*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.*

*We again thank the associate editor and first reviewer for providing helpful comments on our manuscript. Our responses to each of the comments are below.*

**Associate editor 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.

*All text about the case study in the results and methods was moved to a new subsection of the discussion “Case study: Application of the landscape model to the San Gabriel River watershed”. Further, all figures and tables for the case study were moved to the supplement, including substantial portions of the original text to shorten the main text in the discussion. We agree this is a good compromise for the case study that retains the valuable information, while not detracting from the main focus of the manuscript.*   
  
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.

*Please see our responses to 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 objectives were modified based on the suggestions in the detailed comments.*

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 following text was added to line 286: “This post-hoc classification was necessary to determine if observed CSCI scores were under- or over-scoring relative to landscape expectations and can serve to help prioritize management actions. For example, managers may choose to prioritize sites with index scores above or below the landscape model predictions differently than those that are within the expected range.”*

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?

*Many changes were made to the discussion based on the additional comments from the AE and reviewer 1. Please see our responses throughout.*

*In addition, we have added a paragraph comparing model performance to previous studies:*

*“Model performance was comparable to similar studies that focused on developing predictions of biological condition from geospatial data. Hill et al. (2017) developed a national model to predict stream site condition classes that correctly classified sites at about 75\% of locations, depending on region. Importantly, regionally specific models were more accurate than a single national model. For continuous predictions of biointegrity index scores, Carlisle, Falcone, and Meador (2009) developed a model for a large area of the eastern United States. Models for continuous data were able to correctly identify class membership from an a posteriori prediction at about 85\% of sites, which was similar in precision to models that were developed solely for categorical responses. For the landscape model herein, a comparison of the percentage of correctly classified sites that were above the 10th percentile of reference site scores (0.79) for observed data compared to predicted data showed that our model had comparable performance to other studies. The landscape model had 83\% predictive accuracy for classifying sites as altered (<0.79) or unaltered (>0.79) for the statewide results. However, the goal of our model was distinct from previous studies, such that our intent was not to predict bioassessment scores at unsampled locations, but rather to describe variation in scores as a function of land use to identify constraints. Interpretation of predictive accuracies between models should consider the differences in the goals for each model.”*

*A paragraph describing implications of the sensitivity analysis was also added: “Results in Figure 6 also demonstrate the broader implications of how the key decision points affected model results at regional and statewide scales. These results and the functionality provided by SCAPE demonstrate flexibility of the landscape model and the considerations that should be made for regional applications. For example, constraint classifications and the decision points that define them may have little relevance in regions without development gradients that are not well captured by the model (e.g., Sierra Nevada, North Coast). Conversely, the chosen range for the lower and upper expectation of biological integrity that is appropriate for a region is a tradeoff between which constraint classes are most appropriate for a region. Wider intervals force more stream segments into the "possible" constraint classes, whereas smaller intervals provide more separation of segments into the likely constrained or likely unconstrained classes. The specific choice is a management decision and we provide the ability to evaluate tradeoffs both in SCAPE and with our results herein.”*  
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.

*Please see our responses to the reviewer’s concerns.*   
  
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.

*We will defer to the editor on this issue. Although this concern was also raised by our science advisory panel, the suggestion was to clarify the distinction between our use of the term and that used by stream hydrologists, rather than to replace the term entirely. We address this concern early in the introduction and we feel this is an adequate clarification.*  
  
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?

*Sentence changed to “A landscape context is required that describes how likely a site is to achieve biological integrity, which can inform how bioassessment data supports decisions or identifies priorities.”*  
  
Line 72: Do you mean “restoring” streams? The rest of the sentence implies that restoration is the specific management goal.

*Yes, this was changed.*  
  
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.

*This sentence was removed.*  
  
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?

*Suggested changed were made.*  
  
Line 105: Topic sentence is a bit vague. The second sentence is probably a better topic sentence.

*Second sentence was moved to 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.

*The topic sentence was changed to “The goal of this study was to present the development and application of a landscape model to predict a lower and upper bioassessment score that would be expected based on land use.”*  
  
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.

*The objectives were changed based on the suggestions: “Our specific objectives were to 1) develop and validate the landscape model, 2) apply the model results to categorize all stream segments in California into constraint classes, and 3) provide a case study within a single watershed to demonstrate how model predictions and classifications can be used to prioritize management actions at a local scale.”*  
  
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.

*Text was edited as follows: “A CSCI threshold of 0.79, based on the tenth percentile of scores at all reference calibration sites for the original index, has been proposed as a threshold below which a site does not meet designated biological uses (SDRWQB 2016). As described below, the expected CSCI scores obtained from the landscape model were compared to this threshold to identify different constraint classes.”*  
  
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?

*Yes, that is the correct interpretation. Sentence was revised: “Preliminary analyses indicated that these variables produced a predictive model with comparable performance relative to a larger model with additional variables.”*  
  
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? *Changed sentence to: “These variables were chosen specifically as indicators of land-management activities that were most likely to limit the attainability of biological integrity.”*

Line 207: This sentence is also awkward and unclear.

*Sentence was revised: “Landscape variables were preferred over in-stream data because landscape stressors can be more challenging to manage, and we wanted to quantify biological impacts relative to these challenges.”*  
  
Line 210: Is there a geospatial indicator of channel modification? Seems unlikely. Then why bring up the topic? Sentence is also unclear and awkward.

*Initial feedback from our stakeholder group suggested that we not focus only on modified channels because limits on biological integrity are not completely described by channel engineering. Although biological communities in modified channels are often constrained (based on our definition), they are not always constrained. Landscape predictors provided a more inclusive description of the problem. The sentence was modified: “Further, presence or absence of channel modification was not used to quantify limits on biological integrity because landscape predictors were more broadly inclusive of the problem (e.g., modified channels are often but not always constrained, constrained channels are not always modified).”*  
  
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?

*That is correct, sentence was revised: “Overall, the model was associative by design and was intended as a predictive tool that does not describe specific mechanisms of biological alteration.”*  
  
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.

*We agree that an evaluation of the median only partially addresses the range of predictions provided by the quantile models. There are indeed measures of fit for assessing precision of quantile predictions (see Koenker, R., & Machado, J.A.F. (1999). Goodness of fit and*

*related inference processes for quantile regression. Journal of the*

*American Statistical Association, 94(448), 1296–1310), but we are unaware of comparable methods for random forest applications. Although this a potential concern, we do emphasize in the manuscript that the lower and upper bounds on the expected range of scores are flexible and can be varied based on preference or application. The acceptable range is less a science question and more at the discretion of who is applying the results for decision-making. Our sensitivity analysis was meant to provide some insight into how these decision points affected these results.*   
  
Line 261: I think Figure 3 is too busy. Can you dramatically simplify the figure to show just the key explanations of your classification?

*Figure 3 was simplified.*   
  
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.

*Dropped.*  
  
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?

*This final classification scheme is strictly categorical to describe how an observed CSCI score compares to the expected ranges from the model. More importantly, your point about classifying calibration sites is well taken. A simple comparison of the percentage of sites as calibration or validation that were placed in each category suggests there was no bias in applying this scheme to the calibration data:*

|  |  |  |
| --- | --- | --- |
| Site type | Category | Percent of sites |
| Cal | under scoring | 10.3 |
| Val | under scoring | 10.8 |
| Cal | expected | 79.6 |
| Val | expected | 81.3 |
| Cal | over scoring | 10.1 |
| Val | over scoring | 7.85 |

*There are roughly equal percentages of sites for calibration and validation in each category. The analysis was repeated for each region and the results were similar.*

*We also note that the model predictions for the expected range of scores were obtained using the out-of-bag predictions from the random forest models. This eliminated any bias comparing observed CSCI scores for the calibration dataset to those from the model. We have added txt to the methods to make this clear:*

*“All predictions for the calibration dataset were obtained using out-of-bag estimates from the random forest model to prevent bias and over-fitting. Out-of-bag predictions are based on the subset of trees in the random forest model in which calibration data were excluded during training (Mazor et al. 2016; Meinshausen 2017).”*

*The text of this paragraph was also modified: “A categorization scheme was developed for sites where biomonitoring data were available to compare observed CSCI scores to the range of expected scores from the model (Figure 3d). This post-hoc classification was necessary to determine if observed CSCI scores were under- or over-scoring relative to landscape expectations and can serve to help prioritize management actions. For example, managers may choose to prioritize sites with index scores above or below the landscape model predictions differently than those that are within the expected range. Sites with observed scores…”*

NOTE: I am recommending that material from line 291-333 be moved to SI.

*See comments above*.  
  
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.  
*This paragraph was restructured and moved to the new section about the case study in the discussion: “Results from the statewide model were applied in a regional context through local application with a stakeholder group from the San Gabriel River watershed (Los Angeles County, California, Figure S4). The statewide model provides only a range of expected scores for a stream segment. Comparison of observed index scores from an actual biological sample with the results from the model can establish a basis for how managers prioritize sites. For example, managers may prioritize sites with observed scores that are above the modelled expectation differently than those that are scoring within the ranges predicted by the model. Alternatively, a site scoring as expected in an unconstrained segment could be prioritized differently than a site scoring as expected in a constrained segment. As such, the lower San Gabriel watershed is heavily urbanized with many modified channels and managers require prioritization tools to identify where efforts should be focused in the context of landscape development. Information from the landscape model allowed the stakeholder group to develop management priorities based on how actual CSCI scores compared to biological expectations from the model (Figure S5).”*

Line 312: Not sure what a “spreading ground” is.

*This was moved to the supplement, but we have also clarified what this means: “Groundwater recharge areas are present in the middle of the watershed where water is allowed to spread beyond the main channel for subsurface infiltration during high flow events.”*  
  
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?

*Stakeholders first identified their priorities independent of the landscape model and then applies these priorities based on how observed scores compared to modelled expectations. This was clarified:*

*“Management priorities for individual sites that were important for the stakeholder group included the following actions (Table S1, Figure S6):*

*\* Investigate: Conduct additional monitoring at a site or review of supplementary data (e.g., field visits, review aerial imagery);*

*\* Protect: Recommend additional scrutiny of any proposed development and/or projects that could affect a site;*

*\* Restore: Pursue targeted action for causal assessment and/or restoration activity at a site.*

*These priority actions were identified independently from the landscape model and assigned to each site by the stakeholder group based on a comparison of observed CSCI scores and the expected range of scores from the landscape model.”*  
  
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.  
*This sentence was moved to the supplement, but it was revised for clarity: “A template that showed how observed CSCI scores could compare (i.e., under-scored, expected, over-scoring, or above/below biological objective) to segment classifications (i.e., constrained, unconstrained) was provided to the stakeholder to assign priorities among the various outcomes (rows 1-16, Figure S2, left side) that could occur with actual data.”*

Line 349: Simplify to: “There was generally good agreement between observed and predicted CSCI statewide”

*Changed.*  
  
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.”

*Changed.*  
  
Line 353: “Performance was similar with the validation dataset (r=0.72, RMSE=0.18)”.   
Why no slope for the validation dataset?

*Changed and added intercept/slope.*  
  
Line 364: Provide r value for Sierra region, as you have done for other regions.

*Added.*  
  
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.”

*Correct, sentence was changed.*  
  
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.

*The addition to the methods in the response above provides some clarity on why these results are presented (e.g., “A categorization scheme was developed for sites where biomonitoring data were available to compare observed CSCI scores to the range of expected scores from the model (Figure 3d). This post-hoc classification was necessary to determine…”).*

*Also please see the response above about the comparison between calibration/validation sites and out-of-bag estimates.*  
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.

*Yes, please see the addition to the discussion about the implications of these results.*  
  
Lines 399-425: NOTE that I’m recommending most of this material be moved to the SI.  
*Moved to supplement or discussion.*

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.

*Changed to “…tools that provide an estimate of the range of attainable conditions relative to the landscape.”*  
  
Lines 441-446: These sentences are vague and filled with jargon (e.g., temporal and spatial scales, watershed scales, etc.)  
*Text was modified: “The landscape model can place observed scores in an appropriate context relative to their expected condition for the landscape. This information could provide flexibility in the selection of regulatory or management actions at specific sites or within larger regions (e.g., hydrologic subareas), and to further prioritize where and when actions should take place based on the resources needed for protection or restoration actions.”*

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.   
*Modified for clarity: “The landscape model could also help identify where tiered aquatic life uses (TALU, Davies and Jackson 2006) may be needed. However, the model is not intended, nor is it is sufficient, as a standalone tool for this purpose because it lacks specificity as to what uses may apply under different landscape conditions.”*

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.

*See response to next comment.*  
  
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.

*The case study methods and results were shortened and moved to this section. The new subheading is “Case study: Application of the landscape model to the San Gabriel River watershed”.*

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.

*We feel it is critical to emphasize that our model does not just simply describe channel modification, so we have retained most of the text here. However, we have moved the description and figure for Tecolote Creek to supplement to reduce some of the detail in this paragraph.*   
  
Line 585: You already made this point. No need to repeat it here, or perhaps you can remove it from its earlier location.

*Sentence was removed.*  
  
**Reviewer 1 comments:**

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.

*We appreciate the comments regarding the case study and have taken them into consideration for our revisions. As noted by the AE, there was disagreement about the importance of the case study. We have taken the advice of the AE and moved all content from the methods and results about the case study to supplemental material and have also included a shortened version of relevant details as a subsection in the discussion. We hope this is an adequate compromise.*

*However, your comments about the specific language we used to describe the importance of the case study are well taken. We agree they are anecdotal at times and have taken measures to add clarity in the descriptions. Please see some of our responses to specific comments below.*

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.  
*We agree that the use of “prediction interval” to describe the quantile predictions was technically incorrect. A more appropriate description is a range of predictions for the lower and upper limits of the conditional distribution. We have changed all instances in the text that describe prediction interval to a more appropriate description. For example, “This modelling approach can estimate a lower and an upper limit for the conditional distribution of likely scores that might be expected at a site given land use…” or “prediction range” instead of “prediction interval”.*

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"

*Changed.*

44 Strike "that were", strike "clear"

*Removed.*

64 Strike "place" change to "limit the…"

*Changed.*

65 2x negative. Changed to Resource management decisions might be improved if information were…

*Changed.*

79 Sentence beginning "Although…" could use citation

*Added citation to Bernhardt et al. 2007. Bernhardt, E. S., E. B. Sudduth, M. A. Palmer, J. D. Allan, J. L. Meyer, G. Alexander, J. Follastad-Shah, et al. 2007. “Restoring Rivers One Reach at a Time: Results from a Survey of U.S. River Restoration Practitioners.” Restoration Ecology 15 (3): 482–93.* [*https://doi.org/10.1111/j.1526-100X.2007.00244.x*](https://doi.org/10.1111/j.1526-100X.2007.00244.x)*.*

81 "…integrity have been…"

*Changed.*

84 "designation" no -s

*Changed.*

92 "….could be prioritized at less constrained sites where….

*Changed.*

94 "…higher management priority (i.e., for protection) relative to a site that is scoring within the expected ranged based on landscape development."

*Changed.*

96-98 I do not understand what is being said in this sentence. Are you talking random site effects? Maybe re-read and clarify.

*This sentence was revised for clarity: “A predictive model of bioassessment scores that is based on landscape metrics (e.g., imperviousness) could describe constraints on biological integrity, particularly for factors that are difficult to manage and are often associated with instream stressors.”*

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.

*The DPSIR model provides a useful conceptual foundation for this work. It is again referenced in the method when we discuss our choice of predictors.*

119 "…scores that are likely given any landscape context."

*Not sure “any” applies here, since we used specific predictors for urban and agricultural land use.*

120-121 It might be nice, in place of DPSIR, to have a visual conceptual model of how your process works.

*See above comment.*

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.

*See comments above about the case study. But we agree that the value or benefit of the case study and what it addresses should be mentioned. We added an additional sentence for clarification: “Managers currently have no prioritization tools for evaluating the context of biological integrity scores in their watershed.”*

151 Is stream hydrography a stretch? The closest thing you ended up using was canal density. Is that hydrographic?!?

*True, we did not use hydrography as a predictor. However, this statement was meant to indicate we used the NHDPlus data as our base layers for developing the model. We clarified this in the sentence.*

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.

*Changed.*

183 Replace although with while

*Replaced.*

190 Ode et al.

*Changed (also verified other citations with three authors).*

193 StreamCat were - data is plural. You may want to check this throughout.

*Changed.*

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.

*Agreed, this statement is somewhat of a stretch. In hindsight, the StreamCat data were created from multiple layers, each with different dates (e.g., NLCD 2006, 2011). So, it really does not make sense to choose the sample data closest to the StreamCat dates since there is no “closest date”. This statement was removed.*

208 …given that our focus was on constraints to biological condition typically beyond the scope….

*This sentence was modified in response to AE comments.*

212 "…,whereas modified channels are not landscape scale measures."

*Sentence was modified in response to AE comments.*

213 You do focus on ultimate vs proximate causes. I think that is a fine angle.

*True, but it is not a mechanistic model. The sentence was revised as follows: “Overall, the model was associative by design and was intended as a predictive tool that does not describe specific mechanisms of biological alteration.”*

222 What are "robust predictions"?

*This is stats jargon. The sentence was revised: “Random forest models can quantify complex…”*

223 MLR can model complex, non-linear relationships with interactions. So what do you mean?

*Sentence was revised: “Random forest models can quantify complex, non-linear relationships and interactions between variables and can be more effective models with large data relative to more commonly-used approaches, such as …”*

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.

*See response above.*

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).

*Changed to “conditional quantile”.*

240-246 Is there no way to do this without binning? It seems off that there is no continuous solution…

*Perhaps there is a more elegant alternative, but we are confident the approach provided a good representation of development gradients in each region (where the gradients were observed). We have used similar approaches in other studies (e.g., Mazor et al. 2016) and have found the approach to be sufficient.*

242 "…on a random draw of sites from strata of quartiles defined by…"

*Changed.*

244 "…landscape development among regions (i.e.,…" Between is typically two objects and among is for >2, I think…

*Changed.*

246 The remaining 25% of sites were used…. Right? Since you mentioned 75% above.

*Added.*

252 …indicated food predictive ability.

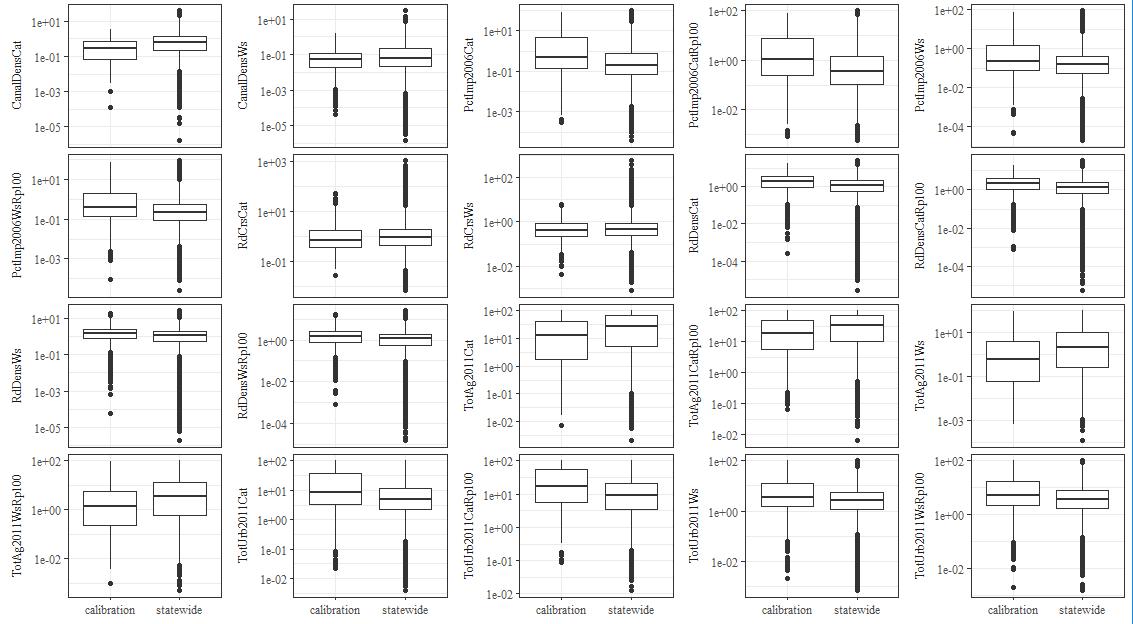
*Changed.*

254-256 I would cut this. Either start with this sentence or drop. I think its redundant.

*Sentence was moved to the start of the paragraph.*

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.

*We had not thought of this issue in our development process, but you raise a good point. Simple boxplot comparisons of the distribution of observations for the predictors between the statewide and calibration dataset shows that the two are similar (below). We also ran individual t-tests for each variable (unequal variance to account for the largely different sample sizes), as well as a multivariate version of the t-test to compare the multivariate means between the groups (Hotelling’s T-squared test). The t-test comparisons indicated that most but not all were significantly different, and the results were confirmed for the Hotelling’s test. However, the power of these tests is incredibly high given the sample sizes (e.g., > 100000 observations in the statewide dataset), so perhaps these tests are at risk of false positive results. Looking at the boxplots, the ranges in values between the datasets are not incredibly different, so we have confidence that we are not over-extrapolating (beyond reason) from our calibration dataset.*



*We have added a sentence on line 260 that addresses this concern: “Ranges in predictor variables between the statewide and calibration datasets were similar, such that over-extrapolation of the model domain to the statewide data was unlikely.”*

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?

*Typo fixed, removed content about reference calibration sties (agree that this is confusing since these are different calibration sites from those in the current manuscript), and no longer describe the ranges as a prediction interval.*

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."

*Changed.*

273 "…depend on the percentile range of score…."

*Changed.*

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.

*The following sentences (lines 275-278) that were added in response to the first round of revisions address this concern. We feel that it is useful to describe these ranges as “certainty” from the context of how acceptable a user might consider a range of expectations to be for a particular application.*

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…

*We agree that some of this is indeed self-revelatory, but we think the analysis provides an indication of how much change might be expected with key differences between regions. The direction of the changes was expected based on how we constructed the model, but the important differences were variation in the changes between regions based on land use differences. We have added a sentence for clarity: “Although some of the results can be assumed (e.g., increasing CSCI thresholds causes more sites to be classified as constrained), we expected differences among regions based on differences in land use.”*

*Note that content for lines 291 to 346 was moved to the discussion or supplement.*

298 How so? How would it be prioritized differently? Is that based on experience or interviewing*? Sentence was modified: “Alternatively, a site scoring as expected in an unconstrained segment could be a higher priority for managers than a site scoring as expected in a constrained segment; the latter may require more resources for comparable improvements in biotic condition.”*

300-301 You lost me here.

*Separated into two sentences for clarity: “The statewide model provides only a range of expected scores for a stream segment. Comparison of observed index scores from an actual biological sample with the results from the model can establish a basis for how managers prioritize sites.”*

309 I think this should be Figure 4b.

*Changed (but now S4b).*

329-331 You lost me here too

*This statement was originally made to clarify that sites were not simply ignored if they were not assigned a priority (restore, investigate, or protect). It was assumed that any routine monitoring or baseline maintenance that occurred at these sites was to continue. This content was moved to the supplement but revised anyway: “For sites without priority assignments, it was assumed that baseline monitoring and maintenance that is currently provided by existing management programs was sufficient for maintaining current biological condition.”*

332-333 I think you already referenced this table above (line 321).

*Yes, but now the citation is only in the supplement so it’s retained.*

350 "…and the predicted medians (r=0.75, RMSE = 0.17; Table 4, Figure S4).

*Text was modified in response to AE comments.*

351 Scratch this sentence. See line above. Just put in parenthetical.

*See response to last comment.*

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?

*There are published methods for evaluating model fit for conditional quantile models (Koenker and Machado 1999. Journal of the American Statistical Assocation. 94:1296:1310). These are similar to pseudo-R2 values for generalized linear models that compare explained deviance (residual variation) to null deviance (unconditional variance). However, the published methods are described for conventional quantile regression and we are unaware of similar methods for quantile regression forests, nor if the former can be meaningfully applied to the latter. For now, we will include only the median performance as is, but we agree that better understanding the performance at conditional quantiles could be informative for future analyses.*

353 "…suggesting minimal median prediction bias…"

*Changed.*

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.

*Changed to “worse performance”.*   
Also, I think you should use your validation results more - that is a more independent, true test of performance, is it not?

*Yes, we agree, but we present both calibration and validation results. We’re not sure what you mean by “use validation results more”.*

367 You could show p values for the slope and intercepts.

*Results for the regression in the supplementary figure are provided in Table 3 in the main text. It is noted in the caption that all p-values are significant for all correlations, intercepts, and slopes. To reduce clutter, we opted not to show the p-values on the table since they were not very informative.*

374 I bet these results are pretty close to what you'd get if you just use percent urban and percent ag….

*Agreed because these were the primary land use gradients we were attempting to describe. A more comprehensive analysis of variable importance could inform further model development to identify the most parsimonious model. We are comfortable with our results as is.*

377 "…within the decile range as often…."

*Changed to “…within the predicted decile range as often….”*

381 Replace "caused by" with "evidence of"?

*Changed.*

382-382 Really? Is the CH underscored @ 13%? I am not sure these statements are consistent across results.

*Rephrased for clarity: “Over-scoring sites were slightly more common in the South Coast and Sierra Nevada regions, whereas under-scoring sites were more common in others (i.e., the Chaparral, Central Valley, and Desert/Modoc regions).”*

386-392 Again, is this all that insightful? It is kind of self-revelatory…

*Yes, but we highlighted the important results as the differences between regions in the following sentences (lines 391 – 398). We feel it’s important to first state the obvious, but perhaps over-looked, results in lines 386-392 as a precedent to understanding the results in the following lines.*

*Note that content for lines 399 – 425 was moved to the discussion or supplement.*

404 "…lower watershed had (predicted or actual) median CSCI scores…." Don't you mean actual observed CSCI scores? I was confused.

*No, this refers to the skew of the quantile predictions. Sentence was revised for clarity: “…had predicted median CSCI scores that were very close to the 10th percentile (i.e., right-skewed quantiles)…”*

407 How was "effectively used" quantified or even qualified? Versus what? What alternatives were tested or compared? Did you interview stakeholders?

*This sentence was removed.*  
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.

*No, there was no before/after comparison, nor was there anything to compare the decisions to. We hope that our revisions to the case study have clarified the role of the landscape model for our local group. Prior to the landscape model, the stakeholders had no means of prioritizing among sites that were differently affected by land use. As such, the model has inherent value because it fills an important information gap.*

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?

*In addition to our above comments regarding the case study, the real value is demonstrating how the statewide results can be applied to inform decision-making at the local-scale. This is stated in several instances in the text (including the revised objectives per recommendations of the AE).*

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.

*We hope that by moving the case study from the methods/results to discussion/supplement have helped address some of your concerns. We do agree that the results could be reinforced through more quantitative measures of success, but we still feel that our experience with this local group is a valuable contribution to the article.*

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.  
*Agreed, sentence was revised: “Management activities for biological integrity could involve the protection of sites meeting or exceeding biological objectives or the restoration of sites that have the potential to meet or exceed biological objectives.”*

457-458 Why? I am interested in your defense.

*Sentence was revised in response to AE: “However, the model is not intended, nor is it is sufficient, as a standalone tool for this purpose because it lacks specificity as to what uses may apply under different landscape conditions.”*

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.

*See responses above regarding the case study. We have also modified much of the text in this paragraph to remove some of the ambiguous and/or qualitative claims.*

485 "…ensured that stakeholders were generally in agreement…" How so? How does that work or is even quantified?

*Modified the text: “The final decision by the group to prioritize management actions for the different sites in broad categories of protect, restore, and investigate was based on group discussions to reach agreement on outcomes from the models. Facilitated discussions that directly engage stakeholders have been suggested by others as effective mechanisms that allow recommendations provided by these tools to be adopted in formal decision-making (Stein et al. 2017).”*

486 "…more likely to adopt the…" How do you know that?

*See response to above comment.*

487 Why is this citation used here? It is on changes in adoption likelihood?

*This paper provides a case study that demonstrates effective engagement of stakeholders, similar to our example.*

493 How did stakeholders interact with these options? Did they change them? How so? Was this quantified somewhere - even narratively?

*Sentence was revised for clarity: “The SCAPE application can be used to select and visualize management priorities…”*

494 Do these really affect the output? They just change the colors. They don't change the values.

*Yes, changing the range of expectations directly changes the classification for any given segment, which in turn can affect the interpretation of an observed score to the classification, which finally can affect the priority assignment.*

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?

*No, but the sentence was revised to clarify what this meant: “…currently under review in California, such as the effect of changing a potential threshold for defining biological use attainment and how the assigned priorities shift accordingly.”*

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?

*The “struggle to prioritize” comes from a lack of context shown in the left side of Figure S5. We modified this statement for clarity: “Without the landscape context provided by the model (i.e., Figure S5, right side), stakeholders struggled to prioritize among sites (i.e., no context for scores, Figure S5, left side).”*

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?

*This statement was removed.*

506-507 I think you need a new section heading here because you leave the last discussion

*This content was moved to the subsection “alternative applications of the landsape model”.*

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.

*Clarified: “…but an engineered channel in a developed landscape will typically be constrained.”*

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….

*We have no reliable data in California that describes where channel engineering has occurred. Although this is obviously a larger concern for urban landscapes, the example in Stein et al. (2013) highlights an important exception that channel modification does not always relate 1:1 to biological alteration. We have also observed this in urban landscapes, but Stein et al. is the only publication we know of where this has been documented in the literature.*

545-547 This sentence doesn't say much in my opinion and we don't really know how much it improved the process.

*Sentence was revised: “Our case study provided an example of how our model helped establish priorities at the local-scale and a similar process could be used for applying different landscape models in other states.”*

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.

*Although there may be some datasets available, they are likely limited in scope and insufficient for calibration of a regional model. StreamCat surrogates may also be possible and this is an avenue worth exploring. The text was revised to suggest this approach as a future research effort: “Accurate data for quantifying these potential stressors are not explicitly available in StreamCat, but surrogates could be explored in future models (e.g., coverage of introduced vegetation classes as a proxy for silviculture). Regardless, investments in improving spatial data could yield significant improvements in further development of bioassessment indices and tools for their interpretation.”*

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

*We have received more than one comment about the novelty of our approach using quantile random forests to develop the landscape model. There are certainly many interesting applications that could be explored beyond our initial approach herein. Your questions are of course warranted and could be applied in future work. Please see our response below.*

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.

*We have added some content to the summary that focuses on the unique aspects of our modeling approach and presents some suggestions for future work: “We demonstrated the use of quantile regression forests to successfully predict a lower and upper range of expected biological index scores that could be observed at a stream segment as a result of landscape development. Although random forest models have been increasingly used in bioassessment applications, our approach is the first to use quantile models to develop biological expectations. As such, additional work could build on this initial approach to apply these models in different locations, to alternative biological response endpoints, or to explore different predictors that capture regionally-specific stressor gradients. The predictive performance of quantile regression forests in bioassessment applications have also not been fully explored, such as understanding the accuracy of predictions or the relative importance of predictors at different quantiles. Our approach suggests these models are promising and future work could focus on the above suggestions to better understand the utility of these tools in applied contexts.”*