*Response to reviewer comments on revised manuscript “Prioritizing management goals for stream biological integrity within the developed landscape context”, by M. W. Beck, R. D. Mazor, S. Johnson, K. Wisenbaker, J. Westfall, P. R. Ode, R. Hill, C. Loflen, M. Sutula, E. D. Stein.*

Dear Marcus Beck,  
  
I have now received reviews of your article from two outside readers, appended below. These reviews indicate that while your article potentially presents an important contribution to the field, revisions are necessary before your article could be accepted for publication.   
  
If you are prepared to undertake these revisions, we look forward to receiving your revised manuscript, and ask that you please directly respond to the readers' comments when you submit it.  
  
Please be advised that your paper might be rejected or be sent for a new review in light of current literature if we do not receive a revised manuscript by Aug 26, 2019. If an extension is needed, please contact me as soon as possible.  
  
To submit a revised version of your article, please go to https://www.editorialmanager.com/fws/ and log in as an Author. You will see a menu item called Submission Needing Revision and will find your submission record there.   
  
Best wishes,  
  
Daren Carlisle  
Associate Editor  
Freshwater Science  
  
Reviewers' comments:  
  
ASSOCIATE EDITOR GENERAL COMMENTS  
  
The revised manuscript contains meaningful revisions based on reviews of the first draft. I agree with the reviewer’s enthusiasm about the paper’s relevance and interest to Freshwater Science readers. After an additional careful reading of the revised manuscript, and based on the comments of the reviewer, I find several remaining issues that need attention. These issues are less about the science and more about the organization and clarity of the manuscript.   
  
First, the reviewer continues to have a major problem with the length, depth, and substance of your case study. In my comments on your first draft I noticed a lack of clarity and detail in this section, but I was not concerned about it’s length. I think you’ve done reasonably well with clarification (but see remaining issues in my specific comments below), but the reviewer makes a convincing argument that the case study narrative seems disproportionately large. I think the solution is relatively simple. First, move all tables and graphics about the case study to Supplementary Information (SI). Then, create a sub-section in the Discussion section entitled “Case Study: Application of the Model etc etc..” In that section, provide a brief description of the watershed and its management challenges, then describe the web tool. All other narrative in the main body about the case study can then be moved to SI. I believe this approach retains the important message in the main body that your models have actually been used in a management context, and makes the fine details available to interested readers.  
  
Second, the writing still requires extensive clarification. Comprehension was often challenging for me; many sentences require clarification or simplification. I indicate where I had the most difficulty in the detailed comments below.  
  
Third, the manuscript requires better organization. Your objectives and methods are not aligned. Nothing was said in the objectives about classifying all CA stream segments. The Methods describe an additional post hoc classification with sites having biological data, but there is no explanation for why this was done and how it is relevant to your objectives. Is this additional classification needed to make the tool usable for decision making? The Discussion section seems to ramble, and is missing key narratives. I think you spend too much space discussing possible alternative applications and nuances of the model, and not enough space discussing your results. Where is the comparison of your model performance to those of previous studies? Where is the discussion of your sensitivity analysis and its implications for applying the model?  
  
The reviewer’s remaining major comment deals with your description of the prediction intervals. I didn’t have this concern because I recall you clarified that your prediction intervals were NOT to be interpreted in a statistical sense. Nevertheless, be sure that you can address this concern by pointing to existing text or adding a bit more if needed for clarification.   
  
Finally, I continually struggled with the language of “constrained,” but the reviewer did not seem to mind. Perhaps I struggled because the term already has a meaning to stream ecologists, but it’s also simply vague. I think you need to reconsider that term—even though you define it early on. It just doesn’t describe what you’re trying to portray. It seems like “range of attainable biological integrity” or something similar is more descriptive. There are just too many possible meanings to “biological constraints…….” Nevertheless, I will defer to the Editor on this issue.   
  
ASSOCIATE EDITOR SPECIFIC COMMENTS  
  
Line 66: It’s not clear what you mean by: “Context is required that describes…how bioassessment data collected over multiple locations and times can be used to support decisions or identifying priorities.” Why would managers or decision makers have a goal of using data from multiple locations/collections?  
  
Line 72: Do you mean “restoring” streams? The rest of the sentence implies that restoration is the specific management goal.  
  
Line 81: I suggest you clarify that these modifications include channelization and/or burial. And it’s not clear how this sentence relates to the previous or the subsequent sentence.  
  
Lines 87-104: The entire paragraph needs more clarification. The topic sentence would improve if you clarified: “…present landscapes are likely to limit management options for restoring biological integrity.” The third sentence is very awkward. I think you should avoid “predicted range” because its meaning is unclear. Perhaps a term like “predicted range of attainability” or “range of ecological attainability” or something similar?  
  
Line 105: Topic sentence is a bit vague. The second sentence is probably a better topic sentence.  
  
Line 123: Even though you defined—still rather vaguely—your concept of “constraints” in a previous paragraph, this topic sentence is still unclear.   
  
Line 126: I don’t think “…using statewide data…” sounds like an objective. After reading the manuscript, it seems to me that you have three major objectives: 1) develop and validate a predictive model, 2) apply the model to classify all stream segments in the state, and 3) provide a case study within a single watershed to illustrate how model predictions & classifications can be used in a decision-making scenario.  
  
Line 174: This sentence needs clarification, particularly the phrase “…identify the likelihood of biological alteration…” Do you mean that this statistically-based value is considered a threshold below which a site is considered impaired / altered?” Or is there some other way this value is used to estimate likelihood of alteration? It sounds like you modeled the actual SCSI values, so why do you need to discuss this threshold? If it is used to interpret / apply the model, it might be best to mention this threshold then rather than here.  
  
Line 204: Not sure what you mean by “adequately described…” Do you mean that preliminary models using additional predictor variables performed no better than models with the current set of predictors?  
  
Line 205: This sentence is vague and too wordy. Do you mean that these predictor variables were selected because you believe they are indicators of the land-management activities that are most likely to limit the attainability of biological integrity?  
  
Line 207: This sentence is also awkward and unclear.  
  
Line 210: Is there a geospatial indicator of channel modification? Seems unlikely. Then why bring up the topic? Sentence is also unclear and awkward.  
  
Line 213: Not sure what you’re trying to say here. Do you mean the model was intended to be a prediction tool that uses landscape drivers and in no way attempts to explore specific causes / mechanisms of biological deterioration?  
  
Line 250: It seems like this validation approach is limited. Your goal is to make inferences about a predicted range at each site, but your validation only measures how well the median of your range predicted the actual CSCI value. I don’t have a suggestion for resolving this, nor am I certain that it’s a problem. But the disconnect between your intended inference and the actual validation procedure is worrisome.  
  
Line 261: I think Figure 3 is too busy. Can you dramatically simplify the figure to show just the key explanations of your classification?  
  
Line 267: Drop “…for the level of landscape development…” from the sentence. It is implicit that the landscape development of each reach was used to generate the prediction.  
  
Line 285: Clarify this sentence. I suggest: “A separate classification was made for sites where biomonitoring data were available.” But were these sites used in the calibration dataset? Is it a problem that you are now using the model that was built on the calibration sites to classify the calibration sites? And for what purpose are you making this additional classification?   
  
NOTE: I am recommending that material from line 291-333 be moved to SI.  
  
Lines 293-301: This paragraph needs clarification. Is the classification system meant to be used ONLY alongside an actual biological sample? It sounds like that is what you are saying. I think this paragraph should clearly explain how you intend managers to use the classification system. That will set up the next paragraph where you describe an example.  
  
Line 312: Not sure what a “spreading ground” is.  
  
Line 319: This sentence is awkward and needs clarification. Did the stakeholders actually use the segment classification system to develop the three priorities, or were these priorities developed independently and subsequently applied to each segment based on its classification from the model output?  
  
Line 326: Unclear sentence. This sentence doesn’t describe the left-hand side of Figure S2. In the figure, you show four possible conditions for each of the your four classes, and these conditions appear to be based on the actual CSCI score of a biological sample collected within each of the four classes. I also question the wisdom of putting this figure in the Supplementary Information rather than the main body.  
  
Line 349: Simplify to: “There was generally good agreement between observed and predicted CSCI statewide”  
  
Line 351: “For the calibration dataset, observed and predicted values were correlated (r=0.75, RMSE=0.17), with an intercept (0.04) and slope (0.93) that indicated minimal bias.”  
  
Line 353: “Performance was similar with the validation dataset (r=0.72, RMSE=0.18)”.   
Why no slope for the validation dataset?  
  
Line 364: Provide r value for Sierra region, as you have done for other regions.  
  
Line 369: I think you’re trying to say: “Statewide, spatial patterns in the predicted limits of biological integrity were similar to patterns in land use.”  
  
Line 377: This paragraph describes the results of your comparison of the stream classification to actual biological assessment scores. You need to explain how this analysis fits into your objectives. Was it a part of model validation? If so, using the calibration sites for this purpose is inappropriate.   
  
Line 385: This material is good. I’m glad you addressed the question of sensitivity of the results to various analysis decisions. But I hope you Discuss the implications of these findings—no matter how obvious—for management decisions based on your models.  
  
Lines 399-425: NOTE that I’m recommending most of this material be moved to the SI.  
  
Line 428: “…landscape context for evaluating observed conditions.” doesn’t clearly communicate what your tool provides. Assessment tools provide (hopefully) an accurate and precise estimate of condition. Your tool estimates the likely range of attainable condition if remediation is implemented.  
  
Lines 441-446: These sentences are vague and filled with jargon (e.g., temporal and spatial scales, watershed scales, etc.)  
  
Lines 455-458: You say the model could be used to identify locations where TALU could apply, but is not intended as a tool for defining tiered uses. I don’t think most readers will understand this nuance.   
  
Line 474: I don’t follow the need for this heading. Most of the Discussion to this point has focused on applications of the model, including the evaluation of management options.   
  
Lines 475-506: Okay. so this is where you discuss the case study. I would rename the subheading so it’s clear you’re talking about the case study.   
Lines 507-525: This is way too much detail. Some of it is potentially relevant for the short section on the case study, but the rest belongs in SI.  
  
Line 585: You already made this point. No need to repeat it here, or perhaps you can remove it from its earlier location.  
  
Reviewer #1: Review of MS2019026  
Prioritizing management goals for stream biological integrity within the developed landscape context - REVISED  
  
Overall Comments for Authors:

I was the first reviewer on the previous version. My overall impression of the model version of this paper has not changed. I think this an interesting application of the quantile modeling to bioassessment data and the idea of biological constraint builds on earlier landscape ecological work related to bioassessment data and provides a tool that has obvious benefits for management. I applaud, I praise, I laud the authors for addressing the comments as they have. I think the explanation of the quantile methods has improved. I think the introduction reads much more cleanly. Thank you for your effort.  
  
I continue to disagree with the associated editor on the inclusion of the case study. I don't honestly think we learn much from that experience here. As you see from my specific comments, there is not enough detail from that experience to have learned much. If the point was to demonstrate that the concept can be applied to help prioritize watersheds - than why is that scientifically insightful and worthy of attention of the FS reader or the greater watershed management audience? I am both the former and the latter and in a scientific article I would want to know more information to conclude that this case study is value added: what is the null model? Did you have them prioritize watersheds before and after and compare to see how much their rankings changed? Where are the details on their deliberation to show how the tool changed thinking? Right now, the main text of this paper simply provides statements like "without this information, stakeholders struggled to prioritize" - what does that mean? How do you quantify struggled? Even anecdotally? The tool helped stakeholders "explore the key decision points that affect the model output". How so? How much did they change things based on changing these decision points? "The final decision of the group to prioritize…was based on an iterative process where ideas were discussed and shared freely among stakeholders." Did you need this tool to do that? How did this tool change that? "This approach ensured that stakeholders were generally in agreement with the final product" - where is the data on this? How do you know this improved this agreement versus other approaches?  
  
I think if you are going to use a case study in a FS scientific article, then the case study of an application of detailed technical approached needs to be so brief as to be inconsequential and non-distracting from the core technical material, or it needs to be rigorous enough in application that it can stand on its own technically. I think if you wanted to include this case study, it could be a much briefer aside where you do not try to make assertions about how the tool improved or changed things without real data on it. Just state briefly that the tool has been applied to help make prioritization decisions in this watershed - then I think you must discuss what this tool replaced in terms of process and that there is a hope it will improve decision-making and that someone hopes to study that. If not, then I don't see the rigor in this case study being useful to your narrative. Reference a technical report for that or get some social scientists to work on really quantifying if these tools indeed improve things - in some way that is more than just speculative.  
  
Don't get me wrong - the model development side of this paper, in my opinion, is an amazing and creative contribution. I just think the case study adds nothing to that. Any tool can be applied. Why is that novel or even worthy of FS reader's attention? Has it contributed, is some quantifiable way, to an improvement? That to me would be more interesting. Right now, I think the case study results are just speculative. That should not hold up publication of the model portion. And, if the AE continues to disagree, then I demur. This is just my opinion. I'd like the quantitative or even semi-quantitative insight to be able to demonstrate to managers that this tool truly improved or changed things. Right now, that information is still speculative, even if the tool is clearly a useful one.

The major technical issue I noticed this time around was the way you describe the quantile RF output. Not sure why I did not notice this before, but the term prediction interval has an existing connotation: for a regression model, if I recall correctly, a prediction interval is where one expects new observation to be located. It is quantified using a t or z score and sample size, etc. It is an extension of the confidence interval. I do not think that is what you mean, so it may be misconstrued. You are estimating quantiles. And, with this machine learning approach, resampling could be done to estimate error around those quantiles. That is not the same as the predicted quantiles themselves being prediction intervals around a single observation. These are population-based quantile estimates. I am not a statistician, but I imagine a real prediction interval around a predicted median would be quite different than the range between predicted deciles, as you have done. Worth checking and rethinking. More in the specific comments.  
  
Specific comments follow by line number (Any statement below should be preceded by an "in my opinion" …):  
Line number Comment  
32 Switch for with "with"  
44 Strike "that were", strike "clear"  
64 Strike "place" change to "limit the…"  
65 2x negative. Changed to Resource management decisions might be improved if information were…  
79 Sentence beginning "Although…" could use citation  
81 "…integrity have been…"  
84 "designation" no -s  
92 "….could be prioritized at less constrained sites where….  
94 "…higher management priority (i.e., for protection) relative to a site that is scoring within the expected ranged based on landscape development."  
96-98 I do not understand what is being said in this sentence. Are you talking random site effects? Maybe re-read and clarify.  
111-113 Is this DPSIR model necessary? I think I only see it here. Kind of comes out of nowhere. And you have a figure on it that is then not really revisited.  
119 "…scores that are likely given any landscape context."  
120-121 It might be nice, in place of DPSIR, to have a visual conceptual model of how your process works.  
130-135 Compare to what? What is the null or existing model against which this new tool is supposed to improve things? What are stakeholders using now? In my experience, it is either "fix the worst sites - 303d sites - first without any context of what uplift can be expected" or its "which watershed is politically the best to work on". And, they used raw water quality or bioassessment scores and decide on that. You also have EPA's Recovery Potential and Healthy Watershed tools that many communities rely on. These are never mentioned or even brought up. So, I think you need some foil for you method or else it's potential benefit is hard to gage.  
151 Is stream hydrography a stretch? The closest thing you ended up using was canal density. Is that hydrographic?!?  
172 "…1.4, which values near 1.0 indicating less deviation from…" You might want to stay away from what a score <> a standard deviation means; but a value of 1.4 might also indicate imbalance. Jury is still out on these type of O/E responses.  
183 Replace although with while  
190 Ode et al.  
193 StreamCat were - data is plural. You may want to check this throughout.  
195 Ok. You say you don't need to match dates because land use did not change dramatically during the period of 16y. On line 146 - you say land cover changed 38% over 27y and you suggest that is a lot - an impressive amount. That could be up to >1%/y. Over 16y, that could be 16% urbanization. If the urban threshold papers(e.g., Cuffney and Qian/King and Baker response) are correct and it takes small changes to shift streams, then I think your argument that not matching sample years may be less defensible. Just saying it is worth a thought. I agree with you, but we need to be honest about our logic.  
208 …given that our focus was on constraints to biological condition typically beyond the scope….  
212 "…,whereas modified channels are not landscape scale measures."  
213 You do focus on ultimate vs proximate causes. I think that is a fine angle.  
222 What are "robust predictions"?  
223 MLR can model complex, non-linear relationships with interactions. So what do you mean?  
227 "This modelling approach generates predicted quantiles of likely scores…"; So here is the statistical language question/comment. Wouldn't a prediction interval, which has a loaded definition already, be around a specific quantile rather than for the individual observation given your modeling approach (which is to predict quantiles rather than specific values)? I'm not sure you are predicting a prediction interval, are you? You are predicting quantiles and you can generate a confidence interval around those (because they are parameters). But, is the distance between your predicted deciles really a prediction interval? I don't think so…you might want to consider different language to be precise. But I may be wrong. A statistician would know.  
235 Bound on the median? These are not bounds on the median. They are predicted quantiles. This is not a confidence interval of your median (which seems like what would be the bounds and which you could calculate with resampling).  
240-246 Is there no way to do this without binning? It seems off that there is no continuous solution…  
242 "…on a random draw of sites from strata of quartiles defined by…"  
244 "…landscape development among regions (i.e.,…" Between is typically two objects and among is for >2, I think…  
246 The remaining 25% of sites were used…. Right? Since you mentioned 75% above.  
252 …indicated food predictive ability.   
254-256 I would cut this. Either start with this sentence or drop. I think its redundant.  
258 I had noticed this earlier either. You use 1.4% of the segments to predict the rest. Is there any concern about applicability of your model to all segments? Can you run a test of what predictor combinations are within the experience of the model and flag segments that you cannot predict, a la the test Van Sickle developed for the O/E code all-subsets DFA version? There are likely sites that just are way too outside your model predictor range set… A multivariate test should work. I think Mazor et al. 2016 may have used one for the CSCI development actually. Not sure.  
264 "We used a CSCI threshold (typo) of 0.79, following previous examples (Mazor…)…and the predicted 10th and 90th percentiles of expected CSCI scores to define an expected range." Typo of threshold.   
Also, I think you can reference Mazor and not mention the 10th percentile of reference calibration sites, since you have calibration sites too?  
Lastly, again, not sure this is a prediction interval. Need a new term - tried expected range - could use predicted range, perhaps?  
267 "Stream segments where the predicted 90th quantile score was below the threshold were considered likely constrained, whereas those where the predicted 10th percentile was above the threshold were considered likely unconstrained."  
273 "…depend on the percentile range of score…."  
274 Word choice on certainty. Band or percentile range. I am not sure this is certainty, which is, again, a statistically loaded term for a variable value.  
272-284 Why does this matter? Isn't it self-revelatory? If you change the width of quantiles, of course more will fall above or below, same with the threshold. I am not sure you actually need to demonstrate this…  
298 How so? How would it be prioritized differently? Is that based on experience or interviewing?  
300-301 You lost me here.   
309 I think this should be Figure 4b.  
329-331 You lost me here too  
332-333 I think you already referenced this table above (line 321).  
350 "…and the predicted medians (r=0.75, RMSE = 0.17; Table 4, Figure S4).  
351 Scratch this sentence. See line above. Just put in parenthetical.  
349-352 I am getting confused on how you evaluated quantile prediction. I guess median vs. observed is a fair estimate of how well the model predicts central tendency, but how do you evaluate how well it is evaluating other, even extreme quantiles? What is your confidence in site quantile estimates? Because isn't the observed score for any site any possible quantile? Medians are expected 50% of the time you sample a site, but wouldn't you be able to use multiple sampled sites to evaluate your other quantile predictions? Someone with better stats thinking than me needs to chime in - but since the quantiles are important - how well are you modeling them?  
353 "…suggesting minimal median prediction bias…"  
362 "slightly"? On p26 you say "poor". I agree with the latter. Or at least put it into context. Slightly lower does not imply the 30% lower performance that it is.  
Also, I think you should use your validation results more - that is a more independent, true test of performance, is it not?  
367 You could show p values for the slope and intercepts.  
374 I bet these results are pretty close to what you'd get if you just use percent urban and percent ag….  
377 "…within the decile range as often…."  
381 Replace "caused by" with "evidence of"?  
382-382 Really? Is the CH underscored @ 13%? I am not sure these statements are consistent across results.  
386-392 Again, is this all that insightful? It is kind of self-revelatory…  
404 "…lower watershed had (predicted or actual) median CSCI scores…." Don't you mean actual observed CSCI scores? I was confused.  
407 How was "effectively used" quantified or even qualified? Versus what? What alternatives were tested or compared? Did you interview stakeholders?  
407-415 Was there a before/after comparison of participants? If so, where are the data on that to compare? How did the tool change their decision making. You just seem to state that it did - but in no specific or quantified way or even qualified way.  
419-425 But this also tracks land cover in this watershed. A big benefit of the tool is not just protecting the best or restoring the worst, but being informed by constraint. I am just not seeing much in these case study results that can be confirmed/tested - it is very circumspect and not even really observationally detailed. What are we really learning about the management process/experience that has changed?   
432-435 I think the case study is so underdeveloped as to be not very useful to any audience. You finish by saying it can help identify where goals could be focused. So, how did it help do that in the case study - with real data on the participants experience.  
438-439 I would add "or exceeding" after meeting - likewise, couldn't you add "…or that could exceed bioobjectives." I mean, we should not race to the bioobjective - we should race to where segments COULD achieve. And if that is HIGHER than the bioobjective, that is what we should be encouraging, don't you think? I think you can emphasize both ends of the spectrum of tool use.  
457-458 Why? I am interested in your defense.  
475-485 So I am wondering where the substance is here? How did this improve or worsen their process? There are a lot of general statements, but not many specifics from the actual participant experience that is quantified or even qualified from interviews, etc.   
485 "…ensured that stakeholders were generally in agreement…" How so? How does that work or is even quantified?  
486 "…more likely to adopt the…" How do you know that?  
487 Why is this citation used here? It is on changes in adoption likelihood?  
493 How did stakeholders interact with these options? Did they change them? How so? Was this quantified somewhere - even narratively?  
494 Do these really affect the output? They just change the colors. They don't change the values.  
498-500 Good. Now did you quantify this change in understanding? Or even record narrative expressions of it? Otherwise, how can we trust this observation?  
503 "…stakeholders struggled to prioritize…" What do you mean? Was this quantified? What were they doing before? Did this new tool really improve this? What was the change in how priorities were made pre and post tool application?  
503-505 This is the kind of thing that would benefit from some data. How many? And did they change their minds? Is their longitudinal data on their decisions?  
506-507 I think you need a new section heading here because you leave the last discussion  
512 - 513 Do you mean that if an engineered channel is in a modified landscape? Because the example does not really follow…. And in Line 518 you say channel modification does not always results in degradation, but here you say it does? I think it is dependent on the landscape context. Just clarify.  
519-520 Okay, so how common are engineered channels in forested landscapes? I doubt that this is universally true…maybe a very small stretch in a forest. But, come on….  
545-547 This sentence doesn't say much in my opinion and we don't really know how much it improved the process.  
563 Did you try and quantify silviculture? Seems like there would be some CA state specific coverage OR it might be captured in the StreamCat variable on introduced vegetation classes.  
588-590 Oooo, that would be complex…different predictors for each quantile? Which makes me wonder, did all your predictors participate equally for every quantile prediction? Does quantile random forests generate the same variable importance information for each quantile? Might have been in SF  
623-628 You spend more than half the summary on the case study. How about focusing and summarizing the unique modeling aspects more thoroughly here and downplaying the case study, seeing as my comments above suggest we don't really have that much solid data on or learn all that much about that process.