EAP24-0641 - "A model-based assessment of anthropogenic disturbance on lotic macroinvertebrate assemblages"

Darin A. Kopp1\*, John L. Stoddard1, Philip R. Kaufmann1,2, Alan T. Herlihy2, Ryan A. Hill1, Meredith M. Brehob3, Robert D Sabo4

**RESPONSE TO EDITOR**

**COMMENT:**

The database presented is intriguing, and the potential applications are relevant. However, it is essential to conduct a thorough review of the document to clarify the novelty and contributions of the methodology developed by the authors. Additionally, further analyses are required to fully utilize their data. It is necessary to test the effectiveness of their model by validating it, as their methods involve extrapolation without proper validation. Furthermore, while the authors propose their method as an alternative to existing approaches, direct comparisons with traditional methods are needed to demonstrate the effectiveness of their approach.

RESPONSE:

Dear, Dr. López-Lozano,

We are pleased to submit a revised version of our manuscript. In this version, we extensively revised the introduction and discussion sections to clarify the novelty and contributions of our approach. We also preformed further analysis to validate our joint species distribution models and to account for potential extrapolation of hindcast predictions using random forest models. Additionally, we place our work in the context of traditional methods and provide a direct comparison to results reported by the National Rivers and Streams Assessment.

The comments provided by both reviewers we helpful for improving our manuscript and are addressed below. We also formatted this version to be consistent with the style of the journal.

Please feel free to contact me with any questions regarding this submission.

Sincerely,

Darin Kopp

**RESPONSE TO REVIEWER #1**

**GENERAL COMMENTS:**

We thank reviewer 1 for their comments. They have improved the quality of our manuscript. We have extensively revised the previous version of the manuscript to incorporate the major and minor comments. In addition, we provide text with referenced line numbers in the revised manuscript for each comment below.

**COMMENT:**

1) Model Extrapolation Without Validation. The authors extrapolate their models to environmental settings with no human disturbances in Steps 2 and 4 described above. Extrapolating a model beyond the range of predictors in the training dataset is inherently risky, particularly with complex algorithms like random forests (RF) and joint species models (JSM). For example, RF predictions are constrained by the extreme values of the response variable in the training dataset, a common limitation of tree-based models. As a result, the six site variables in Step 2 could be underestimated or overestimated, introducing bias into Step 4. Model validation is crucial, especially in extrapolation scenarios. The authors might consider reserving all high-quality, yet not pristine, sites as a test set. If the model demonstrates strong predictive accuracy at these sites, its reliability under zero human disturbance conditions would be more credible. A similar validation could be conducted to the JSM at Step 4.

**RESPONSE:**

Thank you for the comment. The random forest models were fit using the entire national dataset to include the full gradient of anthropogenic predictor variables and validated by withholding a 20% subset of the data (L291-294). Nonetheless, we recognize that changing anthropogenic predictor variables could generate data that were dissimilar to the present-day data used to fit the model. Accordingly, we tested for extrapolation (L310 - 319) following Meyer and Pebesma (2021) and Yuan et al. (2024). Using this approach we flagged 77 sites (~4%) for potential extrapolation (L477-486, Appendix S1: Figure S1) and discuss the implications (L596-612).

Alternatively, the JSDM models were fit at a regional scale and therefore potentially sensitive to extrapolation if hindcast physiochemical variables were outside the range of data used to fit the model. We detected 22 instances (~1%) where the hindcasted values could lead to potential extrapolation of the JSDM (Lines 509-514). Given the rarity of these instances, it is unlikely that this would have large effect on our previous results. Nonetheless, in the revised version, we set the hindcasted values that were beyond the range of present-day conditions to the minimum of the present-day (L512-514).

COMMENT:

2) Environmental Predictor Selection and accuracy in JSM: The authors modeled genus distribution using eight predictors. The selection of some of some predictors for the JSM need to be better justified. Precipitation and temperature are typically key drivers of species distribution, as are stream size and slope. Instead, the authors prioritized six disturbance-related variables. Despite this, the JSM demonstrated notable accuracy in genus richness estimation (after excluding many rare genera), outperforming most models that stack individual species distribution models to estimate species richness. One possible explanation is that the six disturbance-related variables act as surrogates or proxies for other key drivers of species occurrence, as evidenced in Stage-1 modeling. While this relationship may be sufficient for interpolation, it could pose challenges for extrapolation, as the relationships may not hold beyond the training dataset. To address this, some form of validation is necessary (refer to Major Comments-1). In addition, the reported accuracy of genus richness estimation under current environmental conditions, does not appear based on cross-validation or a test dataset. If so, it may be inflated.

RESPONSE:

We appreciate the comment and have revised text for clarification. In brief, the random effects structure of JSDM permitted us to estimate the effects of a relatively small number variables while statistically controlling for other factors (L99-104). Although sample level random effects would be completely confounded with residual error in univariate analyses, they are meaningful in multivariate contexts (L215-226). One potential limitation with this model structure is that it limits our ability to make accurate predictions to new locations units (Abrego and Ovaskainen 2023). Importantly, predicting to new locations was not our objective and we devote several lines of text to this in the discussion (L636-651).

Although predicting to new locations could be limited, a small number of the locations used in our analysis were revisited as part of NRSA (L163-165) and were used to for additional validation of our models (L262-276). 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) we expected variation between samples (L266-272). Thus, we assessed the proportion of values obtained from revisited that were contained by the modeled posterior distribution (L272-274). We found that the values for >90% of values for all regions were within the posterior distribution which this lends plausibility to the posterior distribution reflecting processes characterizing macroinvertebrate assemblages (L465-467, Table 3 and Appendix S1: Figure S2).

COMMENT:

3) The authors present their new approach as an alternative to bioassessment methods based on reference sites. To strengthen the manuscript, they could directly compare the two approaches, as the same set of stream sites has already been evaluated using the standard approach (see: <https://riverstreamassessment.epa.gov/webreport/>). Conducting this additional analysis would highlight the differences between the methods, showcasing their respective advantages and limitations. For example, if the new approach works, higher proportion of the sampling sites can be expected to be ranked as “impaired” This comparison can be used as a limited validation to the whole new approach.

RESPONSE:

We agree that it is insightful to place our results in the context of traditional reference site-based methods used by National Rivers and Streams Assessment. However, because the approaches are fundamentally different (L652—684), a formal comparison is beyond the scope of this manuscript. Instead, we offer several lines of text to highlight differences based on a ranked comparison of the percentage of streams in “poor” condition (L702-724). From these comparisons, the largest disagreement 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 than reference site-based approaches. This is encouraging because we suspected that extensive conversion from grasslands to agriculture in this region has left few sites that are undisturbed (L709-715).

COMMENT:

1) L174-175: The presence or absence of a taxon (genus) is typically determined based on a small field sample (e.g., 500-600 counts in NRSA). This binary data is likely subject to imperfect detectability in most taxa. The authors should discuss the implications of this limitation, particularly how detectability issues might influence their results and interpretations (see Doser et al. 2023. Ecology, <https://gcc02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1002%2Fecy.4137&data=05%7C02%7Ckopp.darin%40epa.gov%7C1081d1a35b214b0d922908dd187b72b8%7C88b378b367484867acf976aacbeca6a7%7C0%7C0%7C638693640674911021%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=Xkj8afdva%2FYWMMEi1K7PYGGwbYd9qY35a3YuvItp0pM%3D&reserved=0>).

RESPONSE

The reviewer is correct that NRSA uses fixed-count subsampling for processing benthic macroinvertebrates. This could potentially increase false absence rates in our analysis (L692-693). We also thank the reviewer for directing us to Doser et al (2023) and Doser et al (2024). Although modeling detecting probabilities would not be possible with NRSA protocols (typically, sites are only visited once), we do mention that future applications might consider modeling taxon-specific detecting probabilities from replicate subsamples (L693-695). Presently, replicates of fixed-count sampling are not part of NRSA protocol.

COMMENT

2) L212-217: The reported correlation (is this Pearson's r?) appears weak. However, pairwise linear correlations may be insufficient for capturing the complexities of the relationships among predictors. Many human disturbances, such as agricultural land use and road density, are almost certainly constrained by geo-climatic variables (e.g., elevation, temperature, and precipitation) in multivariate, nonlinear, and interactive ways. The authors might consider using Variance Inflation Factor (VIF) or Spearman rank correlations for a more robust assessment. Additionally, a stricter test could involve modeling human disturbances as a function of geo-climatic variables using random forests or other machine-learning approaches. Such methods would likely reveal that a significant proportion of human disturbance can be explained by geo-climatic factors.

**RESPONSE**

**The reviewer is correct that we used Pearson’s correlation coefficients to assess pairwise relationships between the predictor variables in random forest models (L297-299) and calculated variance inflation factors(L299). Importantly, we found that all pairwise correlations were <0.7, and variance inflation factors were between 2.6 and 3.9. Indeed, VIF>5 can indicate problems with multicollinearity, however with random forest models (L300-301).**

COMMENT:

3) L399-400 and Table 3: The study uses Jaccard similarity and genus richness to assess the accuracy of JSM. However, the former is notably lower than the latter. The authors should provide an explanation for this discrepancy, considering possible methodological or ecological factors that could contribute to the difference.

RESPONSE

The reviewer is correct the probabilistic adaptation of Jaccard’s similarity suggested that it is easier to predict aggregated, assemblage-level attributes than compositional attributes (L680-684). Nonetheless we present compositional changes separately from changes in richness (Figure 7) such that differences can be assessed for readers. Further, we highlight that this new adaptation could use further research but provides information that is not easily obtained from traditional approaches (L683).

-------------------------------------------------

**RESPONSE TO REVIEWER #2**

GENERAL COMMENTS:

We thank Reviewer 2 for their comments and found them extremely valuable when revising our manuscript. We have extensively revised to include the major and minor comments. We also provide text and reference line numbers below.

COMMENT:

First, it’s curious to me why the authors use a Bayesian-based mixed-effects GLM for the taxa-environment models but then use an ML-regression tree approach for modeling the environmental gradients. These two approaches are quite divergent and it’s unclear how each is best suited to the particular modeling problem they were used to address. The authors comment on the advantages of JSDMs, but the reader would need to be familiar with this approach in order to understand why.

RESPONSE:

Thank you for pointing this ambiguity in the previous version of our manuscript. We have added text in the introduction to more thoroughly outline the benefits of using joint species distribution models (L89-107) and random forest models (L108-125). In brief, JSDMs accommodate sample-level random effects and are fit with bayesian inference. This allowed us to focus on a relatively few physicochemical gradients while statistically controlling for other factors using latent factors. The random forest models were preferred because they can account for complex relationships and a relatively high number of covariates.

COMMENT:

Second, the Introduction needs to better set up how this particular coupled modeling approach solves the problem of inconsistent reference-site quality among regions. It’s also entirely unclear whether this approach actually does any “better” than traditional approaches. Granted, the inferences are potentially more enlightening and interesting than previous work, but the believability of the results needs to be enhanced by clarifying many methodological issues described below.

RESPONSE:

Thank you for this comment. We restructured and rewrote much of the introduction and discussion to improve clarity of the coupled modeling approach (L74-88 and 108-128) and compare our results to those reported by the National Rivers and Streams Assessment (L702-724). In addition, we highlight differences between reference site-based approaches and our model-based approach (L652-701).

COMMENT:

I’m not sure the Introduction sufficiently sets up the specific issues / knowledge gaps addressed in the paper. The first paragraph is good, but needs to more clearly state the dire (?) consequences of inconsistent reference-site quality. That point was only mentioned in passing. For the next paragraphs, I was expecting a brief introduction to the various ways practitioners have circumvented or addressed the issue and why these previous approaches are inadequate—and hence the unique need for the current approach.

The second paragraph, as written, has a lot of duplicative information with the first paragraph. Consider blending or removing the second paragraph entirely.

RESPONSE:

We thank the reviewer for encouraging us to more clearly identify the problems associated with traditional reference site-based approaches. We have extensively revised the introduction. We used stronger language in the first paragraph (L53-73) and omitted paragraph 2 of the previous version. The remaining paragraphs describe the overall approach and the benefits of the JSDM and the physiochemical models.

COMMENT:

My first reaction to reading the third paragraph is “what about the extrapolation problem?” The idea of using empirical models to “zero out” anthropogenic influences on ecosystems is not new. The “ELOHA” concept touted over the last few decades applies the same notion to predictions of natural hydrology. Consider adding a recognition (even if in the Discussion) that the approach has been proposed and is not without limitations? For example, Hill et al. (Freshwater Science 32: 39-55) found that models that zeroed-out human influences on stream temperature produced biased predictions of reference conditions relative to models developed using reference sites. Please address this issue. Also, it’s not clear how this paper differs from (adds to?) the Kopp et al. 2023 paper in Freshwater Science. Seems like the modeling approach in that paper is simply reproduced here to address a specific question. Please clarify.

RESPONSE:

Based on the reviewer’s concerns regarding extrapolation, we have added an additional analysis to explicitly text for extrapolation (L306-319) and believe that defining the domain (area of applicability) of hindcast models will be an essential feature of model-based assessments moving forward (L596-612). Following, Meyer and Pebesma (2021) and Yuan et al (2024) we detected 77 sites (~4%) that had hindcasted values that were sufficiently dissimilar from the data used to fit the model (L477-485, Appendix S1: Figure S1). These sites were excluded, labeled as “Not Assessed” in our analysis (L485). We also clarified that we used z-scores to assess whether the removal of disturbance would change the physiochemical gradients after accounting for unexplained variation in the model or natural variability (L310-327)

We have also added several citations to direct readers to previous efforts that have linked anthropogenic disturbance to in-stream physiochemical conditions for hindcasting (L108-128) and discuss limitations (L596-612). The reviewer is correct that Hill et al. had suggested that hindcasted models could be biased. However, they did not recognize that the reference sites used in the study were a spatially biased subset of all sites that needed to be assessed (Figure 5 in Hill et al 2013). As consequence, it is unsurprising that the reference-site model (one fit using only reference sites) had better performance than the all-sites models, when evaluated against the reference sites but it is not clear whether this is a bias per se. For example, one could argue that the reference-site model would have the same bias if it were used to for prediction at non-reference sites. This is because the two models are fit to balance two different error structures. We contest that all-sites models are better because they include all sites that need to be assessed and used explicit covariates to account for disturbance but acknowledge that validation is challenging with both model-based and reference site-based and should be a focus of future research and conceptual development.

Finally, we have revised the introduction to differentiate this research from our previous work (L74-88, 108-128).

COMMENT:

More detail is needed about the modeling of species(taxon)-environment relations. Make it more clear that species are not being modeled. Rather, genera are being modeled. And somewhere in the Discussion, there should be a short acknowledgement about the limitations of modeling at such a course taxonomic level.

RESPONSE:

Thank you for the comment. We have made it clear that our models focus on genera (L228-231) and discussed the potential limitations associated with the level of taxonomic resolution (L685-701).

COMMENT:

Several basic concepts should be explained in this manuscript so readers aren’t required to consult the 2023 paper. First, which specific eight (Line 163) environmental variables were included in the models? Clarify that it is those listed in Table 1 if that is indeed the case. Second, the dependent variable was presence/absence of each genus, so what was done about taxa that were not identified to that level (e.g., most non-insects)? Third, given that the models were linear in nature, how were non-linear responses considered (e.g., the expected unimodal response along environmental gradients)? Last, were ALL sites used in the modeling or was a validation dataset used?

RESPONSE

Thank you for bringing the cursory treatment of our description in the previous version of the manuscript. We have changed to order of the sections in the methods to improve clarity. Specifically, we make explicit reference to the 8 physiochemical variables included as dependent variables (L211-212) and reference table 1 for readers to obtain further information (L228-231).

Our analysis focused on organisms identified to the genus level because this alleviates some ambiguity associated with aggregating ambiguously identified taxa into operational taxonomic units and improves transferability of our findings (L689-692). We also discuss that this decision potential contributes to false absence in our analysis but, because NRSA identifies most organisms to their genus we conclude that these instances are relatively rare (L686-689).

Previously we found relatively few unimodal genus-environment relationships and therefore assumed linear relationships for this analysis (L237-239).

We used all sites for modeling but included data collected during revisit sites as an additional form of model validation (L245-261).

COMMENT:

The Introduction section argues that anthropogenic stressors and natural environmental gradients are often confounded (e.g. agriculture and topography and climate). How was this issue avoided in this modeling step given that such a limited set of predictors was considered? Is it possible that modeled relationships between a taxon and, say, nutrients, includes the influence of elevation, topography, soils, and climatic conditions? If such confounding is possible, how does it influence the eventual outcome of the analyses?

RESPONSE:

Thank you for giving us the opportunity to clarify. In the revised version of the manuscript (L216-226), we describe that one advantage of JSDM is that they implement site-level random effects, such that unmeasured environmental variables are modeled (statistically accounted for) using latent factors. We also discuss the benefits and limitations of out model structure in the discission (L636-651). In brief, the model structure is beneficial because it allowed our analysis to focus explicitly on physiochemical factors that are commonly altered by human disturbance but limits our ability to make accurate predictions to new locations.

COMMENT:

The modeling of physicochemical gradients also needs more explanation. The anthropogenic factors were “zeroed out” of the random forests by setting each such predictor to its lowest observed value across the nation. Given the strong regional differences in percent agriculture, for example, is this a reasonable approach rather than building separate regional models? It seems that the current approach would create environmental space (e.g., certain regions) within the dataset where the “hindcast” was clearly extrapolating beyond observed conditions. For example, I suspect most NRSA sites in the western US had 0% agriculture in their basins, so the hindcast created from the national model likely set percent agriculture to zero. For sites throughout the plains, midwest, and southeast, however, 0% agriculture was almost never observed. This may be a non-issue, but it seems potentially problematic unless you clearly acknowledge that in many regions your hindcast is an extrapolation.

RESPONSE:

This is a valid point from the reviewer and prompted us to include an explicit test for extrapolation in out hindcast predictions (L306-319). In brief, we intentionally fit models using all available data to include the full gradient of anthropogenic disturbance values (L291-292) but realized that hindcast predictions required a new dataset that could be different from the data used to fit the models. Interestingly, the sites that were flagged for extrapolation were mostly located within the lower Mississippi river basin (L604-605). Because this region is unique and heavily modified it was perhaps unsurprising that removing human disturbance is beyond the domain of the model (L605-607). Further, context that maximizing environmental gradients and then testing for extrapolation may produce more accurate results than fitting localized models with potentially truncated gradients.

COMMENT:

I’m also concerned with the climate analysis. First, clarify that you computed the 10-year average (which years?) climate data for each site. Ten years is not a very long period over which to generate an average climate for the past or the present, and the choice of 1940-1950 seems arbitrary (the presence of nuclear tests notwithstanding). If pre-industrial climate is the needed baseline, 1900-1950 would be a better period over which to average. That said, it is well established that most US regions have experienced wetter conditions since 1970 relative to the 1900-1970 period. But the degree to which the “present” 10 years differs from pre-industrial is entirely dependent on the period over which you computed the averages. As a result, I can’t say that I believe your results about climate until you demonstrate that they’re robust to the time periods you selected to examine.

RESPONSE:

We thank the reviewer for directing our attention to this inconsistency in our analysis. We reanalyzed our data using a 1900-1950 climate baseline (L324-327). Although the mean values for the climate variables were similar to the previous 1940-1950, we found that the standard deviation of precipitation was erroneously log transformed. Correcting this mistake brought our results in line with the reviewer’s expectation (L501-504, L524-526, Figure 5a and 6a) and apologize for this oversight.

COMMENT:

The choice of thresholds for the RDPI is also puzzling. It’s not clear why it’s a problem that most sites had present-day RDPI values > 0. Aren’t you trying to zero-out human influences? Please explain this decision and whether an arbitrarily selected threshold would influence your results and conclusions.

RESPONSE:

Although a RPDI = 0 would indicate that human disturbance is completely absent from the riparian area, this is almost never found. As such, applying this threshold masked any regional variation in disturbance. We did not think it would be informative to report that >90% of sites across all ecoregions have some level of disturbance. Rather we selected 0.33 because it has been used by others to identify least disturbed sites (L343-346) and revealed greater variation among regions. We revised the discussion for clarity (L624-628).

COMMENT:

I found the Discussion to be lacking for several reasons. First, there are many sentences that repeat methods or results information. These are distracting and unnecessary. I would focus the paragraphs on your interpretations only. Second, the limitations of the approach are not described. I suggest an entire section on this topic, particularly devoted to the issues of confounding natural & anthropogenic factors, modeling at the genus level, and considering a very limited set of anthropogenic stressors. Third, the Discussion would ideally describe how this two-stage modeling approach in some way improves upon the more traditional methods. Currently, the statement is made that this approach merely adds to traditional methods. Just about anything can add to current methods. I believe this approach could be a substantial improvement over traditional methods, but the ways this is true are not fully developed in the manuscript.

RESPONSE:

We thank the reviewer for challenging us to improve our discussion and have performed extensive rewrites. We have removed many sentences that repeat results and methods and focus primarily on interpretation. Although we included limitations of our model-based approach, we found that integrating them into specific subsections improved the flow of the writing. Specifically, we addressed the limitations associated with the relatively small number of potential stressors and random effects structure (636-651) and genus level taxonomy (L685-701). In addition, we included several paragraphs to the compare and contrast our model-based approach to reference site-based approaches (L652-701) and a comparison between our results and those reported by the National Rivers and Streams Assessment.

Yours Truly,

Daren Carlisle

Thank you for the thorough review of our manuscript.