Response to reviewer comments

Thank you for giving us the opportunity to revise our manuscript. Reviewer comments were especially constructive and have improved the manuscript. The Reviewers' comments are cited or summarized in *red italicized text*, and our responses are in black Roman text. We include excerpts from our revised manuscript in blue.

Reviewer 2

> **Reviewer 2**: *The reported work only used BMP category and large-scale spatial indicators as predictors of their performance, but the parameters of BMPs (e.g., the size of infiltration) are missing. This may be why the model has a high variance.*

We generally found that specific BMP parameters such as infiltration size, drainage area, infiltration media/soil types, and other factors relevant to BMP performance were not well reported across studies. Our characterization of BMPs is further complicated by inclusion of BMPs that rely on completely different pollutant removal or prevention mechanisms. For example, nutrient management BMPs do not rely on a infiltration area and we would not be able to include these types of management BMPs if infiltration size were a regression moderator. This is certainly a limitation of the study, but the purpose of the project is a broad scale look at BMPs. Further studies investigating specific BMP types with common removal mechanisms are certainly appropriate. To address the Reviewer's concerns, we have incorporated a Study Limitation section to the Discussion:

> A few reviews have noted a common trend of insufficient methodological and site specific data among peer-reviewed BMP performance studies (Grudzinski et al., 2020; Eagle et al., 2017; Liu et al., 2017). We confirm that inconsistent reporting among studies complicates data extraction, effect size calculations, and attributing important sources of variance. Our desire to evaluate the effects of specific BMP parameters was hinderd by the overall lack of reporting of relevant parameters such as drainage area, infiltration media/soil, infiltration volume, riparian/buffer width and area, and other relevant factors. The lack of BMP specific parameters certainly contributes to our relatively high model variance. The International Stormwater BMP Database addresses some of these concerns through a standardized reporting format. Our future efforts will incorporate data from the International Stormwater BMP Database with data retrieved through a systematic review. Additionally, this study was a broad scale look across BMP types which limits factors that cannot be compared across different types of BMPs. For example, livestock management BMPs may not rely on or report parameters such as vegetative buffer width or infiltration area. Inclusion of such factors in our meta-regression approach would necessarily exclude certain types of BMPs. Future BMP-specific meta-analysis might be more useful and informative for practitioners by providing effect sizes of parameters that were necessarily excluded from our study.

> **Reviewer 2**: *The heterogeneity of FIB, TN, TP, TSS are high, the study require subgroup analysis or sensitivity analysis on this issue.*

I appreciate this comment, but want to caution against extensive subgroup analysis/sensitivity analysis. While heterogeneity was indeed high, the primary purpose of the meta-regression models are to investigate how much of the variance that a set of pre-determined moderators explain. Meta-regression were conducted in lieu of traditional subgroup analysis. While additional subgroup analysis could be conducted based on additional moderators, we risk conducting a "data-dredging" study [1]. However, we did add post-hoc sensitivity analysis for runoff source and catchment scale to the paper based on this suggestion. We did not find strong evidence indicating these factors are highly influential to the results we found. This is likely, as you correctly point out, do to the lack of BMP specific parameters that better explain performance than our chosen moderators.

Reviewer 3

> **Reviewer 3**: *Is there any moderating effects of study area size or scale (i.e., field, farm, and watershed) and urban versus agricultural source on the performance of BMPs?*

We added an additional post-hoc sensitivity analysis for both these factors to the study. While this is not the same as fully incorporating the moderators into the original model, the sensitivity analysis indicates that we don't have strong evidence that these terms are highly influential to our results. This is confounded in part by the fact some of our BMP types are only applied to urban or agricultural BMPs and highlights a need for future meta-analysis to be applied for specific BMP types instead of across multiple BMP types.

> **Reviewer 3**: *What is the temporal coverage of the aridity index? Does it reflect long-term climate patterns or only a single year? Using a snapshot of the climate variable would be less optimal.*

It is annually averaged from 1970-2000 which should suffice as representative for climate patterns in this study. We have added clarification in the text which now reads:

> We mapped study location coordinates to aridity index values published in the "Global Aridity Index and Potential Evapotranspiration Database - Version 3" (Global-AI_PET_v3) which provides gridded 30 arc-second annual average (annual averaged from 1970-2000) precipitation and potential evapotranspiration estimates.

> **Reviewer 3**: *Since the locations of the studies are not precisely defined in many cases, please provide more details on how to spatially link studies to the aridity index map to obtain aridity values.*

We added the following text to describe how spatial locations were derived if they were not provided in the study:

> Study locations were recrorded as latitude and longitude coordinates as described in studies. If coordinates were not provided, the study team used Google Maps to locate approximate study location using site descriptions (county, city, municipal buildings, etc.) from the study and recorded the coordinates.

> **Reviewer 3**: *Have you examined the effects of soil biophysical properties in addition to aridity? Can you obtain more detailed spatial locations of the various studies so to link these studies to local soil conditions in addition to climate?*

We did not include soil properties with the assumption that many infiltration based BMP modify soil as part of BMP construction. It is certainly possible that native soil conditions provide some additional moderating effect, but we did not obtain this data based on our assumptions.

> **Reviewer 3**: *Could you please elaborate more on the topic "scaling BMP pollutant reductions to basin wide water quality improvement remains a substantial challenge". This is a very interesting point, what modeling or mixed approaches exist to achieve this goal?*

We have expanded on this slightly in the conclusion. Essentially it is difficult to empirically link water quality improvements to BMP implementation in many studies due to short study lengths, underlying measurement and analytic uncertainty, and difficulty establishing adequate controls. Many watershed scale studies therefore rely on numeric watershed models to demonstrate BMP performance (see the USDA CEAP program for example).

> **Reviewer 3**: *Please expand the abbreviation "ROSES" in the main text.*

Added.

Reviewer 3: *Change "Standard deviation of control measreuments" to "... measurements" in Table 2.*

Fixed.

Reviewer 3: *Change "and outlier" to "an" in Figure S9 to S14.*

Fixed.

Reviewer 3: *In the sentence "Any study values outside of the full model confidence intervals would be considered an outlier.", by "study values", do you mean the intercept estimate of a study?*

I worded this poorly and have updated the figure captions with the following:

> Any intercept estimates outside of the full model confidence intervals implies the excluded study is a potential outlier.

Reviewer 3: *Explain the abbreviated parameters in Figure 1 – 5 (e.g. TS, TP, TN, etc.)*

Added definitions to all the figure captions.

Reviewer 3: *Delete one of the "the" at line 277.*

Fixed.

Reviewer 4

The statistical analysis was rigorous and thorough. That, alone, warrants a recommendation for acceptance of this manuscript. My primary concern is that the models were fitted using only one of the four BMP variables – BMP Subcategory. Were models with the other three BMP variables evaluated? In the mid-Atlantic US region, critically important BMPs for rural and agricultural lands are vegetative buffer strips, and for urban areas constructed wetlands are quite important. For example, much work has been done by researchers in North Carolina on metrics of vegetative buffer strips re-

lated to their effectiveness. Evaluation of models with specific practice characterization would have been informative.

We agree with the overall comment here. This paper was our first cut analysis across a broad range of BMPs. Our approach requires somewhat broad variables used in the regression models so we do not eliminate studies that do not include specific BMP data, such as buffer strip width. Further meta-analysis should focus on specific practices where relevant specific BMP parameters can be properly evaluated. We've added additional context in the Study Limitations section.

That leads me to my second concern. Only the last paragraph of the manuscript is presented as a summary. I recommend that this paragraph be reconstituted as a separate "Conclusion" section and that it be more thorough than this simple summary paragraph. For example, it would be useful to know what the authors expected to find in their modeling studies versus what actually was found.

We have incorporated a conclusions section and expanded a bit more on our expectations as well as incorporated suggestions from other reviewers.

Reviewer 5

Reviewer 5: *Add spatial extent covered in abstract*

Added the spatial extent.

Reviewer 5: *Add sediments or suspended sediments to keywords if there is space*

Added.

Reviewer 5: *1st Paragraph of introduction (Line 28), add a sentence about sediment impairments to waters similar to the ones discussing FIBs and nutrients impairments*

Added the following:

> Approximately 15% of river and streams have excessive sedimentation, which leads to twice the likelihood of a stream to have poor biological condition (EPA, 2017).

> **Reviewer 5**: *Include small paragraph in discussion about how search terms of "BMPs" or "best management practices" and inclusion criteria (concentrations) may have not captured some studies. For example, if the study used the name of the practice instead of BMP. Also, there are a suite of agricultural drainage practices but studies generally focus on load reduction (from flow reduction) compared to concentration reduction (see Frankenberger et al., 2023 https://elibrary.asabe.org/abstract.asp?aid=54314 who's review has a cluster of those studies in the Midwest, which is blank on this map).*

This escaped our notice, thanks for raising this relevant shortcoming. We added the following to the Study Limitations section within the discussion:

> One reviewer also noted some examples of missing studies, attributable to our selection of search terms and inclusion criteria (Table 1). In particular, our search query did not include specifc practice names, instead using "BMP" or "best management practice" as a keyword search. This choice may have led us to miss studies that only included the practice name. Or inclusion criteria of concentration based studies may have also lead to the exclusion of studies that focused on load reductions (through flow reductions) but may have included data on relevant concentration effects.

> **Reviewer 5**: *Lines 61 – 64 consider adding an example BMP for instances where effluent is and is not related to influent concentration*

We modified the lines to include the following:

> Barrett (2005) demonstrated that percent pollutant reduction is often a function of influent quality. Specifically, for certain types of BMPs percent removal is low at low influent concentrations, and increases with increasing influent concentrations. Horvath et al. (2023) found that influent phosphorus concentrations had some explanatory ability for BMP performance among grass strips, bioretnetion, and grass swale BMPs. However, for some types of BMPs, such as media filters and permanently pooled retention basins (Barrett, 2005) and pollutant parameters, effluent concentration is unrelated to influent concentration.

> **Reviewer 5**: *Line 136 – specify version of R used*

Added.

Reviewer 5: *Line 149 – check for consistency in defining equation terms (CV v. CV2)*

Fixed. Now reads as below:

$$v(ROM) = \frac{\sum_{i=1}^{K}\left(CV^2_{control,i}\right)/K}{n_{control}} + \frac{\sum_{i=1}^{K}\left(CV^2_{experiment,i}\right)/K}{n_{experiment}},$$

where v represents the sampling variance, $CV^2_{control,i}$ and $CV^2_{experiment,i}$ are the squared coefficients of variation from the ith study for studies 1, 2, ..., K.

Reviewer 5: *Line 160 – mean annual potential (or reference) evapotranspiration*

Fixed.

Reviewer 5: *Line 191 – "published after between 1999 and 2023" confusing wording, 2000-2022?*

Fixed.

Reviewer 5: *Line 194 – PO~4*

Fixed.

Reviewer 5: *Lines 213 – 214 – clarify "may not be reliable" in general, or just for lower aridity?*

It would not be reliable when extrapolated to lower aridity regions due to lack of model coverage. Clarified this in text.

> **Reviewer 5**: *Lines 61 – 64 consider adding an example BMP for instances where effluent is and is not related to influent concentration*

Added.

> **Reviewer 5**: *Lines 242-254: Nice discussion on spatial coverage with the caveat that we are not aware of the actual distribution of BMPs. However, I am curious, are BMPs over-represented or just predominately located in humid regions? Could it be related to larger population concentrations along the humid east coast than the arid west? Are detention practices common in locations with little rainfall and high potential loss of water to evaporation? See my recommendation for Figure 7 to potentially add a 3rd distribution of aridity index of urban & agricultural land covers across the US.*

We have added a third distribution, based on your suggestion below, looking at just urban and agricultural land and still find some discrepancies in the distribution of aridity values.

> **Reviewer 5**: *Lines 349 – 351: Based on your experience trying to synthesis information across studies in this review process, do you have recommendations for minimum reporting requirements or standards for future studies on BMP concentration reductions? If so, please list in text or perhaps in a box.*

We added the following text to the conclusions:

> To improve future data synthesis efforts, we highly reccomend future BMP studies follow the reporting guidelines provided in Eagle et al. (2017). In particular provide clearly defined water quality parameters, tabular data (either in the manuscript, as supplemental material, or in an open data repository), and error estimation at minimum. Control and treatment means should be made available in reports, not just the transformed efficiency or percent change values. Furthermore clearly defined controls and treatments are fundamental for comparing across studies. Finally, a major shortcoming in our synthesis was driven by the lack of reported covariate data. Although data such as soil type may not be an within study experimental covariate, documenting of these types of study variables is useful for data synthesis efforts.

> **Reviewer 5**: *Consider defining abbreviated terms in the figure caption for clarity (e.g, DIN, FIB)*

We have added terms to all the figure captions.

> **Reviewer 5**: *Table 1 - Excluision Criteria → Exclusion Criteria - Under Inclusion Criteria/Outcomes include sediments in addition for FIB and nutrients - Under Exclusion criteria/Outcomes "Studies not explicitly linking reductions to a specific BMP..." it may be more clear if you say linking "concentration reductions"*

The table has been updated with the suggested langauge.

> **Reviewer 5**: *Figure 1 - While the color palette may be appropriate for normal, trichromatic, and dichromatic conditions there is very little contrast in the monochromatic or grey-scale view. If article is laser printed or photocopied, it is difficult to distinguish between the stacked elements of the bar graphs. Maybe white or black outline separating stacked sections could help. - Please increase the length of the y-axis for panel C to for larger font size/less crowding*

We moved panel c to a new figure (now figure 2). The parameters are not plotted on individual facets with the overall distribution in light grey in the background to reduce reliance on color for parameter identification. We also improved legibility of the y-axis.

> **Reviewer 5**: *Figure 3 - Consider adjusting the background grey value for no-data in maps. In grey-scale/monochromatic view the mid-tone greens are indistinguishable from the no-value grey. Labeling the states with the number of studies does help.*

We changed to a transparent background for states with no data.

> **Reviewer 5**: *Figure 4, 5, 6 - Consider adding including the range in sampling variance in the caption or provide a visual key to demonstrate the values represented by differently sized points.*

Legends indicating sampling variance have been added to figures.

> **Reviewer 5**: *Figure 5 - Add panel letters to figure caption to improve clarity. - Consider adding note for aridity index (annual precipitation/potential evapotranspiration) in the x-label or in the figure caption that higher values indicate more humid conditions.*

Thank you for the suggestions. We have incorporated all the suggestions to the figure.

> **Reviewer 5**: *Figure 7 - It is difficult to see the tail of the US. Aridity Index values, consider having a slight offset from zero above the x-axis/y-axis intercept. - Consider having a 3rd distribution shown – U.S. Aridity Index Values for urban and agricultural lands, with other land-covers filtered out, if these are the land-cover types that are associated with the placement of BMPs you analyzed. You could use a product with the same resolution as your Aridity Index values such as the 30m National Land Cover Database (https://www.usgs.gov/centers/eros/science/national-land-cover-database#overview) - Consider adding note for aridity index (annual precipitation/potential evapotranspiration) in the x-label or in the figure caption that higher values indicate more humid conditions.*

Thank you for the suggestions. We have incorporated all the suggestions to the figure. We switched from overlapping densities to a ridge plot with each category labeled on the y-axis since it gets difficult to discern densities with three overlapping plots.

Bibliography

[1] S. G. Thompson and J. P. T. Higgins, "How Should Meta-regression Analyses Be Undertaken and Interpreted?", *Statistics in Medicine*, vol. 21, no. 11, pp. 1559–1573, Jun. 2002, doi: 10.1002/sim.1187.