentrations in the lower Mississippi and SUBD along the southern coast (Appendix S1: Figure S1). To avoid the potential for bias associated with extrapolation, we designated them as “Not Assessed” in our analysis.

For sites that were not potentially susceptible to extrapolation, removing anthropogenic disturbance from the physiochemical environment generally decreased nutrient concentrations, salinity, and riparian disturbance and increased substrate diameter (Figure 4). Change in NTL and PTL concentrations were most evident in Temperate Plains and Southern Plains and change in CL was most evident in the Northern Appalachians. Substrate diameter was noticeably coarser after hindcasting for all regions except Xeric, Southern Appalachians, and Northern Appalachians. ~~Table 3. Regression coefficients between predicted and observed richness and compositional similarity. Occurrence probabilities thresholds (Pr) were used to exclude taxa that with low predicted occurrence probabilities. We considered $R^2 \geq 0.2$, $-1.5 \leq$ intercept ≤ 1.5, and $0.85 \leq$ slope ≤ 1.15 are indicative of adequate model performance (Linke et al. 2005). Values in parentheses are the 5th and 95th percentile of estimates from the 3000 posterior samples.~~

	~~Pr~~	~~Intercept~~	~~Slope~~	~~R^2~~	~~Mean Jaccard Similarity~~
~~CPL~~	~~0.05~~	~~0.37 (-0.37 - 1.02)~~	~~1.00 (0.96 - 1.05)~~	~~0.85 (0.82 - 0.87)~~	~~0.49 (0.24 - 0.65)~~
~~NAP~~	~~0.05~~	~~-0.45 (-2.02 - 1.02)~~	~~1.03 (0.99 - 1.08)~~	~~0.74 (0.71 - 0.78)~~	~~0.51 (0.29 - 0.63)~~
~~NPL~~	~~0.05~~	~~-0.07 (-1.21 - 1.04)~~	~~1.02 (0.96 - 1.09)~~	~~0.76 (0.71 - 0.81)~~	~~0.56 (0.33 - 0.73)~~
~~SAP~~	~~0.05~~	~~0.40 (-0.66 - 1.41)~~	~~1.00 (0.97 - 1.04)~~	~~0.85 (0.82 - 0.87)~~	~~0.5 (0.25 - 0.65)~~
~~SPL~~	~~0.10~~	~~0.48 (-0.45 - 1.29)~~	~~1.02 (0.97 - 1.09)~~	~~0.79 (0.75 - 0.83)~~	~~0.48 (0.2 - 0.68)~~
~~TPL~~	~~0.05~~	~~0.06 (-0.91 - 0.94)~~	~~1.01 (0.97 - 1.06)~~	~~0.82 (0.79 - 0.84)~~	~~0.52 (0.28 - 0.72)~~
~~UMW~~	~~0.05~~	~~-0.74 (-2.52 - 0.99)~~	~~1.04 (0.98 - 1.10)~~	~~0.73 (0.66 - 0.78)~~	~~0.52 (0.33 - 0.66)~~
~~WMT~~	~~0.10~~	~~0.24 (-1.24 - 1.59)~~	~~1.04 (0.99 - 1.10)~~	~~0.74 (0.70 - 0.78)~~	~~0.55 (0.32 - 0.72)~~
~~XER~~	~~0.05~~	~~-0.12 (-1.11 - 0.74)~~	~~1.03 (0.98 - 1.08)~~	~~0.83 (0.80 - 0.86)~~	~~0.48 (0.18 - 0.66)~~

Observed and hindcasted values for sulfate, and mean summer air temperature were generally similar for all ecoregions.

Anthropogenic disturbance affected physiochemical gradients differently depending on the ecoregion (Figure 5a). We found 60.8% (95% Confidence Interval, hereafter CI = 52.8-

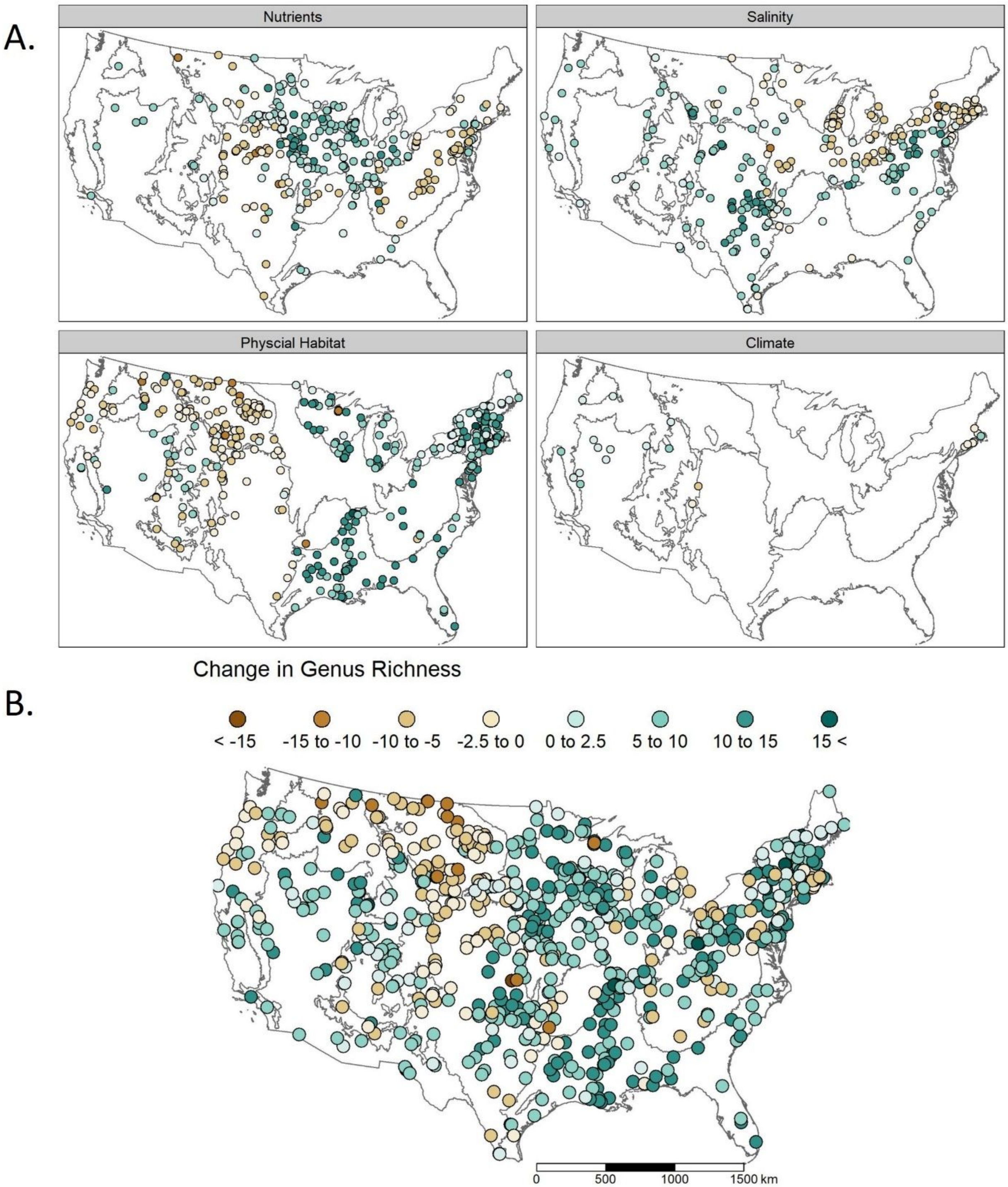
68.8%) of streams in Temperate Plains and 38.4% (CI = 32.0-44.9%) of streams Upper Midwest had elevated NTL relative to hindcasted estimates. In the Northern Appalachians, 41.4% (CI = 34.5-48.2%) of streams were found to have potentially elevated CL concentrations. Streams with excess fine substrates were most prevalent in the Temperate Plains (29.2%, CI = 21.2-37.1%) and Northern Plains (21.4%. CI = 14.1-28.9%). We also found similarities among regions. RPDI and TPRCP were the most prevalent disturbance among all ecoregions while disturbed SO4 and MSAT values were the least prevalent. Finally, anthropogenic disturbance could alter multiple physiochemical variables simultaneously (Figure 5b). In general sites with >2 physiochemical variables disturbed by anthropogenic activities were located in Northern Appalachians, Southern Plains, Temperate Plains, and Upper Midwest.

Macroinvertebrate assemblage response to undisturbed physiochemical environment

~~We evaluated how~~ Macroinvertebrate models were fit for each region separately and we assessed whether removing anthropogenic disturbance would produce estimates that exceed the range of values used to fit each model. Hindcast predictions exceeded the range of present-day conditions at 1 location in CPL, SAP and XER, 6 locations in TPL, and 12 locations in WMT. In these instances, we reset the hindcasted values to be within the range (minimum) of data used to fit the mode to avoid potential extrapolation.

Hindcast predictions for each category of physiochemical variables generated different outcomes for macroinvertebrate assemblages ~~would respond to removing anthropogenic disturbance from the abiotic variables using 5 scenarios. The effects of each scenario varied among regions and could increase or decrease genus richness (Figure 5a).~~ (Figure 6a). For example, ~~the nutrient scenario (~~removing anthropogenic disturbance from NTL and PTL~~), tended~~

~~to increase~~ concentrations, produced macroinvertebrate assemblages with higher genus richness in the Temperate Plains ~~but decrease~~and lower richness at some sites in the Southern Plains and Southern Appalachians ecoregions~~. The salinity scenario~~ (Figure 6a). On the other hand, removing disturbance from CL and SO4 tended to increase genus richness in the Southern Plains but ~~decrease~~decreased richness at sites in the Northern Appalachians. ~~The~~Hindcasted physical habitat ~~scenario, tended to increase richness in the Coastal Plains, Upper Midwest and Northern Appalachians and decrease richness in the Northern Plains and Western Mountains. Finally, the climate scenario had a relatively small effect on~~ variables, mostly affected macroinvertebrate assemblages in the Northern Plains by decreasing richness but also increased genus richness in Coastal Plains, and Northern Appalachians and Upper Midwest (Figure 6a). TPRCP generally contributed to changes in macroinvertebrate assemblages because relatively few sites had ~~differences between observed and 1940-1950 averages for~~ MSAT ~~and TPRCP~~temperature anomalies exceed ~~>~~ 2 standard deviations~~. The pristine scenario changed all abiotic gradients to reflect the removal of disturbance and~~ from 1900-1950 mean. In the Southern Appalachians genus richness decreased from present day conditions while in the Northern Plains genus richness generally increased from present day conditions. Removing hindcasting all physiochemical variables typically increased genus richness ~~(Figure 5b). Collectively these results suggest that human alteration of nutrient and salinity concentrations and habitat typically reduces macroinvertebrate assemblage richness.~~in all regions except for the Northern Plains and Southern Appalachians (Figure 6b).



~~Figure 5: Difference in macroinvertebrate genus richness predicted from present-day and after removing anthropogenic disturbance from each category of environmental variable (A) and all variables simultaneously (B). The points in each panel are sites that had a change in genus richness with >75% support after removing disturbance.~~

Identifying genera that had higher present-day occurrence probabilities (increasers), or higher hindcast occurrence probabilities (decreasers) ~~elucidates how major taxonomic groups may respond to disturbance (Table 4). For example, if disturbance was removed from the physiochemical variables, our results indicate that occurrence probabilities of Ephemeroptera, Plecoptera and Trichoptera (EPT) would on average increase at 26% of the sites in CPL. Alternatively, members of Mollusca in NPL tend to increase at 33% of the sites. More generally, insects in CPL, NAP, SAP, TPL, UMW and XER tended to be decreasers while non-insects tended to be increasers at more sites in all regions except CPL.~~

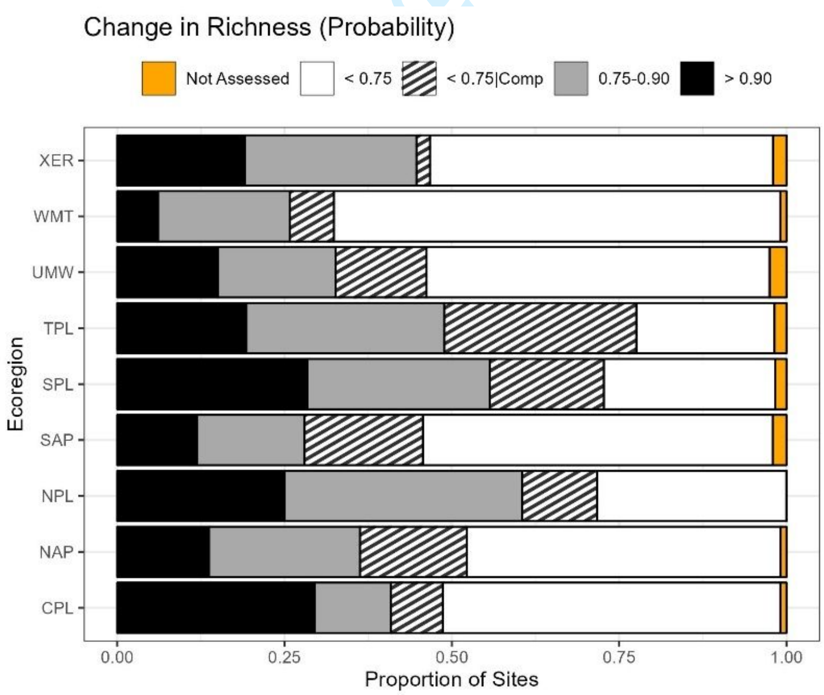
indicated that insects tended to be decreasers while non-insects tended to be increasers, but this was not consistent across all regions (Table 4). For example, in the Southern Plains and Western Mountains, insects were almost equally likely to be increasers (0.1 and 0.6, respectively) or decreasers (0.09 and 0.03, respectively) and in the Northern Plains both insects and non-insects tended to be increasers at a larger proportion of sites (0.19 and 0.22). Among insects, members of chironomidae were generally most likely to be increasers. Among non-insects, in contrast, members of mollusca were most likely increasers (Table 4). Thus, higher present-day genus richness compared to hindcast estimates could be due to genera that are typically considered to be tolerant of human activities.

Using the probabilistic sites, we aggregated our results to reflect the total population of streams and rivers in each ecoregion and detected that 14.3 – 75.5% of streams have present-day assemblages that differ from assemblages expected from the hindcasted physiochemical environment (Figure 7). With respect to changes in richness with support >0.75, the Northern Plains (58.3%), Temperate Plains (54.8%), and Southern Plains (54.2%) the most streams with potentially altered macroinvertebrate assemblages while the Western Mountains (9.2%), Upper Midwest (19.6%), and Northern Appalachians (29.7%) have the fewest. Compositional change without a corresponding change in genus richness could increase the percentage of streams with altered assemblages. Evidence for compositional change was most pronounced in Temperate Plains (20.7%), Southern Appalachians (20.7 %) and Upper Midwest (18.8%). All regions had a percentage of streams that could not be assessed because of potential extrapolation of hindcasted physiochemical variables, incomplete data or macroinvertebrate assemblages consisting of only rare genera (i.e. prevalence < 10%). Among them, the Coastal Plains (12.9%) and Xeric (11.1%) had the largest percentage of streams that were not assessed, thus the total number of sites assessed was less than the total number of probabilistic sites surveyed (n = 1748).

~~Table 4: Mean proportion of sites in each taxonomic group that was identified as an increaser or decreaser. Increasers (I) are genera that have a significantly higher probability of occurrence under present day conditions compared to hindcasted conditions; decreasers (D) are genera that have a significantly lower probability of occurrence under present-day conditions compared to hindcasted conditions. The values in the table are the mean proportion of sites for genera within the major taxonomic group. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = xeric. EPT = Ephemeroptera, Plecoptera and Trichoptera.~~

	CPL		NAP		NPL		SAP		SPL		TPL		UMW		WMT		XER	
	D	I	D	I	D	I	D	I	D	I	D	I	D	I	D	I	D	I
Insects	0.17	0.08	0.07	0.05	0.08	0.19	0.10	0.05	0.09	0.10	0.17	0.11	0.12	0.07	0.03	0.06	0.11	0.06
—EPT	0.26	0.00	0.08	0.04	0.12	0.14	0.12	0.04	0.09	0.03	0.24	0.04	0.16	0.01	0.05	0.05	0.13	0.01
—CHIRONOMIDAE	0.13	0.10	0.07	0.05	0.02	0.19	0.05	0.06	0.03	0.13	0.11	0.14	0.06	0.13	0.02	0.07	0.07	0.10
—Other Insects	0.17	0.06	0.07	0.07	0.06	0.24	0.11	0.05	0.21	0.03	0.20	0.11	0.10	0.02	0.03	0.02	0.13	0.01
Non-Insects	0.14	0.13	0.04	0.17	0.05	0.22	0.07	0.11	0.05	0.17	0.12	0.13	0.07	0.13	0.04	0.08	0.04	0.12
—ARTHROPODA	0.23	0.12	0.04	0.18	0.09	0.01	0.08	0.09	0.07	0.07	0.20	0.24	0.12	0.00	0.05	0.04	0.04	0.04
—MOLLUSCA	0.08	0.10	0.05	0.21	0.01	0.33	0.09	0.14	0.09	0.09	0.11	0.08	0.08	0.05	0.01	0.07	0.01	0.20
—Other Non-Insects	0.19	0.16	–	0.13	0.01	0.15	0.01	0.11	0.01	0.24	0.06	0.13	0.01	0.24	0.04	0.12	0.06	0.11

~~We identified sites that had a Jaccard similarity < 0.9 to assess compositional changes at sites that did not change in richness. Changes in composition were most pronounced in TPL, SPL, SAP, and NAP (> 15% of sites). Adding sites with evidence of compositional change to sites that had evidence of a change in genus richness revealed that removing anthropogenic genus richness or composition at 32-78% of the sites within an ecoregion (Figure 6). Macroinvertebrate assemblages would change the least in the WMT ecoregion and the most in the NPL, TPL and SPL. Considering only changes in richness revealed similar patterns but the proportion of sites identified to potentially change depended on the level of support. Considering sites with > 0.75 support for differences in richness revealed that disturbance could change the number of genera at 26-61% while the more conservative threshold (i.e. > 0.9 support) revealed that 6-30% sites might change in richness.~~



~~Figure 6: Proportion of sites where genus richness or composition could change if anthropogenic disturbance was removed from the physiochemical environment. Shaded bars indicate that the probability of mean present-day richness differs from hindcast with > 0.90 (Black) or 0.75-0.90~~

~~(Gray) support. Hatched bars indicate that compositional similarity was < 0.9 but support for a difference in richness were < 0.75. White bars indicate the proportion of sites that had < 0.75 support for change in richness and >0.9 compositional similarity. Yellow bars indicate the proportion of sites that could not be assessed because of insufficient data.~~

Discussion

Assess the effects of anthropogenic disturbance on biological assemblages is challenging because few minimally disturbed sites remain (Stoddard et al. 2006). Sites in least-disturbed are often adopted as an alternative but are difficult to consistently define, vary in quality, and can be spatially aggregated (Hawkins et al. 2010, McNellie et al. 2020). The model-based framework we developed contributes to a suite of other approaches intended to circumvent the need for reference sites (Chessman and Royal 2004, Elias et al. 2016, Yuan et al. 2024). Specifically, we evaluated the effects of human disturbance on several physiochemical variables and, in turn, addressed whether altered physiochemical conditions affect macroinvertebrate assemblages. Using this approach, we found that anthropogenic disturbance can affect multiple physiochemical variables simultaneously and that the effects on any single factor can vary among ecoregions. We also found that removing or reducing disturbance could change genus richness at >50% of the streams in some ecoregions and up to 75% if compositional change is considered. Collectively, our framework offers a promising alternative to evaluating the effects of specific disturbances on macroinvertebrate assemblages that does not rely on reference sites.

Hindcasting physiochemical conditions

Random forest models have been used by others to predict physiochemical variables if human disturbance was reduced or removed (Yuan et al. 2024). In general, using models to infer physiochemical conditions eliminates the need to identify reference sites or collect excessive data from undisturbed locations (Herlihy and Sifneos 2008, Soranno et al. 2011). In this study, random forest models improve on earlier regression-based hindcasting approaches (Dodds and Oakes 2004, Herlihy and Sifneos 2008) because they incorporate a relatively large number of natural and anthropogenic predictor variables and accommodate their complex relationships with the response variables (Yuan et al. 2024). Importantly, because reference sites may not be representative of all sites that need to be assessed including natural and anthropogenic variables enabled us to make site-specific predictions rather than relying on an estimated mean value (i.e. the intercept from a multiple regression as a function of only disturbance) for an entire region (Dodds and Oakes 2004, Herlihy and Sifneos 2008). Thus, our hindcasting approach provides a first-order approximation for what conditions at a site could be if they are presently unknown.

Furthermore, random forest models permitted us to evaluate the relative importance of specific anthropogenic activities and characterize potential site and regional scale differences after reducing or removing them. For example, higher than expected total nitrogen in Temperate Plains and Upper Midwest and the relatively high importance of percent agriculture landcover our model suggests that nutrient inputs associated with agricultural activities may be elevating in-stream nutrient concentrations (Lin et al. 2021, Sabo et al. 2023). Similarly, elevated chloride concentrations in the Northern Appalachians and the relative high importance of road density in the model, could reflect the contribution of road salts to freshwater salinization (Kaushal et al. 2023).

Certainly, inferring physiochemical conditions for different levels of human disturbance depends on the quality and structure of the model. Although our models were evaluated using an independent validation dataset, hindcasting required a dataset that reduced or eliminated the effect of predictor variables associated with human disturbance. As consequence, the dataset used for hindcasting could be dissimilar from the data used to calibrate the model and potentially generate errors associated with extrapolation (Meyer and Pebesma 2021). For most regions, the hindcasting dataset was sufficiently similar to the training dataset such that we could assess the alteration of physiochemical variables for majority of sites. Nonetheless 18% of sites surveyed the Coastal Plains were flagged as a result of testing for extrapolation. Considering that many of the sites were located in the lower Mississippi River Basin, a unique and heavily modified system, it is perhaps unsurprising that removing human disturbances from this system is beyond the domain of our model. Testing for potential extrapolation in the context of hindcasting is novel and, although the threshold to determine whether our results were prone to extrapolation has been used previously (Yuan et al. 2024), more research should be devoted to optimizing and interpreting this threshold. When locations or streams are potentially susceptible to extrapolation, it might be more reasonable to select values that represent management goals, best professional judgement or those that maximize societal benefits (Bouleau and Pont 2015).

To assess how human-related disturbance effects physiochemical environment, we used standardized anomalies (i.e. z-scores), to account for either unexplained variation in the model values or variation along a baseline. Thresholds are of central importance for the communication and evaluation of human disturbances communication (Wood 2008). Here, we used a ±2SD threshold for deciphering whether human disturbances alter physiochemical conditions (Kilgour and Stanfield 2005). Although this threshold is intuitive, it could potentially be too strict for some variables. For example, values for MSAT at many locations did not exceed the threshold in our analysis, suggesting that 2018 and 2019 temperature averages are consistent with 1900-1950 averages, given interannual variation. However, warming of 1SD has been implicated in reductions of insects in agricultural landscapes (Outhwaite et al. 2022) and there is evidence for increasing summer air temperatures in the United States of approximately 0.09°C per decade since 1901 (USEPA 2024a). On the other hand, to evaluate riparian disturbance, we used a threshold of 0.33 which suggests that human disturbance is not completely absent for all sites that were below this threshold. Indeed, nearly every site in our analysis has RPDI >0 such that this threshold did not informatively differentiate regional variation in disturbance. Further, this is the threshold used by others to identify least disturbed sites (Kaufmann et al. 2022b, USEPA 2023) and we sought to evaluate how this could affect microinvertebrate assemblages. Although thresholds were necessary to identify which regions may have relatively more disturbed physiochemical conditions that others, future applications may select different thresholds, report continuous scores, or vary the amount of disturbance to meet specific applications (Yuan et al. 2024).

Application for biological assessment without reference sites

~~Modeling has improved our ability to quantify the effects of anthropogenic disturbance on biological assemblages (Hawkins et al. 2000, Hawkins 2006, Bailey et al. 2014). However, these approaches have traditionally relied on collecting reference assemblages from sites that are representative of a least-disturbed state. Because streams and rivers are among the most degraded ecosystems (Dudgeon et al. 2006, Vörösmarty et al. 2010), sites in least disturbed condition can be spatially aggregated or vary in quality among ecoregions (Herlihy et al. 2008, McNellie et al. 2020). This makes it challenging to elucidate the full extent of anthropogenic disturbance on biological assemblages. Two-stage modeling approaches that combines estimates of abiotic conditions without human disturbance taxon-environment relationships could improve our understanding of how disturbance affects macroinvertebrate assemblages (Yuan et al. 2024). Our~~

~~approach builds on previous approaches by using a JSDM framework. We found that anthropogenic disturbance can simultaneously affect multiple physiochemical gradients and that removing disturbance could change genus richness at >50% of the sites and up to 75% if compositional change is considered. Typically, sites increased in richness under hindcast conditions, indicating that genus richness is presently lower than it would be without human disturbance. Because JSDM also generate occurrence probabilities for each taxon, we assessed taxon-specific differences between hindcasted and present-day conditions predictions for each site. In general, we found that insect taxa typically have lower predicted probabilities for more sites under present-day conditions compared to non-insects. This could support the notion that insect taxa tend to decrease in the presence of anthropogenic disturbance while non-insect taxa increase (Rumschlag et al. 2023).~~

~~*Hindcasting physiochemical conditions*~~

~~Anthropogenic disturbance affects macroinvertebrate assemblages by modifying the in stream physiochemical environment~~ Using hindcast physiochemical variables to predict the assemblage expected to occur if anthropogenic disturbance was removed allowed us to infer the consequences of altering a relatively small number of physiochemical variables. Central to our application is that multivariate models can implement meaningful site-level random effects that statistically control for unmeasured environmental variables and potential biotic interactions ~~(Tang et al. 2020, Fergus et al. 2023)~~(Warton et al. 2015, Ovaskainen et al. 2016, Kopp et al. 2023)~~. We used a modeling approach to relate major anthropogenic disturbances in the watershed to several in stream physiochemical variables that influence the distribution of benthic macroinvertebrates (Kopp et al. 2023) and therefore circumvent the need to define background~~

~~physiochemical conditions using reference sites (Dodds and Oakes 2004, Herlihy and Sifneos 2008). Unsurprisingly, anthropogenic disturbance variables were ranked among the most important predictors for nutrients, salinity, and substrate diameter, supporting the notion that these environmental gradients are susceptible to alteration by anthropogenic activities (Paulsen et al. 2008, Lin et al. 2021, Zak et al. 2021, Kaufmann et al. 2022, Kaushal et al. 2023, Sabo et al. 2023). Overall, our results suggest that nutrient and salinity concentrations could be lower and substrate dimeter could be coarser if anthropogenic disturbances were removed.~~

~~The proportion of sites where we inferred altered physiochemical conditions varied among the ecoregions and could indicate regional differences in the extent and magnitude of specific anthropogenic activities, as well as the types of disturbance and their effects on their physiochemical and biotic conditions. For example, the larger portion of sites with higher than expected total nitrogen in Temperate Plains and Upper Midwest suggests that nutrient inputs associated with agricultural activities may be elevating in stream nutrient concentrations~~. Although, this feature kept our model sufficiently tractable it limits our ability to make accurate predictions to new locations (Abrego and Ovaskainen 2023). However, predicting to new locations was not our objective. Rather, our models establish empirical relationships between macroinvertebrate occurrences and then evaluate how those assemblages might differ if select physiochemical conditions changed while all else remained constant. Furthermore, because the data we used for our analysis was collected as part of a probabilistic survey, designed to be representative of the population of stream and rivers (Olsen and Peck 2008), our site-specific inferences can be aggregated to elucidate regional and sub-continental patterns. Indeed, it may be undesirable to refit models every time an assessment is needed, and future efforts should focus on improving predictive abilities by including immutable factors in addition to stressor gradients (Yuan et al. 2024).

Importantly our approach differs significantly from traditional, reference site-based approaches ~~(Lin et al. 2021, Sabo et al. 2023)~~(Hawkins et al. 2000, Herlihy et al. 2008, USEPA 2023)~~. Similarly, chloride concentrations exceed hindcasted estimates for the majority of streams in Northern Appalachians, potentially reflecting the contribution of road salts in freshwater salinization (Kaushal et al. 2023). The Temperate Plains had a relatively large proportion of sites with finer substrates and reduced natural riparian vegetation and emphasizes the importance of maintaining native vegetation in riparian areas (Matthews 1988, Dodds et al. 2004). Finally, regional sulfate concentrations in streams can be influenced by acid mine drainage, fertilizer leaching, and industrial wastewater inputs (Herlihy et al. 1990, Zak et al. 2021). We found that approximately 10% of sites in the Southern Appalachians and Upper Midwest could have lower concentrations of sulfate, possibly implicating these activities in elevating sulfate concentrations. The difference between present-day and 1940-1950 climate averages were generally less than 2 standard deviations, indicating similarities in mean summer temperature and total annual precipitation. However, averages may not capture changes in variability or extreme events and could potentially oversimplify several other important aspects of climatic change that have occurred in the United States (USEPA 2024). Importantly, using landscape variables to predict in-stream physiochemical provides a more direct linkage between human disturbance and biological assemblages than using landcover variables alone (Kilgour and Stanfield 2005, Yuan et al. 2024).~~

. Foremost, our efforts focus explicitly on a relatively small number of potential stressors whereas reference-site approaches focus implicitly on a theoretically larger, but undefined number of stressors. For example, the difference between a test site and reference sites could be related to a number of other stressors that were not used as biotic screens but co-occur with them (Herlihy et al. 2008, Paulsen et al. 2008). Alternatively, in our model-based approach, differences between present day and hindcast assemblages are only related to changes in the physiochemical variables included as fixed effects. Although the former may elucidate general disturbance effects, it is difficult to attribute differences between assemblages to a specific environmental disturbance (Paulsen et al. 2008). On the other hand, our model-based approach enhances interpretations with respect to specific physiochemical variables but may omit important anthropogenic stressors that were not explicitly included in the model. Thus, the choice between reference site-based and model-based assessments may be contingent on whether the environmental gradients that are commonly disturbed by human activities can be identified and appropriately modeled.

The JSDMs were also fitted using bayesian inference and, as such, yielded posterior distributions that can be used for hypothesis testing (Johnson et al. 2022). Specifically, we evaluate whether the effects of human disturbance on physiochemical variables were sufficient to alter benthic assemblages on a site-specific basis. In contrast, traditional reference site-based approaches evaluate each test site based on quantiles of the distribution obtained from reference sites (Herlihy et al. 2008, USEPA 2023). In this regard, JSDMs and bayesian inference may be advantageous in interpretating biological condition estimates because they account for uncertainty in predictions for each site. Furthermore, we also included all genera from the regional taxa pool while some typical reference site-based approaches include only taxa that occur at reference sites. This implies that in the absence of human disturbance those taxa should occur at all locations regardless of other taxa. This is problematic because genera that tolerate anthropogenic disturbance probably evolved under similar conditions that occurred naturally or have remarkable plasticity (Wiens et al. 2010, Heino et al. 2013). Finally, our model-based approach can reveal compositional changes. Although the probabilistic adaptation of Jaccard's similarity calculation suggests that predicting individual genera is more challenging than aggregated genus richness, this analysis provided additional information that is generally not easily available from other approaches (Hawkins 2006, Van Sickle 2008) and enhanced our ability to identify sites that are potentially affected by human disturbance.

Selecting the appropriate taxonomic level that specimens should be identified to is a critical decision in biological assessment (Chessman et al. 2007). We focused exclusively on taxa that were identified to the genus level and excluded those that could not be unambiguously identified. This possibly increases false absences rates in our study, but NRSA identifies most organisms to their genus such that these instances are relatively rare. Further, common taxonomy alleviates some ambiguity associated with aggregating unresolved taxa into operational taxonomic units (Yuan et al. 2008) and improves transferability of our genus-environment relationships to other studies. Fixed-count subsampling, performed as part of the standardized NRSA protocol, could also increase false absences in our study. Although, explicitly modeling taxon-specific detection probabilities from replicate subsamples may be an interesting avenue for future research (Doser et al. 2023, Doser et al. 2024), these data are presently unavailable. Finally, focusing on genera could also reduce interspecific variation, but species-level taxonomy was not available for this dataset. In general, finer taxonomic resolution would be substantially more expensive and perhaps only yield marginal benefits (Chessman et al. 2007). Nonetheless, metabarcoding approaches may have shown promise for bioassessments (Smucker et al. 2024) and could avoid some of the limitations associated with selecting the appropriate taxonomic resolution.

Although the model-based approaches may have some advantages over reference site-based approaches, they require further investigation before they can be fully assimilated into biomonitoring programs. Nevertheless, the National Rivers and Streams Assessment reports biological condition estimates at for the 2018-19 survey using a multi-metric index (USEPA 2024b) and it is worthwhile to compare our results to those. The results from NRSA are available at https://riverstreamassessment.epa.gov/dashboard. Based on a ranked comparison, the largest disagreements pertained to the percentage of stream miles in poor condition in the Plains ecoregions. Specifically, our model-based approach showed more streams to be altered in the Northern Plains, Southern Plains and Temperate Plains. Indeed, the US Great Plains (i.e. Temperate Plains, Northern Plains and Southern Plains) have undergone extensive conversion from grasslands to agriculture such that there may be few sites that are undisturbed (Samson and Knopf 1994, Dodds et al. 2004, Olimb and Robinson 2019). Because of the large extent of anthropogenic activities in these regions, it is likely that the reference sites used to assess biological condition are potentially lower quality (Herlihy et al. 2008). Conversely, NRSA ranked the Coastal Plains as having a larger percentage of streams in poor condition than our model-based approach. Unfortunately, because of potential extrapolation associated with our model-based assessment, the Coastal Plains also had a relatively large number of streams that could not be assessed. The rankings for the remaining ecoregions tended be similar and both methods agreed that streams in the Western Mountains were the least-disturbed among the ecoregions. Indeed, about 75% of the land area of this ecoregion is in federal ownership, which could convey some protection to streams and rivers (Jenkins et al. 2015). Furthermore, Western

Mountains have a relatively shorter history of anthropogenic disturbance than other ecoregions in the United States.

Macroinvertebrate response to anthropogenic disturbance

~~Substituting hindcast physiochemical conditions into the JSDMs elucidated how macroinvertebrate assemblages may change if human disturbance was removed. Hindcast physiochemical conditions tended to increase estimates of genus richness, suggesting that in many ecoregions fewer genera persist when anthropogenic disturbances alter physiochemical conditions. Indeed, other studies that have found either modest increases in macroinvertebrate richness~~ Beyond applications for bioassessment, our model-based approach contributes to a growing literature devoted to understanding of recent trends in macroinvertebrates ~~(Gebert et al. 2022, Rumschlag et al. 2023)~~(Crossley et al. 2020, Jähnig et al. 2021, Gebert et al. 2022, Spake et al. 2022, Rumschlag et al. 2023) ~~or no clear evidence for a widespread decline of insects (Crossley et al. 2020). However, few datasets span more than several decades which limits our ability to assess trends beyond a relatively recent historical time period (Blüthgen et al. 2022) and could potentially reflect a shifting baseline effect (Humphries and Winemiller 2009, Soga and Gaston 2018). An advantage of our approach is that space-for-time substitutions leverage spatial variability to increase the length of the environmental gradient being assessed such that our pristine scenario represents an earlier period before major anthropogenic disturbances. Further, the survey design implemented by NRSA reduces confounding spatial and temporal dimensions that could be present in datasets compiled from multiple sources~~ . In our study, hindcasting estimates tended to show increases in genus richness, suggesting that in many ecoregions fewer genera persist when physiochemical conditions are disturbed. Other studies have reported either modest increases in macroinvertebrate richness (Gebert et al. 2022, Rumschlag et al. 2023) or no clear evidence for a widespread decline of insects (Crossley et al. 2020). Because few datasets span more than several decades, an advantage of our space-for-time approach is that it could reflect a period before major anthropogenic disturbances (Blüthgen et al. 2022). In addition, the survey design and consistent methodology implemented by NRSA reduces the potential for confounding spatial and temporal dimensions that could be present in datasets that were compiled from multiple sources ~~(Jähnig et al. 2021,~~ (Jähnig et al. 2021, Blüthgen et al. 2022~~, Boyd et al. 2023)~~, Boyd et al. 2023). Nonetheless, although our results suggest that sites typically support fewer macroinvertebrate genera under present-day conditions, we did detect increased in genus richness at several sites in the Northern Plains and Western Mountains which suggests that macroinvertebrate response to disturbance could be context dependent ~~(Powell et al. 2023)~~(Powell et al. 2023).

We compared genus-specific occurrence probabilities under present-day and hindcasted conditions to identify which genera contribute to assemblage level patterns. Insects in most ecoregions tended to decrease with disturbance while non-insects tended to increase. ~~To elucidate which~~This suggests that insect genera may, in general, be less tolerant to anthropogenic disturbance (Jähnig et al. 2021) and aligns with other studies that have documented declines in insect richness (Rumschlag et al. 2023). Among insects, Ephemeroptera, Plecoptera and Trichoptera (EPT) are often used as indicators of anthropogenic disturbance (Stoddard et al. 2008) and may have already lost a considerable proportion of species (Sánchez-Bayo and Wyckhuys 2019). This is especially concerning because we found that EPT occurrence probabilities could continue to decrease in many regions of the US. Among non-insects, non-arthropods included members of Annelida, Nemertea, and Platyhelminthes tended to increase at a relatively large proportion of sites which is consistent with high pollution tolerance values typically assigned to these taxa (Carlisle et al. 2007, Griffith 2023).

Because our approach was based on quantifying genus-environment relationships along environmental gradients that are typically altered by human activity, we could also compare the relative effects of changing specific categories of physiochemical variables ~~were most important in structuring~~ on macroinvertebrate assemblages~~, we created scenarios by changing either NTL and PTL (Nutrient Scenario), CL and SO4 (Salinity Scenario), RPDI and SUBD (Habitat Scenario), or MSAT and TPRCP (Climate Scenario). The increase in estimated~~. For example, higher genus richness under present-day conditions in the Northern Plains ~~and Western Mountains ecoregions was also present in the~~ compared to hindcasted habitat ~~scenario and~~variables suggests that disturbance may have increased ~~taxa richness by disturbing the riparian area or substrate diameter. Riparian disturbance was present at majority of these sites and the index we used to measure it consists of several direct measures of human disturbance (USEPA 2017a). Cattle~~genera richness. Interestingly, cattle in pasture and rangelands were ~~one of~~among the most common types of disturbance measured by ~~riparian disturbance index~~RPDI for ~~these regions~~this region ~~(USEPA 2023)~~(USEPA 2023) and could potentially increase ~~taxa~~genera richness by suppressing riparian forest cover and, in turn, elevating primary production ~~(Mittelbach et al. 2001, Tonkin et al. 2013)~~(Mittelbach et al. 2001, Tonkin et al. 2013). Conversely, in the Northern Appalachians and Coastal Plains, where the presence of trash, landfills or buildings were the most common factors ~~contributing to the riparian disturbance index~~for RPDI, we found that genus richness was presently lower which is consistent with the negative effects of urbanization on stream macroinvertebrate assemblages ~~(Morse et al. 2003).~~(Morse et al. 2003). We also found different effects with respect to changes in precipitation whereby genus richness increased in the Northern Plains and decreased in Southern Appalachians. Collectively, these ~~results elucidate~~ patterns reveal which environmental variables ~~could~~may be most important ~~for~~in structuring microinvertebrate assemblages and in what contexts~~.~~

~~We compared taxon-specific occurrence probabilities under present-day and hindcasted conditions to identify taxa that were considered to increase or decrease under anthropogenic disturbance. Insects in most ecoregions tended to decrease with disturbance while non-insects tended to increase. This suggests that insect taxa may, in general, be comparatively less tolerant to anthropogenic disturbance (Jähnig et al. 2021) and aligns with other studies that have documented declines in insect richness (Rumschlag et al. 2023). Among insects, Ephemeroptera, Plecoptera and Trichoptera (EPT) are often used as indicators of anthropogenic disturbance (Stoddard et al. 2008) and considered to have already lost a considerable proportion of species (Sánchez-Bayo and Wyckhuys 2019). We found that the occurrence probabilities for the EPT taxa~~ – an interpretation that ~~still exist could decrease at approximately 25% of sites in the Temperate Plains. Indeed, these taxa may~~would not be ~~especially sensitive to nutrient additions (Nessel et al. 2023) and the nutrient scenario we created here resulted in decreased richness at majority of these sites without removing disturbance from any other environmental gradient. Among non-insects, non-arthropods included members of Annelida, Nemertea, and Platyhelminthes tended to increase at a relatively large proportion of sites which is consistent with high pollution tolerance values typically assigned to these taxa (Carlisle et al. 2007, Griffith 2023).~~

~~The proportion of sites that changed in richness or composition could reveal differences among ecoregions in the effects of anthropogenic disturbance on macroinvertebrate assemblages.~~

~~We found disturbance effected macroinvertebrate assemblages at the greatest proportion of sites in the Northern Plains, Southern Plains and Temperate Plains compared to other regions. The US Great Plains have undergone extensive conversion from grasslands to agriculture such that there may be few sites that are undisturbed (Samson and Knopf 1994, Dodds et al. 2004, Olimb and Robinson 2019). Conversely, we found that present-day assemblages in the Western Mountains were generally similar to hindcast predictions. About 75% of the land area of this ecoregion is classified as federal ownership land which could convey some protection to streams and rivers (Jenkins et al. 2015). Further, this region has a relatively short history of anthropogenic disturbance compared to other regions of the United States, such that there may be a relatively large number if streams and rivers that are undisturbed. Our JSDM approach also allowed us to assess compositional change in addition to changes in richness resulting from disturbance. We found that assemblage composition at some sites in all regions could be affected by anthropogenic disturbance without a subsequent change in richness (Van Sickle 2008), which suggests that tolerant taxa may have replaced sensitive taxa at many sites (Jähnig et al. 2021, Rumschlag et al. 2023).~~

~~Application for biological assessment without reference sites~~

~~Identifying minimally disturbed reference sites to assess the effects of anthropogenic disturbance on biological assemblages is challenging because there are few existing sites in this condition (Stoddard et al. 2006). Least-disturbed is often adopted as a suitable alternative but is difficult to consistently define, vary in quality, and can be spatially aggregated (Hawkins et al. 2010, McNellie et al. 2020). These challenges have motivated several efforts to explore alternative approaches for assessing biological condition without relying on reference sites~~

~~(Chessman and Royal 2004, Kilgour and Stanfield 2005, Elias et al. 2016, Yuan et al. 2024). Beyond elucidating the potential effects of anthropogenic disturbance on macroinvertebrate assemblages, our approach complements these efforts by developing a two-stage modeling approach that combines random forest predictions of physiochemical conditions in the absence of anthropogenic disturbance~~possible with ~~taxon-environment relationships quantified using joint species distribution modeling.~~

~~Specifically, our efforts include the entire regional taxa pool and incorporated taxon-specific responses to physiochemical variables that are often altered by human activities. This is consistent with Yuan et al. (2024) but differs from~~ traditional ~~approaches that quantify relationships between taxa that occur at least-disturbed reference sites and natural (immutable) environmental gradients (e.g. watershed area, elevation, lithology) (Hawkins et al. 2000, Clarke et al. 2003, Hallstan et al. 2012). By focusing on physiochemical variables that are commonly altered by human disturbance, we can more directly infer the consequences of disturbance, or the potential effectiveness of efforts to restore or remediate specific stressors, on macroinvertebrate assemblages. We also used site-level random effects to statistically control for unmeasured environmental variables or biotic interactions using JSDMs (Kopp et al. 2023). This had the added benefit of keeping our model sufficiently tractable but assumes that only the environmental variables we included in our analysis act as agents of change for the assemblage and limits our ability to make accurate predictions to new locations. Indeed the data we used for our analysis were collected as part of a probabilistic survey specifically designed to be representative of the population of stream and rivers (Olsen and Peck 2008) thus our site-specific analysis should capture general patterns across CONUS. Nonetheless future efforts to develop~~

~~these models for biological assessment purposes will benefit from improving the predictive capability of these models~~reference site-based approaches.

Conclusions

We used a ~~two-stage modeling process~~model-based approach to assess the potential effects of anthropogenic disturbance on physiochemical gradients and benthic macroinvertebrate assemblages. Our approach combines ~~taxon~~genus-environment relationships with estimates of a number of important dimensions in physiochemical ~~conditions~~condition after removing anthropogenic disturbance. ~~Increasingly~~The number of sites in minimally disturbed condition are ~~disappearing and~~progressively diminishing, so methods that circumvent the need for reference sites ~~in the~~for biological ~~assessment~~assessments of streams and rivers are crucial to understanding the extent of anthropogenic impacts. Importantly, ~~this~~our framework could provide an avenue to conduct biological assessment without depending on least disturbed reference sites.

Acknowledgements

We thank S. Paulsen, L. Riato, E. Fergus, K. Blocksom, and D. Peck for valuable discussions throughout the development of this project. D. Benkendorf~~ and~~, S. Rumschlag, D. Carlisle and an anonymous reviewer provided comments on earlier versions of this manuscript. The information in this document has been funded entirely by the U.S. Environmental Protection Agency, in part through appointments to the Internship/Research Participation Program at the

Office of Research and Development, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency. Any mention of trade names, products or services does not imply an endorsement by the U.S. Government or the U.S. Environmental Protection Agency. The EPA does not endorse and commercial products, services, or ~~enterprises.~~

enterprise

References:

~~Bailey, R. C., S. Linke, and A. G. Yates. 2014. Bioassessment of freshwater ecosystems using the Reference Condition Approach: comparing established and new methods with common data sets. Freshwater Science **33**:1204-1211.~~

Abrego, N., and O. Ovaskainen. 2023. Evaluating the predictive performance of presence–absence models: Why can the same model appear excellent or poor? Ecology and Evolution **13**:e10784.

Blüthgen, N., M. Staab, R. Achury, and W. W. Weisser. 2022. Unravelling insect declines: can space replace time? Biology Letters **18**:20210666.

Bouleau, G., and D. Pont. 2015. Did you say reference conditions? Ecological and socio-economic perspectives on the European Water Framework Directive. Environmental Science & Policy **47**:32-41.

Boyd, R. J., G. D. Powney, and O. L. Pescott. 2023. We need to talk about nonprobability samples. Trends in Ecology & Evolution **38**:521-531.

Breiman, L. 2001. Random Forests. Machine Learning **45**:5-32.

Buss, D. F., D. M. Carlisle, T.-S. Chon, J. Culp, J. S. Harding, H. E. Keizer-Vlek, W. A. Robinson, S. Strachan, C. Thirion, and R. M. Hughes. 2014. Stream biomonitoring using macroinvertebrates around the globe: a comparison of large-scale programs. Environmental Monitoring and Assessment **187**:4132.

Calabrese, J. M., G. Certain, C. Kraan, and C. F. Dormann. 2014. Stacking species distribution models and adjusting bias by linking them to macroecological models. Global Ecology and Biogeography **23**:99-112.

Carlisle, D. M., M. R. Meador, S. R. Moulton, and P. M. Ruhl. 2007. Estimation and application of indicator values for common macroinvertebrate genera and families of the United States. Ecological Indicators **7**:22-33.

~~Chen, K., J. R. Olson, J. J. Vander Laan, R. A. Hill, B. Wang, and C. P. Hawkins. 2019. Improving the performance of ecological indices by balancing reference site quality and representativeness. Hydrobiologia **837**:177-194.~~

Chessman, B., S. Williams, and C. Besley. 2007. Bioassessment of streams with macroinvertebrates: effect of sampled habitat and taxonomic resolution. Journal of the North American Benthological Society **26**:546-565.

Chessman, B. C., and M. J. Royal. 2004. Bioassessment without reference sites: use of environmental filters to predict natural assemblages of river macroinvertebrates. Journal of the North American Benthological Society **23**:599-615.

~~Clarke, R. T., J. F. Wright, and M. T. Furse. 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modelling **160**:219-233.~~

Crossley, M. S., A. R. Meier, E. M. Baldwin, L. L. Berry, L. C. Crenshaw, G. L. Hartman, D. Lagos-Kutz, D. H. Nichols, K. Patel, S. Varriano, W. E. Snyder, and M. D. Moran. 2020. No net insect abundance and diversity declines across US Long Term Ecological Research sites. Nature Ecology & Evolution **4**:1368-1376.

Deflem, I. S., E. Bennetsen, Ø. H. Opedal, F. C. F. Calboli, O. Ovaskainen, G. Van Thuyne, F. A. M. Volckaert, and J. A. M. Raeymaekers. 2021. Predicting fish community responses to environmental policy targets. Biodiversity and Conservation **30**:1457-1478.

Dodds, W. K., K. Gido, M. R. Whiles, K. M. Fritz, and W. J. Matthews. 2004. Life on the Edge: The Ecology of Great Plains Prairie Streams. BioScience **54**:205-216.

Dodds, W. K., and R. M. Oakes. 2004. A technique for establishing reference nutrient concentrations across watersheds affected by humans. Limnology and Oceanography: Methods **2**:333-341.

Doser, J. W., A. O. Finley, and S. Banerjee. 2023. Joint species distribution models with imperfect detection for high-dimensional spatial data. Ecology **104**:e4137.

Doser, J. W., A. O. Finley, M. Kéry, and E. F. Zipkin. 2024. spAbundance: An R package for single-species and multi-species spatially explicit abundance models. Methods in Ecology and Evolution **n/a**.

Dudgeon, D., A. H. Arthington, M. O. Gessner, Z. Kawabata, D. J. Knowler, C. Lévêque, R. J. Naiman, A. H. Prieur-Richard, D. Soto, M. L. Stiassny, and C. A. Sullivan. 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. Biol Rev Camb Philos Soc **81**:163-182.

Dumelle, M., T. Kincaid, A. R. Olsen, and M. Weber. 2023. spsurvey: Spatial Sampling Design and Analysis in R. Journal of Statistical Software **105**:1 - 29.

Elias, C. L., A. R. Calapez, S. F. P. Almeida, B. Chessman, N. Simões, and M. J. Feio. 2016. Predicting reference conditions for river bioassessment by incorporating boosted trees in the environmental filters method. Ecological Indicators **69**:239-251.

Elith, J., and J. R. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics **40**:677-697.

Fergus, C. E., J. R. Brooks, P. R. Kaufmann, A. T. Herlihy, R. A. Hill, R. M. Mitchell, and P. Ringold. 2023. Disentangling natural and anthropogenic effects on benthic macroinvertebrate assemblages in western US streams. Ecosphere **14**:e4688.

Franklin, J. 2010. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press, Cambridge.

Gebert, F., M. K. Obrist, R. Siber, F. Altermatt, K. Bollmann, and N. Schuwirth. 2022. Recent trends in stream macroinvertebrates: warm-adapted and pesticide-tolerant taxa increase in richness. Biology Letters **18**:20210513.

Griffith, M. B. 2023. Development of national stressor-specific genus sensitivity values. Science of The Total Environment **895**:165121.

~~Hallstan, S., R. K. Johnson, E. Willén, and U. Grandin. 2012. Comparison of classification-then-modelling and species-by-species modelling for predicting lake phytoplankton assemblages. Ecological Modelling **231**:11-19.~~

Guisan, A., W. Thuiller, and N. E. Zimmermann. 2017. Habitat Suitability and Distribution Models: With Applications in R. Cambridge University Press, Cambridge.

Hawkins, C. P. 2006. QUANTIFYING BIOLOGICAL INTEGRITY BY TAXONOMIC COMPLETENESS: ITS UTILITY IN REGIONAL AND GLOBAL ASSESSMENTS. Ecological Applications **16**:1277-1294.

Hawkins, C. P., R. H. Norris, J. N. Hogue, and J. W. Feminella. 2000. DEVELOPMENT AND EVALUATION OF PREDICTIVE MODELS FOR MEASURING THE BIOLOGICAL INTEGRITY OF STREAMS. Ecological Applications **10**:1456-1477.

Hawkins, C. P., J. R. Olson, and R. A. Hill. 2010. The reference condition: predicting benchmarks for ecological and water-quality assessments. Journal of the North American Benthological Society **29**:312-343.

~~Hering, D., A. Borja, J. Carstensen, L. Carvalho, M. Elliott, C. K. Feld, A.-S. Heiskanen, R. K. Johnson, J. Moe, D. Pont, A. L. Solheim, and W. v. de Bund. 2010. The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. Science of The Total Environment **408**:4007-4019.~~

Heino, J., D. Schmera, and T. Erős. 2013. A macroecological perspective of trait patterns in stream communities. Freshwater Biology **58**:1539-1555.

Herlihy, A. T., P. R. Kaufmann, and M. E. Mitch~~, and D. D. Brown. 1990. Regional estimates of acid mine drainage impact on streams in the mid-atlantic and Southeastern~~. 1991. Stream chemistry in the eastern United States~~. Water, Air, and Soil Pollution **50**:91-107~~: 2. Current sources of acidity in acidic and low acid-neutralizing capacity streams. Water Resources Research **27**:629-642.

Herlihy, A. T., S. G. Paulsen, J. V. Sickle, J. L. Stoddard, C. P. Hawkins, and L. L. Yuan. 2008. Striving for consistency in a national assessment: the challenges of applying a reference-condition approach at a continental scale. Journal of the North American Benthological Society **27**:860-877.

Herlihy, A. T., and J. C. Sifneos. 2008. Developing nutrient criteria and classification schemes for wadeable streams in the conterminous US. Journal of the North American Benthological Society **27**:932-948.

Herlihy, A. T., J. L. Stoddard, and C. B. Johnson. 1998. The Relationship Between Stream Chemistry and Watershed Land Cover Data in the Mid-Atlantic Region, U.S. Water, Air, and Soil Pollution **105**:377-386.

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. JAWRA Journal of the American Water Resources Association **52**:120-128.

Hughes, R. M., and D. V. Peck. 2008. Acquiring data for large aquatic resource surveys: the art of compromise among science, logistics, and reality. Journal of the North American Benthological Society **27**:837-859.

~~Humphries, P., and K. O. Winemiller. 2009. Historical Impacts on River Fauna, Shifting Baselines, and Challenges for Restoration. BioScience **59**:673-684.~~

~~Jackson, S., and W. Davis. 1994. Meeting the Goal of Biological Integrity in Water-Resource Programs in the US Environmental Protection Agency. Journal of the North American Benthological Society **13**:592-597.~~

Jähnig, S. C., V. Baranov, F. Altermatt, P. Cranston, M. Friedrichs-Manthey, J. Geist, F. He, J. Heino, D. Hering, F. Hölker, J. Jourdan, G. Kalinkat, J. Kiesel, F. Leese, A. Maasri, M. T. Monaghan, R. B. Schäfer, K. Tockner, J. D. Tonkin, and S. Domisch. 2021. Revisiting global trends in freshwater insect biodiversity. WIREs Water **8**:e1506.

Jenkins, C. N., K. S. Van Houtan, S. L. Pimm, and J. O. Sexton. 2015. US protected lands mismatch biodiversity priorities. Proceedings of the National Academy of Sciences **112**:5081-5086.

Johnson, A. A., M. Q. Ott, and M. Dogucu. 2022. Bayes Rules!: An Introduction to Applied Bayesian Modeling Chapman and Hall/CRC.

Kaufmann, P. R., R. M. Hughes, S. G. Paulsen, D. V. Peck, C. W. Seeliger, T. Kincaid, and R. M. Mitchell. 2022a. Physical habitat in conterminous US streams and Rivers, part 2: A quantitative assessment of habitat condition. Ecological Indicators **141**:109047.

Kaufmann, P. R., R. M. Hughes, S. G. Paulsen, D. V. Peck, C. W. Seeliger, M. H. Weber, and R. M. Mitchell. ~~2022~~2022b. Physical habitat in conterminous US streams and rivers, Part 1: Geoclimatic controls and anthropogenic alteration. Ecological Indicators **141**:109046.

Kaufmann, P. R., P. Levine, E.G. Robison, C. Seeliger, and D.V. Peck. 1999. Quantifying Physical Habitat in Wadeable Streams. EPA/620/R-99/003, U.S. Environmental Protection Agency, Washington, D.C.

Kaushal, S. S., G. E. Likens, P. M. Mayer, R. R. Shatkay, S. A. Shelton, S. B. Grant, R. M. Utz, A. M. Yaculak, C. M. Maas, J. E. Reimer, S. V. Bhide, J. T. Malin, and M. A. Rippy. 2023. The anthropogenic salt cycle. Nature Reviews Earth & Environment **4**:770-784.

Kilgour, B., and L. Stanfield. 2005. A new approach for defining reference conditions in streams. AGU Spring Meeting Abstracts.

Kopp, D. A., J. L. Stoddard, R. A. Hill, J. M. Doyle, P. R. Kaufmann, A. T. Herlihy, and S. G. Paulsen. 2023. Joint species distribution models reveal taxon-specific sensitivities to potential anthropogenic alteration. Freshwater Science **42**:268-284.

~~Lewis, S. L., and M. A. Maslin. 2015. Defining the Anthropocene. Nature **519**:171-180.~~

Lin, J., J. E. Compton, R. A. Hill, A. T. Herlihy, R. D. Sabo, J. R. Brooks, M. Weber, B. Pickard, S. G. Paulsen, and J. L. Stoddard. 2021. Context is Everything: Interacting Inputs and Landscape Characteristics Control Stream Nitrogen. Environmental Science & Technology **55**:7890-7899.

Linke, S., R. H. Norris, D. P. Faith, and D. Stockwell. 2005. ANNA: A new prediction method for bioassessment programs. Freshwater Biology **50**:147-158.

~~Matthews, W. J. 1988. North American Prairie Streams as Systems for Ecological Study. Journal of the North American Benthological Society **7**:387-409.~~

McNellie, M. J., I. Oliver, J. Dorrough, S. Ferrier, G. Newell, and P. Gibbons. 2020. Reference state and benchmark concepts for better biodiversity conservation in contemporary ecosystems. Global Change Biology **26**:6702-6714.

Meinshausen, N. 2017. quantregForest: Quantile Regression Forests. R package version 1.3-7.

Meyer, H., and E. Pebesma. 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution **12**:1620-1633.

Mitchell, R. M., A. T. Herlihy, and R. M. Hughes. 2025. Striving for consistency in a national lake assessment: Defining reference status and littoral macroinvertebrate condition in lakes across the conterminous United States. Ecological Indicators **170**:112992.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. WHAT IS THE OBSERVED

RELATIONSHIP BETWEEN SPECIES RICHNESS AND PRODUCTIVITY? Ecology **82**:2381-2396.

Morse, C. C., A. D. Huryn, and C. Cronan. 2003. Impervious Surface Area as a Predictor of the Effects of Urbanization on Stream Insect Communities in Maine, U.S.A. Environmental Monitoring and Assessment **89**:95-127.

~~Nessel, M. P., T. Konnovitch, G. Q. Romero, and A. L. González. 2023. Decline of insects and arachnids driven by nutrient enrichment: A meta-analysis. Ecology **104**:e3897.~~

O'brien, R. M. 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity **41**:673-690.

Olimb, S. K., and B. Robinson. 2019. Grass to grain: Probabilistic modeling of agricultural conversion in the North American Great Plains. Ecological Indicators **102**:237-245.

Olsen, A. R., and D. V. Peck. 2008. Survey design and extent estimates for the Wadeable Streams Assessment. Journal of the North American Benthological Society **27**:822-836.

Olson, J. R., and S. M. Cormier. 2019. Modeling Spatial and Temporal Variation in Natural Background Specific Conductivity. Environmental Science & Technology **53**:4316-4325.

Olson, J. R., and C. Hawkins, P. . 2013. Developing site-specific nutrient criteria from empirical models. Freshwater Science **32**:719-740.

Olson, J. R., and C. P. Hawkins. 2012. Predicting natural base-flow stream water chemistry in the western United States. Water Resources Research **48**.

Outhwaite, C. L., P. McCann, and T. Newbold. 2022. Agriculture and climate change are reshaping insect biodiversity worldwide. Nature **605**:97-102.

Ovaskainen, O., and N. Abrego. 2020. Joint Species Distribution Modelling: With Applications in R. Cambridge University Press, Cambridge.

Ovaskainen, O., D. B. Roy, R. Fox, and B. J. Anderson. 2016. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. Methods in Ecology and Evolution **7**:428-436.

Ovaskainen, O., G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters **20**:561-576.

Paulsen, S. G., A. Mayio, D. V. Peck, J. L. Stoddard, E. Tarquinio, S. M. Holdsworth, J. V. Sickle, L. L. Yuan, C. P. Hawkins, A. T. Herlihy, P. R. Kaufmann, M. T. Barbour, D. P. Larsen, and A. R. Olsen. 2008. Condition of stream ecosystems in the US: an overview of the first national assessment. Journal of the North American Benthological Society **27**:812-821.

Peters, J., B. D. Baets, N. E. C. Verhoest, R. Samson, S. Degroeve, P. D. Becker, and W. Huybrechts. 2007. Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling **207**:304-318.

Powell, K. E., T. H. Oliver, T. Johns, M. González-Suárez, J. England, and D. B. Roy. 2023. Abundance trends for river macroinvertebrates vary across taxa, trophic group and river typology. Glob Chang Biol **29**:1282-1295.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems **9**:181-199.

Reynoldson, T. B., R. H. Norris, V. H. Resh, K. E. Day, and D. M. Rosenberg. 1997. The Reference Condition: A Comparison of Multimetric and Multivariate Approaches to

Assess Water-Quality Impairment Using Benthic Macroinvertebrates. Journal of the North American Benthological Society **16**:833-852.

Rumschlag, S. L., M. B. Mahon, D. K. Jones, W. Battaglin, J. Behrens, E. S. Bernhardt, P. Bradley, E. Brown, F. De Laender, R. Hill, S. Kunz, S. Lee, E. Rosi, R. Schäfer, T. S. Schmidt, M. Simonin, K. Smalling, K. Voss, and J. R. Rohr. 2023. Density declines, richness increases, and composition shifts in stream macroinvertebrates. Science Advances **9**:eadf4896.

Sabo, R. D., C. M. Clark, J. Bash, D. Sobota, E. Cooter, J. P. Dobrowolski, B. Z. Houlton, A. Rea, D. Schwede, S. L. Morford, and J. E. Compton. 2019. Decadal Shift in Nitrogen Inputs and Fluxes Across the Contiguous United States: 2002–2012. Journal of Geophysical Research: Biogeosciences **124**:3104-3124.

Sabo, R. D., C. M. Clark, D. A. Gibbs, G. S. Metson, M. J. Todd, S. D. LeDuc, D. Greiner, M. M. Fry, R. Polinsky, Q. Yang, H. Tian, and J. E. Compton. 2021. Phosphorus Inventory for the Conterminous United States (2002–2012). Journal of Geophysical Research: Biogeosciences **126**:e2020JG005684.

Sabo, R. D., B. Pickard, J. Lin, B. Washington, C. M. Clark, J. E. Compton, M. Pennino, B. Bierwagen, S. D. LeDuc, J. N. Carleton, M. Weber, M. Fry, R. Hill, S. Paulsen, A. Herlihy, and J. L. Stoddard. 2023. Comparing Drivers of Spatial Variability in U.S. Lake and Stream Phosphorus Concentrations. Journal of Geophysical Research: Biogeosciences **128**:e2022JG007227.

Samson, F., and F. Knopf. 1994. Prairie conservation in North America. BioScience **44**:418-421.

Sánchez-Bayo, F., and K. A. G. Wyckhuys. 2019. Worldwide decline of the entomofauna: A review of its drivers. Biological Conservation **232**:8-27.

Scherrer, D., H. K. Mod, and A. Guisan. 2020. How to evaluate community predictions without thresholding? Methods in Ecology and Evolution **11**:51-63.

~~Soga, M., and K. J. Gaston. 2018. Shifting baseline syndrome: causes, consequences, and implications. Frontiers in Ecology and the Environment **16**:222-230.~~

Smucker, N. J., E. M. Pilgrim, C. T. Nietch, L. Gains-Germain, C. Carpenter, J. A. Darling, L. L. Yuan, R. M. Mitchell, and A. I. Pollard. 2024. Using DNA metabarcoding to characterize national scale diatom-environment relationships and to develop indicators in streams and rivers of the United States. Science of The Total Environment **939**:173502.

Soranno, P. A., T. Wagner, S. L. Martin, C. McLean, L. N. Novitski, C. D. Provence, and A. R. Rober. 2011. Quantifying regional reference conditions for freshwater ecosystem management: A comparison of approaches and future research needs. Lake and Reservoir Management **27**:138-148.

Spake, R., M. P. Barajas-Barbosa, S. A. Blowes, D. E. Bowler, C. T. Callaghan, M. Garbowski, S. D. Jurburg, R. van Klink, L. Korell, E. Ladouceur, R. Rozzi, D. S. Viana, W.-B. Xu, and J. M. Chase. 2022. Detecting Thresholds of Ecological Change in the Anthropocene. Annual Review of Environment and Resources **47**:797-821.

Stoddard, J. L., A. T. Herlihy, D. V. Peck, R. M. Hughes, T. R. Whittier, and E. Tarquinio. 2008. A process for creating multimetric indices for large-scale aquatic surveys. Journal of the North American Benthological Society **27**:878-891.

Stoddard, J. L., D. P. Larsen, C. P. Hawkins, R. K. Johnson, and R. H. Norris. 2006. SETTING EXPECTATIONS FOR THE ECOLOGICAL CONDITION OF STREAMS: THE CONCEPT OF REFERENCE CONDITION. Ecological Applications **16**:1267-1276.

Tang, T., R. J. Stevenson, and J. B. Grace. 2020. The importance of natural versus human factors for ecological conditions of streams and rivers. Science of The Total Environment **704**:135268.

Tikhonov, G., Ø. H. Opedal, N. Abrego, A. Lehikoinen, M. M. J. de Jonge, J. Oksanen, and O. Ovaskainen. 2020. Joint species distribution modelling with the r-package Hmsc. Methods in Ecology and Evolution **11**:442-447.

Tonkin, J. D., R. G. Death, and K. J. Collier. 2013. Do productivity and disturbance interact to modulate macroinvertebrate diversity in streams? Hydrobiologia **701**:159-172.

USEPA. 2017a. National Rivers and Streams Assessment 2018-19: Field Operations Manual–Wadeable. EPA-841-B-17-003a. U.S. Environmental Protection Agency, Office of Water, Washington, DC. .

USEPA. 2017b. National Rivers and Streams Assessment 2018-19: Laboratory Operations Manual. EPA841-B-17-004. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA. 2023. National Rivers and Streams Assessment 2018-2019 Technical Support Document. EPA 841-R-22-005. U.S. Environmental Protection Agency**,** Office of Water and Office of Research and Development, Washington, D.C. .

USEPA. ~~2024~~2024a. Climate change indicators in the Unites States. EPA 430-R-24-003.

USEPA. 2024b. National Rivers and Streams Assessment: The Third Collaborative Survey. U.S. Environmental Protection Agency, Office of Water and Office of Research and Development.

Van Sickle, J. 2008. An index of compositional dissimilarity between observed and expected assemblages. Journal of the North American Benthological Society **27**:227-235.

Vörösmarty, C. J., P. B. McIntyre, M. O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, S. E. Bunn, C. A. Sullivan, C. R. Liermann, and P. M. Davies. 2010. Global threats to human water security and river biodiversity. Nature **467**:555-561.

Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. 2015. So Many Variables: Joint Modeling in Community Ecology. Trends in Ecology & Evolution **30**:766-779.

Wiens, J. J., D. D. Ackerly, A. P. Allen, B. L. Anacker, L. B. Buckley, H. V. Cornell, E. I. Damschen, T. Jonathan Davies, J.-A. Grytnes, S. P. Harrison, B. A. Hawkins, R. D. Holt, C. M. McCain, and P. R. Stephens. 2010. Niche conservatism as an emerging principle in ecology and conservation biology. Ecology Letters **13**:1310-1324.

Wilkinson, D. P., N. Golding, G. Guillera-Arroita, R. Tingley, and M. A. McCarthy. 2021. Defining and evaluating predictions of joint species distribution models. Methods in Ecology and Evolution **12**:394-404.

~~Yuan, L. L. 2006. Theoretical predictions of observed to expected ratios in RIVPACS-type predictive model assessments of stream biological condition. Journal of the North American Benthological Society **25**:841-850.~~

Wood, G. 2008. Thresholds and criteria for evaluating and communicating impact significance in environmental statements: 'See no evil, hear no evil, speak no evil'? Environmental Impact Assessment Review **28**:22-38.

Yuan, L. L., D. M. Carlisle, R. M. Mitchell, and A. I. Pollard. 2024. Using models of local environmental conditions for biological assessment. Freshwater Science:000-000.

Yuan, L. L., C. P. Hawkins, and J. V. Sickle. 2008. Effects of regionalization decisions on an O/E index for the US national assessment. Journal of the North American Benthological Society **27**:892-905.

Zak, D., M. Hupfer, A. Cabezas, G. Jurasinski, J. Audet, A. Kleeberg, R. McInnes, S. M. Kristiansen, R. J. Petersen, H. Liu, and T. Goldhammer. 2021. Sulphate in freshwater ecosystems: A review of sources, biogeochemical cycles, ecotoxicological effects and bioremediation. Earth-Science Reviews **212**:103446.

~~Supplementary Information~~

~~A model-based assessment of anthropogenic disturbance on lotic macroinvertebrate assemblages~~

~~Darin A. Kopp[1,2*], John L. Stoddard[2], Philip R. Kaufmann[2,3], Alan T. Herlihy[3], Ryan Hill[2], Meredith Brehob[5], Robert Sabo[6]~~

~~[1]Oak Ridge Institute for Science and Education Participant c/o US EPA, ORD, Corvallis, Oregon~~

~~[2]US EPA, Office of Research and Development, Corvallis, Oregon~~

~~[3]Department of Fisheries, Wildlife, & Conservation Sciences, Oregon State University, Corvallis, Oregon~~

~~[4]Current affiliation: Oregon Department of Fish and Wildlife, Salem, Oregon~~

~~[5]Oak Ridge Institute for Science and Education Participant c/o US EPA, ORD, Washington, D.C.~~

~~[6]US EPA, Office of Research and Development, Center for Public Health and Environmental Assessment, DC, Washington, USA~~

~~*Correspondence: kopp.darin@epa.gov~~

~~Table S1: Geoclimatic and anthropogenic variables used in random forest models. Variables were selected based on their hypothesized relationship with total nitrogen, total phosphorus, chloride, sulfate and substrate diameter. StreamCat data is publicly accessible at https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset (Hill et al. 2016)~~

~~Variable~~	~~Description~~	~~Source~~
~~BFIWs~~	~~The component of streamflow that can be attributed to ground-water discharge~~	~~StreamCat, Hill et al. 2016~~
~~ClayWs~~	~~Mean % clay content of soils in watershed~~	~~StreamCat~~
~~ElevWs~~	~~Mean elevation of watershed (m)~~	~~StreamCat~~
~~KffactWs~~	~~Mean of STATSGO Kffactor~~	~~StreamCat~~
~~DepClWs~~	~~Mean atmospheric chloride deposition 2018~~	~~National Atmospheric Deposition Program~~
~~DepPWs~~	~~Atmospheric phosphorous deposition 2013~~	~~Sabo et al 2023~~
~~NWs~~	~~Rock derived nitrogen in watershed~~	~~StreamCat~~
~~P2O5Ws~~	~~Mean % of lithological phosphorous oxide (P2O5) content in surface or near surface geology~~	~~StreamCat~~
~~PctAlluvCoastWs~~	~~% of watershed area classified with lithology type as alluvium and fine-textured coastal zone sediment~~	~~StreamCat~~
~~PctHbWetWs~~	~~% of watershed area classified as herbaceous wetland land cover~~	~~StreamCat~~
~~PermWs~~	~~Mean permeability (cm/hour) of soils (STATSGO) within watershed~~	~~StreamCat~~
~~RunoffWs~~	~~Mean runoff (mm) within watershed~~	~~StreamCat~~
~~SandWs~~	~~Mean % sand content of soils (STATSGO) within watershed~~	~~StreamCat~~
~~StreamPower~~	~~The amount of energy the water exerts on the sides and bottom of a stream~~	~~(((TPRCP - ET)/1000)*(0.032*WSArea)^0.5) * slope~~
~~SWs~~	~~Mean % of lithological sulfur (S) content in surface or near surface geology within watershed~~	~~StreamCat~~
~~WsAreaSqKm~~	~~Area of watershed (square km)~~	~~StreamCat~~
~~CoalMineDensWs~~	~~Density of coal mines sites within watershed (mines/square km)~~	~~StreamCat~~
~~DepNWs~~	~~Mean atmospheric nitrogen deposition in watershed 2018~~	~~National Atmospheric Deposition Program~~
~~DepSWs~~	~~Mean atmospheric sulfur deposition in watershed 2018~~	~~National Atmospheric Deposition Program~~
~~MineDensWs~~	~~Density of coal mines sites within watershed (mines/square km)~~	~~StreamCat~~
~~N_inputs~~	~~Sum of anthropogenic nitrogen inputs: N_Fert_FarmWs, N_Fert_UrbanWs, N_Human_WasteWs, N_Livestock_WasteWs~~	~~Sabo et al. 2023~~
~~P_inputs~~	~~Sum of anthropogenic phosphorus inputs: P_f_fertilizerWs, P_human_wasteWs, P_livestock_WasteWs, P_nf_fertilizerWs~~	~~Sabo et al. 2023~~
~~PctCropWs~~	~~% of watershed area classified as crop land use~~	~~StreamCat~~
~~PctCropWsRp100~~	~~% of watershed riparian area classified as crop land use~~	~~StreamCat~~
~~PctNatWs~~	~~% of watershed with natural vegetation cover~~	~~StreamCat~~
~~TPRCP~~	~~Mean total annual precipitation in watershed~~	~~PRISM~~
~~RdDensWs~~	~~Density of roads (2010 Census Tiger Lines) within watershed (km/square km)~~	~~StreamCat~~
~~MSAT~~	~~Mean summer (July/August) air temperature~~	~~PRISM~~
~~W1_HAG~~	~~Agricultural disturbance adjacent to a stream~~	~~NRSA Field Data~~
~~NABD_NrmStorWs~~	~~Volume all reservoirs per unit area of watershed (cubic meters/square km)~~	~~StreamCat~~

~~Table S2: Variable importance rankings for random forest models and values used to estimate physiochemical conditions in the absence of anthropogenic disturbance. The 5 most important variables for each model are bold.~~

Predictor Variable	NTL	PTL	CL	SO4	SUBD	Value Without Disturbance
Geoclimatic						
BFIWs	6	13	**5**	9	–	–
ClayWs	–	10	7	11	–	–
ElevWs	7	**5**	8	8	–	–
KffactWs	14	12	15	12	9	–
DepClWs	–	–	9	–	–	–
DepPWs	–	**4**	–	–	–	–
NWs	11		12	**4**	–	–
P2O5Ws	15	6	18	16	–	–
PctAlluvCoastWs	–	–	–	–	**5**	–
PctHbWetWs	9	7	10	17	–	–
PermWs	13	11	17	13	–	–
RunoffWs	**2**	**2**	**4**	**1**	**2**	–
SandWs	–	–	–	–	7	–
StreamPower	–	–	–	–	**1**	–
SWs	12	16	6	**2**	–	–
WsAreaSqKm	16	14	13	15	6	–
Anthropogenic						
CoalMineDensWs	–	–	19	7	13	0
DepNWs	**5**	–	–	–	–	5 kgN/ha/yr*
DepSWs	–	–	14	10	–	1.65 kgS/ha/yr*
MineDensWs	–	–	16	18	11	0
N_inputs	**4**	–	–	–	–	0
P_inputs	–	9	–	–	–	0
PctCropWs	**1**	**1**	**3**	**3**	–	0
PctCropWsRp100	–	–	–	–	**4**	0
PctNatWs	–	–	–	–	**3**	100
TPRCP	10	15	11	14	–	Average 1940-1950
RdDensWs	**3**	8	**1**	6	10	0
MSAT	8	**3**	**2**	**5**	–	Average 1940-1950
W1_HAG	–	–	–	–	8	0
NABD_NrmStorWs	–	–	–	–	12	0

~~* Estimates for atmospheric nitrogen and sulfur deposition were obtained from Clark et al. 2018. Although they used 0.4 kgN/ha and 0.1kgS/ha no sites had deposition values below this level. Instead, we selected the higher values of 3-5KgN/ha which could be reasonable estimates for background deposition values before 1900. For S deposition, other efforts have estimated pre-industrial S deposition at 0.32–2.98 kgS/ha (Granat et al. 1976, Fakhraei et al. 2016) and we used the used middle number of this interval.~~

~~Table S3. Regional means of present-day (PD) and hindcasted (HC) values for each environmental gradient. Parentheses are 10th and 90th quantiles of observed hindcasted values. NTL = Total Nitrogen, PTL = Total Phosphorus, CL = Chloride, SO4 = Sulfate, SUBD = Substrate Diameter, RPDI = Riparian Disturbance index, MSAT = Mean Summer Air Temperature, PRCP = Total Annual Precipitation.~~

	NTL (ug/L)		PTL (ug/L)		CL (mg/L)		SO4 (mg/L)		SUB (mm)		RPDI		MSAT (DegC)		TPRCP (mm)	
	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC
CPL	780 (305-2452.2)	363 (289-558)	82.6 (23.08-245.52)	46.74 (25.73-77.65)	9.89 (2.93-64.71)	4.06 (2.92-7.65)	7.49 (1.1-44.55)	3.56 (1.93-14.73)	0.35 (0.01-2.21)	6.17 (1.96-17.71)	0.5 (0.03-1.79)	0.33 (0.03-0.33)	27.46 (25.54-28.54)	27.02 (24.24-27.85)	1613.83 (1186.55-1989.95)	1219.48 (1014.36-1680.13)
NAP	431 (187.4-1309.4)	263 (214-301.6)	23.76 (7.38-75.51)	17.37 (9.52-29.31)	16.73 (0.47-72.34)	1.97 (0.3-5.76)	5.58 (1.71-16.05)	3.24 (1.74-6.24)	11.66 (0.08-121.32)	16.64 (6.1-52.76)	0.65 (0-2.02)	0.33 (0-0.33)	21.49 (19.01-23.67)	20.43 (18.31-21.78)	1295.41 (1104.73-1627.51)	1088.38 (946.14-1233.49)
NPL	879 (266-2188)	456 (202-687.5)	85.76 (18.4-361.74)	76.25 (33.96-108.45)	12.09 (2.35-50.42)	10.24 (1.53-18.33)	486.24 (19.03-1350.29)	324.9 (9.33-653.05)	0.51 (0.01-17.94)	3.44 (1.13-23.73)	1.27 (0.38-2.08)	0.33 (0.33-0.33)	20.52 (18.35-22.48)	21.19 (19.1-23.77)	437.42 (320.7-631.93)	360.52 (282.21-428.35)
SAP	557.5 (173.5-1868)	265 (169-346.5)	29.3 (9.84-128.09)	26.5 (14.84-45.6)	5.89 (1.35-47.42)	2.45 (1.04-3.98)	7.46 (1.55-80.35)	4.3 (1.89-7.34)	16.28 (0.35-206.31)	15.51 (7.79-51.83)	0.76 (0.05-2.29)	0.33 (0.05-0.33)	24.76 (22.1-26.72)	24.07 (21.37-26.37)	1475.78 (1106.55-1809.01)	1110.34 (916.58-1283.79)
SPL	1175 (424.8-3229.4)	411 (263.4-668.4)	150 (25.42-808.04)	57 (36.71-86.67)	30.36 (4.55-383.92)	10.27 (3.97-18.72)	88.89 (7.92-905.64)	66.72 (9.5-511.06)	0.35 (0.02-20.65)	4.8 (0.95-12.72)	1.04 (0.22-2.35)	0.33 (0.22-0.33)	25.88 (21.93-29.65)	26.16 (22.05-28.9)	616.17 (394.96-1060.16)	534.19 (313.6-973.8)
TPL	1816 (659-8927.6)	411 (329.1-558)	167 (53.96-444.72)	50 (42-68.09)	18.9 (7.24-59.44)	5.22 (3.43-8.97)	34.4 (12.05-270.07)	12.8 (5.57-164.48)	0.35 (0.01-15.97)	11.31 (3.82-18.96)	0.83 (0.12-2.28)	0.33 (0.12-0.33)	23.33 (20.96-25.27)	23.53 (21.24-25.41)	1004.22 (677.24-1387.75)	775.73 (514.38-922.32)
UMW	1168 (355-5250)	350 (265-418)	57.4 (17.1-165.47)	35.55 (16.51-56.07)	9.97 (0.71-29.93)	2.56 (0.34-4.41)	9.28 (1.11-41.51)	3.71 (1.57-8.29)	0.35 (0.01-12.33)	6.94 (4.59-28.61)	0.62 (0-2.14)	0.33 (0-0.33)	20.84 (18.14-22.56)	20.98 (18.42-22.39)	960.16 (788.56-1393.8)	793.53 (622.66-877.31)
WMT	133 (50-386.8)	133 (65-216)	21.84 (7.09-82.73)	21.14 (8.61-58.72)	0.92 (0.12-9.18)	0.52 (0.15-2.53)	2.97 (0.47-26.75)	2.87 (0.93-19.97)	38.31 (0.98-202.46)	34.89 (6.85-82.76)	0.48 (0-1.72)	0.33 (0-0.33)	18.08 (13.98-22.29)	17 (12.61-21.33)	669.31 (349.23-1707.15)	701.68 (387.1-1812.51)
XER	340 (110-1258.8)	218 (133-396.4)	53.87 (14.35-258.99)	45.76 (15.65-90.21)	5.35 (0.8-143.8)	2.9 (0.73-14.32)	27.8 (1.6-403.9)	22.32 (2.54-186.07)	2.02 (0.02-57.52)	12.71 (3.53-45.29)	1.16 (0.12-2.61)	0.33 (0.12-0.33)	22.76 (18.79-30.29)	21.36 (17.45-27.99)	287.58 (169.27-503.23)	336.66 (220.95-694.06)

Tables

Table 1: Sites, genera, and environmental variables included in the analysis. Revisit sites are locations that were revisited during the survey to assess within year variability. Values for the environmental variables are median values for each region. Range is given in parentheses. NTL = Total Nitrogen, PTL = Total Phosphorous, CL = Chloride, SO4 = Sulfate, RPDI = Riparian Disturbance Index, SUBD = Substrate Diameter, TPRCP = Total Precipitation and MSAT = Mean Summer Air Temperature. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric.

Region	Site (#)	Genera (#)	Revisit Sites (#)	NTL (ug/L)	PTL (ug/L)	CL (mg/L)	SO4 (mg/L)	RPDI	SUBD (mm)	TPRCP (mm)	MSAT (ºC)
CPL	226	71	36	786 (91-7713)	84.81 (5.2-3922.25)	10.06 (0.52-4797.9)	7.83 (0.11-1900.89)	0.5 (0-5.39)	0.35 (0.01-341.65)	1613.83 (388.66-2397.1)	27.46 (23.75-31.68)
NAP	228	127	29	431 (81-6413)	23.76 (3.12-587.2)	16.73 (0.07-668.16)	5.58 (0.06-467.93)	0.65 (0-5.24)	11.66 (0.01-864.9)	1295.41 (888.91-1883.17)	21.49 (17.24-24.72)
NPL	152	76	7	881.5 (71-15675)	85.61 (3.53-9248.31)	12.09 (0.11-1251.58)	486.24 (3.5-4079.78)	1.27 (0-4.49)	0.51 (0.01-560.5)	437.42 (157.84-1055.11)	20.52 (15.62-23.46)
SAP	266	111	32	557.5 (36-18700)	29.3 (3.79-4050)	5.89 (0.36-197.39)	7.46 (0.58-397.69)	0.76 (0-4.56)	16.28 (0.01-5656.85)	1475.78 (889.64-2451.81)	24.76 (19.19-27.73)
SPL	174	59	5	1165 (145-21175)	147.12 (5.21-4351.7)	30.36 (0.43-5220)	88.89 (1.96-3716.6)	1.04 (0-5.88)	0.35 (0.01-5656.85)	616.17 (240.1-1404.61)	25.88 (11.96-32.75)
TPL	223	74	29	1806 (236-16219)	165.08 (11.45-1066.87)	18.9 (1.44-736.62)	34.4 (5.79-1386.77)	0.83 (0-5.47)	0.35 (0.01-5656.85)	1004.22 (376.7-1801.08)	23.33 (18.28-27.09)
UMW	201	104	10	1168 (195-17675)	57.4 (8.16-659)	9.97 (0.01-306.69)	9.28 (0.04-160.4)	0.62 (0-6.43)	0.35 (0.01-812.79)	960.16 (496.44-1718.2)	20.84 (17.05-23.43)
WMT	225	94	19	133 (22-4719)	21.84 (2.71-569.8)	0.92 (0.04-521.64)	2.97 (0.07-1682.05)	0.48 (0-3.95)	38.14 (0.01-1288.61)	669.31 (175.89-3946.33)	18.08 (11.71-29.4)
XER	196	75	16	344.5 (46-8000)	54.62 (4.29-4667.41)	5.35 (0.1-1867.57)	27.8 (0.02-3286.24)	1.16 (0-4.36)	2.02 (0.01-368.11)	287.58 (100.46-1176.56)	22.76 (14.69-36.11)

Table 2: Random Forest model performance metrics for testing and out-of-bag training datasets. NTL = Total Nitrogen, PTL = Total Phosphorus, CL = Chloride, SO4 = Sulfate, SUBD = Substrate Diameter. RMSE = Root mean squared error of random forest models fitted with ln(x + 1) (NTL, PTL, and CL) or ln(x) (SO4) or Log10(SUBD) transformations.

Variable	R^2_{train}	$RMSE_{train}$	R^2_{test}	$RMSE_{test}$
ln(NTL+1) (ug/L)	0.73	0.61	0.72	0.60
ln(PTL+1) (ug/L)	0.60	0.78	0.69	0.70
ln(CL+1) (mg/L)	0.73	0.75	0.68	0.85
ln(SO4) (mg/L)	0.77	0.96	0.78	0.97
log_{10}(SUBD) (mm)	0.46	0.98	0.51	1.00

Table 3. Regression coefficients between predicted and observed richness and compositional similarity. Occurrence probabilities thresholds (Pr) were used to exclude genera with low predicted occurrence probabilities. We considered $R^2 \geq 0.2$, $-1.5 \leq$ intercept ≤ 1.5, and $0.85 \leq$ slope ≤ 1.15 are indicative of adequate model performance. Values in parentheses are the 5th and 95th percentile of estimates from the 3000 posterior samples. Compositional similarity was measured using a probabilistic adaptation of Jaccard index measured for each site. Parentheses are the 5th and 95th percentile of values for all sites.

		Genus Richness				Composition	
	Pr	Intercept	Slope	R^2	Revisit Sites within Posterior Distribution (%)	Mean Jaccard Similarity	Mean Jaccard Similarity at revisit sites
CPL	0.05	0.37 (-0.37 - 1.02)	1.00 (0.96 - 1.05)	0.85 (0.82 - 0.87)	97	0.52 (0.26 - 0.7)	0.68 (0.24-0.95)
NAP	0.05	-0.45 (-2.02 - 1.02)	1.03 (0.99 - 1.08)	0.74 (0.71 - 0.78)	90	0.52 (0.31 - 0.64)	0.35 (0.17-0.65)
NPL	0.05	-0.07 (-1.21 - 1.04)	1.02 (0.96 - 1.09)	0.76 (0.71 - 0.81)	100	0.58 (0.34 - 0.77)	0.39 (0.12-0.71)
SAP	0.05	0.40 (-0.66 - 1.41)	1.00 (0.97 - 1.04)	0.85 (0.82 - 0.87)	97	0.52 (0.26 - 0.67)	0.56 (0.21-0.87)
SPL	0.10	0.48 (-0.45 - 1.29)	1.02 (0.97 - 1.09)	0.79 (0.75 - 0.83)	100	0.52 (0.25 - 0.72)	0.61 (0.52-0.98)
TPL	0.05	0.06 (-0.91 - 0.94)	1.01 (0.97 - 1.06)	0.82 (0.79 - 0.84)	100	0.55 (0.31 - 0.77)	0.59 (0.18-1)
UMW	0.05	-0.74 (-2.52 - 0.99)	1.04 (0.98 - 1.10)	0.73 (0.66 - 0.78)	100	0.55 (0.34 - 0.69)	0.53 (0.27-0.9)
WMT	0.10	0.24 (-1.24 - 1.59)	1.04 (0.99 - 1.10)	0.74 (0.70 - 0.78)	100	0.60 (0.36 - 0.75)	0.45 (0.16-0.79)
XER	0.05	-0.12 (-1.11 - 0.74)	1.03 (0.98 - 1.08)	0.83 (0.80 - 0.86)	94	0.50 (0.23 - 0.71)	0.48 (0.07-0.83)

Table 4: Mean proportion of sites in each taxonomic group that was identified as an increaser or decreaser. Increasers (I) are genera that have a significantly higher probability of occurrence under present day conditions compared to hindcasted conditions; decreasers (D) are genera that have a significantly lower probability of occurrence under present-day conditions compared to hindcasted conditions. The values in the table are the mean proportion of sites for genera within the major taxonomic group. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = xeric. EPT = Ephemeroptera, Plecoptera and Trichoptera.

	CPL		NAP		NPL		SAP		SPL		TPL		UMW		WMT		XER	
	D	I	D	I	D	I	D	I	D	I	D	I	D	I	D	I	D	I
Insects	0.16	0.08	0.08	0.05	0.1	0.17	0.1	0.11	0.09	0.12	0.16	0.11	0.11	0.12	0.03	0.06	0.09	0.05
EPT	0.24	0.01	0.07	0.04	0.17	0.11	0.13	0.08	0.09	0.04	0.23	0.04	0.14	0.03	0.04	0.06	0.1	0.03
CHIRONOMIDAE	0.11	0.09	0.09	0.05	0.04	0.18	0.04	0.12	0.02	0.17	0.12	0.14	0.06	0.16	0.02	0.07	0.06	0.1
Other Insects	0.21	0.07	0.08	0.08	0.07	0.21	0.13	0.12	0.23	0.02	0.18	0.11	0.09	0.04	0.02	0.03	0.11	0.01
Non-Insects	0.14	0.16	0.06	0.18	0.06	0.19	0.11	0.13	0.06	0.13	0.15	0.15	0.08	0.14	0.03	0.08	0.04	0.14
ARTHROPODA	0.21	0.11	0.07	0.16	0.11	0.02	0.16	0.09	0.05	0.06	0.23	0.35	0.1	0.04	0.04	0.03	0.05	0.04
MOLLUSCA	0.08	0.11	0.03	0.28	0.02	0.3	0.1	0.15	0.09	0.06	0.21	0.05	0.14	0.07	0.01	0.06	0.01	0.26
Other Non-Insects	0.15	0.22	--	0.14	0.05	0.13	0.04	0.16	0.02	0.19	0.05	0.15	0.01	0.25	0.04	0.14	0.04	0.14

Figure Captions

Figure 1: Ecoregions and survey locations for the National Rivers and Streams Assessment 2018-2019 survey. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric

Figure 2: Conceptual diagram of posterior distributions generated from bayesian inference. The grey is a posterior distribution of genus richness generated after removing anthropogenic disturbances (Hindcasted) and the white is a posterior distribution generated from present-day conditions. The dotted line is the expected value (i.e. mean) of the present day posterior distribution and dark shading represents the lower 10% or the upper 90% of the HC posterior distribution. At site A, the PD mean is within the <10% of the HC distribution and indicates that PD richness is likely lower than HC richness. Alternatively, at site B, PD richness is >90% of the HC distribution and indicates that PD richness is likely higher. At site C, the two posterior distributions are similar such that these is likely no difference between PD and HC.

Figure 3: Partial dependence plots showing the effects of anthropogenic variables on in-stream physiochemical factors. For visualization, each anthropogenic factor was rescaled between 0-1 and labeled low, medium, and high. CoalMineDen = coal mine density, MineDen = gravel mine density, MSAT = mean summer air temperature, N_dep = atmospheric nitrogen deposition, N_input = anthropogenic nitrogen inputs, P_input = anthropogenic phosphorous inputs, PctCrop = percent crop in the watershed, PctCropRP = percent crop in the riparian area, PctNatRP = percent natural vegetation in riparian area, RdDen = road density, S_dep = atmospheric sulfur deposition, TPRCP = total precipitation, and W1_HAG = agricultural disturbance adjacent to stream reach. NTL = Total Nitrogen (ug/L), PTL = Total Phosphorus (ug/L), CL = Chloride (mg/L), SO4 = Sulfate (mg/L), SUBD = Substrate Diameter Log10(mm).

Figure 4: Observed versus hindcasted values for each environmental gradient. Points are regional means and vertical and horizontal bars represent the 10^{th} and 90^{th} quantiles of observed values or hindcasted values within each region, respectively. For RPDI, all hindcast values were < 0.33 and the ecoregions were plotted separately for visualization. The dashed line is the 1:1 relationship for all plots except for RPDI where it represents the 0.33 threshold applied to all ecoregions. NTL = Total Nitrogen (ug/L), PTL = Total Phosphorus (ug/L), CL = Chloride (mg/L), SO4 = Sulfate (mg/L), SUBD = Substrate Diameter (mm), RPDI = Riparian Disturbance Index, MSAT = mean summer air temperature (ºC) and TPRCP = total precipitation (mm). CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric. See Appendix S1: Table S4 for observed and hindcasted mean values and quantiles for each region.

Figure 5: A) Proportion of sites where hindcasted abiotic conditions were >2 standard deviations from the observed value for Total Nitrogen (NTL), Total Phosphorus (PTL), Chloride (CL), Sulfate (SO4), Substrate Diameter (SUBD), mean summer air temperature (MSAT), and Total Precipitation (TPRCP) and > 0.33 for Riparian Disturbance Index (RPDI). B) The number of environmental variables affected by human disturbance. Locations where the effects were not detected are plotted separately. CPL = Coastal Plains, NAP = Northern Appalachians, NPL = Northern Plains, SAP = Southern Appalachians, SPL = Southern Plains, TPL = Temperate Plains, UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric.

Figure 6: Difference in macroinvertebrate genus richness after removing anthropogenic disturbance from each category of physiochemical variables (A) and all variables simultaneously (B). The points in each panel are sites that had a change in genus richness with >75% support after removing disturbance.

Figure 7: Percentage of the population of streams where genus richness or composition could change given a hindcasted physiochemical environment. Shaded bars indicate that the probability of mean present-day richness differs from hindcast with > 0.90 (Black) or 0.75-0.90 (Gray) support. Hatched bars indicate that Jaccard similarity was < 0.9 but support for a difference in richness were < 0.75. White bars indicate the proportion of sites that had < 0.75 support for change in richness and >0.9 compositional similarity. Yellow bars indicate the proportion of sites that could not be assessed because of insufficient data or potential extrapolation from predicting hindcast physiochemical conditions. Regions are arranged according to richness. NPL = Northern Plains, TPL = Temperate Plains, SPL = Southern Plains, XER = Xeric, SAP = Southern Appalachians, CPL = Coastal Plains, NAP = Northern Appalachians, UMW = Upper Midwest and WMT = Western Mountains.