Freshwater Science

Prioritizing management goals for stream biological integrity within the developed landscape context

--Manuscript Draft--

Manuscript Number:	2019026R1
Full Title:	Prioritizing management goals for stream biological integrity within the developed landscape context
Short Title:	Stream priorities in developed landscapes
Article Type:	Regular
Corresponding Author:	Marcus Beck Southern California Coastal Water Research Project Costa Mesa, CA UNITED STATES
Corresponding Author's Institution:	Southern California Coastal Water Research Project
First Author:	Marcus W Beck
Order of Authors:	Marcus W Beck
	Raphael D Mazor
	Scott Johnson
	Karin Wisenbaker
	Joshua Westfall
	Peter R Ode
	Ryan Hill
	Chad Loflen
	Martha Sutula
	Eric D. Stein
Order of Authors Secondary Information:	
Manuscript Region of Origin:	UNITED STATES
Abstract:	Stream management goals for biological integrity may be difficult to achieve in developed landscapes where channel modification and other factors constrain in-stream conditions. To evaluate potential constraints on biological integrity, we developed a statewide landscape model for California that estimates ranges of likely scores for a macroinvertebrate-based index that are typical at a site with the observed level of landscape alteration. This context can support prioritization decisions for stream management, like identifying reaches for restoration or enhanced protection based on how observed scores relate to the model expectations. Median scores were accurately predicted by the model for all sites in California with bioassessment data (Pearson correlation r = 0.75 between observed and predicted for calibration data, r = 0.72 for validation). The model also predicted that 15% of streams statewide are constrained for biological integrity within their present developed landscape, particularly for urban and agricultural areas in the South Coast, Central Valley, and Bay Area regions. We worked with a local stakeholder group from the San Gabriel River watershed (Los Angeles County, California) to evaluate how the statewide model could support local management decisions. To achieve this purpose, we created an interactive application, the Stream Classification and Priority Explorer (SCAPE), that compares observed scores with expectations from the landscape model to assign priorities. We observed model predictions consistent with the land use gradient from the upper to lower watershed, where potential limits to achieving biological integrity were more common in the heavily urbanized lower watershed. However, most of the sites in the lower watershed scored within their expected ranges, and were therefore given a low priority for restoration. In contrast, two low-scoring sites in the undeveloped upper watershed were prioritized for causal assessment and possible future

	restoration, whereas three high-scoring sites were prioritized for protection. The availability of geospatial and bioassessment data at the national level suggests that these tools can easily be applied to inform management decisions at other locations where altered landscapes may limit biological integrity.
Suggested Reviewers:	Andy Rehn California Department of Fish and Wildlife Andy.Rehn@wildlife.ca.gov
	Terry Flemming USEPA Fleming.Terrence@epa.gov
	Kai Chen Nanjing University ckai2005@gmail.com
	William Bouchard MN Pollution Control Agency will.bouchard@state.mn.us
Opposed Reviewers:	Lester Yuan USEPA yuan.lester@epa.gov already reviewed
Response to Reviewers:	

SOUTHERN CALIFORNIA COASTAL WATER RESEARCH PROJECT
A Public Agency for Environmental Research

May 16th, 2019

Dr. Charles Hawkins
Chief Editor
Freshwater Science

I am pleased to resubmit our manuscript, “Prioritizing management goals for stream biological integrity within the developed landscape context,” to be considered as an original research article in Freshwater Science.

We again appreciate the substantial comments on our submission provided by our handling editor, Dr. Daren Carlisle, and those from our first reviewer. We have made substantial revisions to the manuscript to address these concerns and a point by point response is provided with our re-submission. In short, we have taken the advice to move the case study to the discussion and have moved all relevant figures and tables to the supplement, including extraneous text. We have also replaced any use of the term “prediction interval” with a more appropriate description, as recommended by reviewer one. There are also edits to the discussion and conclusion to place our results in a broader context, including implications of the sensitivity analysis, comparison with previous studies, and next steps.

Our organization agrees to submit payment for page charges if the paper is published. We appreciate the opportunity to publish our work in FWS.

Marcus W. Beck (marcusb@sccwrp.org), 714-755-3217
S. California Coastal Water Research Project
3535 Harbor Blvd. Suite 110
Costa Mesa, CA 92626

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, which can 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 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 et al. (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 captured well by the model (e.g., Sierra Nevada, North Coast). Conversely, the chosen range for the lower and upper expectation of biological integrity 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 be used to identify 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 changes 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 at a stream reach 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 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?
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 text 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 applied 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 then 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-scoring, 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.

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.
We feel 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 with large datasets 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.

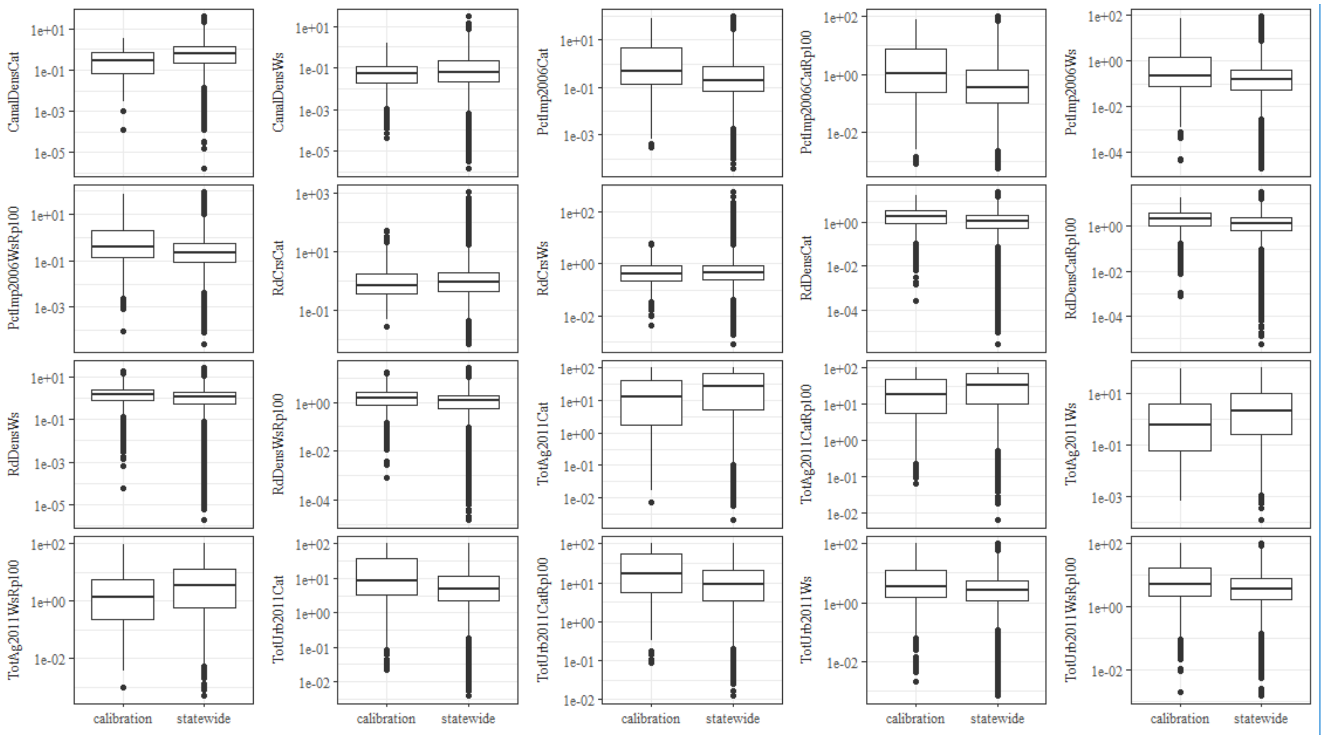
244 "…landscape development among regions (i.e.,…" Between is typically two objects and among is for >2, I think…
Changed.

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 sustaining 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. Also, please see our addition to the conclusions that suggests next steps for quantile regression forests.

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, but please see our addition to the conclusions on next steps for further developing these models.

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 how outcomes from the model could be applied. 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 had limited information 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 California 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 present 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."

Running head: Stream priorities in developed landscapes

Prioritizing management goals for stream biological integrity within the developed landscape context

Marcus W. Beck (marcusb@sccwrp.org)[1*]

Raphael D. Mazor (raphaelm@sccwrp.org)[1]

Scott Johnson (scott@aquaticbioassay.com)[2]

Karin Wisenbaker (karin@aquaticbioassay.com)[2]

Joshua Westfall (jwestfall@lacsd.org)[3]

Peter R. Ode (peter.ode@wildlife.ca.gov)[4]

Ryan Hill (hill.ryan@epa.gov)[5]

Chad Loflen (Chad.Loflen@waterboards.ca.gov)[6]

Martha Sutula (marthas@sccwrp.org)[1]

Eric D. Stein (erics@sccwrp.org)[1]

[*] Corresponding author

[1] Southern California Coastal Water Research Project, 3535 Harbor Blvd., Costa Mesa, CA 92626

[2] Aquatic Bioassay & Consulting Laboratories, Inc. 29 North Olive, Ventura, CA 93001

[3] Sanitation Districts of Los Angeles County, 1955 Workman Mill Road, Whittier 90601

[4] California Department of Fish and Wildlife, Office of Spill Prevention and Response, 2005 Nimbus Road, Rancho Cordova, CA 95670

[5] Department of Forest Engineering, Resources, and Management, Oregon State University c/o Western Ecology Division, 140 Peavy Hall, 3100 SW Jefferson Way, Corvallis, OR 97333

[6] San Diego Regional Water Quality Control Board, Healthy Waters Branch, Monitoring and Assessment Research Unit, 2375 Northside Drive, San Diego, CA 92108

Abstract

Stream management goals for biological integrity may be difficult to achieve in developed landscapes where channel modification and other factors constrain in-stream conditions. To evaluate potential constraints on biological integrity, we developed a statewide landscape model for California that estimates ranges of likely scores for a macroinvertebrate-based index that are typical at a site with the observed level of landscape alteration. This context can support prioritization decisions for stream management, like identifying reaches for restoration or enhanced protection based on how observed scores relate to the model expectations. Median scores were accurately predicted by the model for all sites in California with bioassessment data (Pearson correlation r = 0.75 between observed and predicted for calibration data, r = 0.72 for validation). The model also predicted that 15% of streams statewide are constrained for biological integrity within their present developed landscape, particularly for urban and agricultural areas in the South Coast, Central Valley, and Bay Area regions. We worked with a local stakeholder group from the San Gabriel River watershed (Los Angeles County, California) to evaluate how the statewide model could support local management decisions. To achieve this purpose, we created an interactive application, the Stream Classification and Priority Explorer (SCAPE), that compares observed scores with expectations from the landscape model to assign priorities. We observed model predictions consistent with the land use gradient from the upper to lower watershed, where potential limits to achieving biological integrity were more common in the heavily urbanized lower watershed. However, most of the sites in the lower watershed scored within their expected ranges, and were therefore given a low priority for restoration. In contrast, two low-scoring sites in the undeveloped upper watershed were prioritized for causal assessment and possible future restoration, whereas three high-scoring sites were prioritized for protection.

The availability of geospatial and bioassessment data at the national level suggests that these tools can easily be applied to inform management decisions at other locations where altered landscapes may limit biological integrity.

Key words: Bioassessment, biotic integrity, streams, urbanization, modified channels, landscape stressors, random forests, prioritization, data visualization, stakeholder group

Introduction

The widespread use of bioassessment data to assess ecological condition of aquatic environments is a significant advance over chemical or physical methods of assessment, yet managers and stakeholders require contextual information for synthesizing and interpreting biological information. The reference condition concept that is built into many biological indices provides a broad context for observed condition relative to unaltered habitats for a particular region (Reynoldson et al. 1997; Stoddard et al. 2006). However, achieving a reference condition of biological integrity (i.e., having structure and function comparable to natural habitat for the same region, Karr et al. 1986) may be challenging if landscape conditions (e.g., watershed imperviousness) limit the spatial and temporal scales that can be effectively managed (Chessman and Royal 2004; Chessman 2014). Resource management decisions could be improved if information is available that describes these limitations. 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 be used to identify priorities.

Prioritizing among sites that are affected by landscape alteration is a critical challenge for managers in urban and agricultural settings (Walsh et al. 2005; Beechie et al. 2007; Paul et al.

2008). In developed landscapes, the majority of stream miles are in poor biotic condition (USGS 1999; Finkenbine et al. 2000; Morgan and Cushman 2005). Restoring streams in urban or agricultural settings can be costly, success is not universally defined, and achieving regional reference-like conditions may be difficult (Bernhardt et al. 2007; Kenney et al. 2012; Shoredits and Clayton 2013). Conventional approaches to protect and restore biological integrity have commonly focused on direct improvements at the site level to mitigate instream stressors (Carline and Walsh 2007; Lester and Boulton 2008; Roni and Beechi 2012; Loflen et al. 2016), in addition to upstream preventive measures that may be incentivized or enforced through regulation. Although these approaches can lead to improvements in ecological condition(e.g., Bernhardt et al. 2007), there is no universal remedy for achieving biological integrity in streams. In urban areas, protective thresholds for biological integrity have been debated (Cuffney et al. 2011). For biological integrity, several states have implemented a tiered aquatic life use or alternative use designation as potential approaches to account for baseline shifts in ecosystem condition from channel modification (e.g., FDEP 2011; USEPA 2013; MBI 2016; permitted under section 303(c)(2) of the Clean Water Act). Other approaches may include site-specific criteria or alternative thresholds with implementation clarifications (e.g., SDRWQB 2016).

Herein, we define constrained streams as those where present landscapes are likely to limit management options for restoring biological integrity. This definition describes a biological expectation and is distinct from the classical definition used in the general stream ecology literature (e.g., a physically constrained channel in the morphological sense). By describing an expected range of biological conditions due to factors that constrain biointegrity and may be difficult to manage, efforts to improve or protect condition could be prioritized at less constrained sites where alternative or more easily managed factors are influencing condition. For example, a monitoring site with an observed biological index score that is above an expected range of attainability could be assigned a higher management priority (i.e., for protection) relative to a site that is scoring within the expected range based on landscape development. 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. Analysis methods that characterize biotic and abiotic factors that limit assemblage composition have been explored by others (i.e., limiting factor theory, Chessman et al. 2008; Chessman 2014). Similar concepts have been applied in a landscape context to describe variation in biological communities and metrics at different spatial scales (Waite 2013; Waite et al. 2014), although they have not been developed to describe constraints as defined above.

Consistent and empirical links between land use thresholds and poor biotic integrity have been identified in many cases (Allan et al. 1997; Wang et al. 1997; Clapcott et al. 2011) and previous modelling efforts have successfully used geospatial data to predict stream condition at regional or national scales (Vølstad et al. 2004; Carlisle et al. 2009; Brown et al. 2012; Hill et al. 2017). Many of these models are based on the understanding of relationships between stream condition and watershed characteristics (Hynes 1975; Johnson et al. 1997; Richards et al. 1997), which can be broadly conceptualized within the Driver-Pressure-Stress-Impact-Response (DPSIR) framework that describes relationships between the origins and consequences of environmental problems (Smeets and Weterings 1999). However, past efforts have primarily focused on characterizing condition at unsampled locations, often predicting the most likely condition by estimating averages. Alternative modelling approaches, such as quantile-based methods (e.g., Cade and Noon 2003), could be used to predict a range of expectations for biotic integrity from geospatial data. This approach differs fundamentally from previous efforts of estimating average condition by providing an estimate of the minimum and maximum scores that are likely given the landscape context. Once the responses of macroinvertebrate communities to landscape changes at large spatial scales are understood, expectations can be compared to field samples and sites can be prioritized by local managers based on deviation from the expectation.

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 at a stream reach based on land use. 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. The model was developed and applied to all streams and rivers in California, specifically focusing on the potential of urban and agricultural land use to constrain biological condition. The case study demonstrated how the statewide model could be used to classify and prioritize at the regional scale using guidance from a local stakeholder group from a heavily urbanized watershed. Managers currently have no prioritization tools for evaluating the context of biological integrity scores in their watershed. An interactive software application, the Stream Classification and Priority Explorer (SCAPE), was developed for our case study to help stakeholders choose regional management priorities from the statewide landscape model.

Methods

Study area and data sources

California covers 424,000 km^2 of land with extreme diversity in several environmental gradients, such as elevation, geology, and climate (Figure 1a, Ode et al. 2016). Temperate rainforests occur in the north (North Coast region, NC), deserts and plateaus in the northeast and southeast (Deserts and Modoc Plateau region, DM), and Mediterranean climates in coastal regions (Chaparral and South Coast regions, CH and SC). The Central Valley region (CV) is largely agricultural and drains a large mountainous area in the east-central region of the state (Sierra Nevada region, SN). Urban development is concentrated in coastal areas in the central (San Francisco Bay Area, Chapparal region) and southern (Los Angeles, San Diego metropolitan area, South Coast) regions of the state. Landscape alteration has been relatively recent, with one estimate that developed lands increased in California by 38% from 1973 to 2000 (Sleeter et al. 2011). Silviculture and logging activities have also occurred in forested regions (SN, NC). For analysis, the state was evaluated as a whole and by the major regions described above (Ode et al. 2011).

The landscape model was developed using land use data, national stream hydrography layers, and biological assessments. A general assumption was that water quality issues could be conceptually linked to societal and economic drivers, reflected through the link between land use and stream biotic integrity (e.g., under the DPSIR framework, Figure 2, Smeets and Weterings 1999). Stream data from the National Hydrography Dataset Plus (NHD-plus, McKay et al. 2012) were used to identify stream segments in California for modelling biological integrity. The NHD-plus is a surface water framework that maps drainage networks and associated features (e.g., streams, lakes, canals, etc.) in the United States. Stream segments designated in the NHD-plus were used as the discrete spatial unit for modelling biological integrity. Here and throughout, “segment” is defined based on NHD-Plus flowlines. Hydrography data were combined with landscape metrics available from the StreamCat Dataset (Hill et al. 2016) that provided estimates of land use at the riparian zone (i.e., a 100-m buffer on each side of the stream segment), the catchment (i.e., nearby landscape flowing directly into the immediate stream segment, excluding upstream segments), and the entire upstream watershed for each NHD-Plus segment. Many of the metrics in StreamCat were derived from the 2006 National Land Cover Database (Fry et al. 2011).

The California Stream Condition Index (CSCI, Mazor et al. 2016) was used as a measure of biological condition in California streams. The CSCI is a predictive index that compares the observed taxa and metrics at a site to those expected under least disturbed reference conditions (Stoddard et al. 2006). Expected values at a site are based on models that estimate the likely macroinvertebrate community in relation to factors that naturally influence biology, e.g., watershed size, elevation, climate, etc. (Moss et al. 1987; Cao et al. 2007). The index score at a site can vary from 0 to ~ 1.4, with values near 1 indicating less deviation from reference state. Because the index was developed to minimize the influence of natural gradients, the index scores have consistent meaning across the state (Mazor et al. 2016). 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.

Benthic macroinvertebrate data were used to calculate 6270 individual CSCI scores at nearly 3400 unique sites between 2000 and 2016 (Figure 1b). We aggregated data collected under more than 20 federal, state, and regional bioassessment programs. Some of these programs employed a spatially balanced probabilistic design (e.g., the statewide Perennial Stream Assessment, Rehn 2015; the Stormwater Monitoring Coalition's survey of southern California streams, Mazor 2015), while other programs used different designs for project-specific purposes (such as the statewide Reference Condition Monitoring Program, Ode et al. 2016). Most of these programs targeted perennial streams, although an unknown number of intermittent streams with flows lasting into the normal sampling period were included (Mazor et al. 2014), particularly in more arid southern California. Because these programs are extensive, most regions and stream-types where perennial wadeable streams are located were represented in the calibration data set.

Field samples were collected during base flow conditions typically between May and July following methods in Ode et al. (2016). Bioassessment sites were snapped to the closest NHD-plus stream segment in ArcGIS (ESRI 2016). In cases where multiple sites were located on the same segment, the most downstream site was selected for model calibration under the assumption that the landscape data in StreamCat were most relevant to this site. One sample date was chosen randomly for sites with multiple dates so that one CSCI score was matched to a site. This created a final dataset of 2620 unique field observations used to calibrate and validate the landscape model.

Building and validating the landscape model

Expected CSCI scores were modelled using estimates of canal/ditch density, imperviousness, road density/crossings, and urban and agricultural land use for each stream segment (Table 1, Figure S1). StreamCat was used as the only source for predictor variables because of consistent methods and linkage to NHD-Plus flowlines (Hill et al. 2016). Preliminary analyses indicated that these variables produced a predictive model with comparable performance relative to a larger model with additional predictors. These variables were chosen specifically as indicators of land-management activities that were most likely to limit the attainability of biological integrity. 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. 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). Overall, the model was associative by design and was intended as a predictive tool that does not describe specific mechanisms of biological alteration. We assumed that deviation of observed scores from the model predictions (i.e., residuals) could be used to describe in-stream factors associated with condition for follow-up analysis.

The model was developed using quantile regression forests to estimate ranges of likely CSCI scores in different landscapes (Meinshausen 2006, 2017). Random forests are an ensemble learning approach to predictive modelling that aggregates information from a large number of regression trees and have been used extensively in bioassessment applications (Carlisle et al. 2009; Chen et al. 2014; Mazor et al. 2016; Fox et al. 2017). Random forest models can quantify complex, non-linear relationships and interactions between variables and can be more effective with large datasets relative to more commonly-used approaches, such as multiple regression (Breiman 2001; Hastie et al. 2009). Quantile models, such as quantile regression forests, evaluate the conditional response across the range of values that are expected, in contrast to conventional models that provide only an estimate of the mean response (Cade and Noon 2003). 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, which can be used to identify sites where that range includes management targets. Quantile regression forests were used to predict CSCI scores in each stream segment at five percent increments (i.e., 5th, 10th, etc.) from the 5th to 95th percentile of expectations. The statewide validated model (described below) was used to predict percentile expectations of CSCI scores at all stream segments where predictors were available. For example, the 50th percentile prediction was the most likely score for a stream segment given observed values for landscape variables, whereas a lower (e.g., 5th percentile) and upper (95th percentile) conditional quantile (and points in between) were also predicted. The quantregForest package (Meinshausen 2017) for the R Statistical Programming Language (RDCT 2018) was used to develop the landscape model using the default settings.

We stratified sample data to ensure sufficient representation of landscape gradients across major regions in the state (Figure 1). Calibration data for the landscape model were obtained from a random selection of 75% of segments with observed CSCI scores, where the selection was based on a random draw from strata of quartiles defined by increasing watershed imperviousness relative to each region (n = 1965 segments). This ensured that the model was calibrated with data that covered the variation of landscape development among regions (i.e., regions with low development were not under-represented and those with high development were not over-represented). The remaining 25% of sites were used for model validation (n = 655). Where multiple samples were available at a single site, one sample was selected at random for both calibration and validation purposes. 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).

Model performance metrics were chosen to evaluate both predictive ability of the landscape model and potential for bias which may vary depending on different land use gradients across the state. Performance was assessed for the statewide dataset and within each major region by comparing differences between observed CSCI scores and median predictions at the same locations. Differences were evaluated using Pearson correlations and root mean squared errors (RMSE); high correlation coefficients and low RMSE values indicated good predictive ability. Regression analysis between observed and predicted scores was used to assess potential bias based on intercept and slope values differing from 0 and 1, respectively.

Statewide application of the landscape model

We applied the landscape model to 138716 stream segments statewide to estimate the extent of streams in one of four different constraint classes: likely unconstrained, possibly unconstrained, possibly constrained, and likely constrained (Table 2). 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. The classification process is described in Figure 3a through c. Classifications were based on the comparison of a CSCI threshold representing a management goal and the predicted range or predicted median score at a segment. These two decision points (i.e., the threshold and the size of the predicted range) were critical in defining segment classifications. We used a CSCI threshold of 0.79 following previous examples (Mazor et al. 2016; SDRWQB 2016) and a lower and upper bound from the 10th to the 90th percentiles of expected CSCI scores. 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 (Figure 3c). The remaining sites were classified as possibly unconstrained or possibly constrained, based on whether the median expectation was above or below the threshold respectively (Table 2).

The influence of the key decision points on the extent of segment classifications created by the landscape model was evaluated. Stream segment classifications depend on the percentile range of score expectations (or certainty) from the landscape model (Figure 3b) and the CSCI threshold for evaluating the overlap extent (Figure 3c). For the certainty range, these bounds do not describe statistical certainty in the traditional sense (e.g., confidence interval), but rather a desired range that is defined as a potentially acceptable lower and upper limit around the median prediction for a CSCI score given landscape development. Eight different ranges of values for the score expectations from wide to narrow were evaluated at five percent intervals, i.e., 5th-95th, 10th-90th, …, 45th-55th. Different CSCI thresholds were also evaluated using values of 0.63, 0.79, and 0.92, corresponding to the 1st, 10th, and 30th percentile of scores at reference calibration sites used to develop the CSCI (Figure 1b, Mazor et al. 2016). The percentage of stream segments in each class statewide and by major regions were estimated for each of the twenty-four scenarios (width by threshold combinations). 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.

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, which can 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 above the upper limit of the segment expectation (e.g., above the 90th percentile of expected scores) were considered "over-scoring" and sites below the lower limit (e.g., 10th percentile) were considered "under-scoring". If neither "over-scoring" nor "under-scoring", the site was considered as "expected" within the context of the landscape model.

Results

Model performance

There was generally good agreement between observed and predicted CSCI scores statewide (Table 3, Figure S2). 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 median prediction bias. Performance was similar with the validation dataset (r = 0.72, RMSE = 0.18, intercept = 0.07, slope = 0.90).

Overall, the model performed well in regions with a mix of urban, agricultural, and open land (e.g., South Coast and Chaparral regions), whereas performance was weakest in regions without strong development gradients (e.g., Sierra Nevada region) (Table 3, Figure S2, S3). Performance for the Chaparral and South Coast regions were comparable or slightly improved compared to the statewide dataset for both the calibration (r = 0.71, 0.75, respectively) and validation (r = 0.74, 0.72) datasets. Model predictions for the Central Valley, Desert/Modoc, and North Coast regions had worse performance compared to the statewide results, with correlations of approximately 0.66, 0.50, and 0.55 with observed values in the calibration dataset and 0.49, 0.55, and 0.55 in the validation dataset. Model performance was weakest for the Sierra Nevada region (calibration r = 0.45, validation r = 0.21), where timber harvesting, rather than urban or agricultural development, is the most widespread stressor. A slight bias in model predictions was observed for the Central Valley and North Coast, where the former was over-predicted and the latter was under-predicted (Figure S2).

Statewide patterns in stream constraints

Statewide, spatial patterns in the predicted limits of biological integrity were similar to patterns in land use (Figure 4). A majority of stream segments statewide were classified as possibly constrained (11% of all stream length) or possibly unconstrained (46%), whereas a minority were likely constrained (4%) or likely unconstrained (39%) (Table 4). Likely unconstrained streams were common in the Sierra Nevada (50%), North Coast (46%), and Desert/Modoc (46%) regions, whereas likely constrained were relatively abundant in the Central Valley (22%) and South Coast (15%) regions. However, constrained and unconstrained streams were both found in every region (Figure 4)

Observed CSCI scores were within the predicted decile range as often as expected (i.e., 80% statewide, based on the 10th and 90th conditional quantiles), and over-scoring sites were roughly as common (9%) as under-scoring sites (10%) (Table 5). Similar patterns were observed within regions, although a slightly larger percentage of sites in the Central Valley were under-scoring compared to the other regions, which may have been evidence of a slight bias of over-predicting in this region. 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).

Changing key decision points of the landscape model affected the estimates of the extent of streams in each class (Figure 5). Unsurprisingly, decreasing the certainty of predictions from the landscape model by narrowing the quantile range (5th-95th to 45th-55th) shifted a number of streams from the possible to likely category in both constrained and unconstrained segments. Similarly, changing the CSCI threshold from relaxed to more conservative (0.63 to 0.92) increased the number of streams classified as possibly or likely constrained and decreased the number of streams as possibly or likely unconstrained. However, the effects of these decision points varied greatly by region. For example, over 80% of segments in the Central Valley were classified as likely constrained using a high CSCI threshold with the narrowest range of predictions, whereas less than 1% of segments were in this category using a low CSCI threshold with the widest range of predictions. Opposite trends were observed in regions with reduced land use pressures. For example, almost all stream segments in the North Coast and Sierra Nevada regions were classified as likely unconstrained using a low CSCI threshold and narrow range of predictions.

Discussion

Managing for biological integrity requires the use of 1) assessment tools that can accurately evaluate condition, and 2) tools that provide an estimate of the range of attainable conditions relative to the landscape. The landscape model was developed with these needs in mind to better inform application of the CSCI for decision-making relative to landscape constraints on biological condition. Statewide application of the model demonstrated where streams are likely constrained on a regional basis, whereas application in a case study (described below) demonstrated how the model can be used by local stakeholders to prioritize management actions that are informed by landscape context. The landscape model can inform the interpretation of biotic condition and is a decision-making tool that can help identify where management goals could be focused.

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 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 et al. (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.

Case study: Application of the landscape model to the San Gabriel River watershed

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 a higher priority for managers than a site scoring as expected in a constrained segment; the latter may require more resources for comparable changes in biotic condition. As such, the lower San Gabriel watershed is heavily urbanized with many modified channels (Figure S4b) and managers require prioritization tools to identify where efforts should be focused among many sites 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).

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 first identified independent from the landscape model and then 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. In general, stakeholders assigned higher priority for all three actions to sites in likely unconstrained segments where CSCI scores were over- or under-scoring or at sites that were possibly unconstrained but the observed CSCI scores were below the biological threshold (dotted line in Figure S6, Table S1). Constrained sites were given lower priority overall or restoration actions were recommended as a lower priority despite low CSCI scores. Continuing current practices (e.g., routine monitoring, neither of the above actions) was also identified by the stakeholders as necessary for these low priority sites. Recommended actions to investigate were applied to both over-scoring and under-scoring sites, protect was given a high priority exclusively at over-scoring sites, and restore was more common at under-scoring sites.

The landscape model is primarily an exploratory tool to help identify patterns among monitoring sites where more intensive analyses may be appropriate. This application was tested through engagement of our local stakeholder group. Rather than identifying individual sites in need of specific management actions, the group used the landscape model to characterize patterns on the landscape that were consistent with the recommended management priorities. In doing so, the group explored and discussed potential management actions relative to the landscape characteristics of the watershed. 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 how outcomes from the model could be applied. 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). However, the recommended actions have relevance only in the interests of the San Gabriel Regional Monitoring Program. Localized applications of the statewide model must engage stakeholders in a similar process to develop recommendations that are specific to regional needs at the watershed scale (Brody 2003; Reed 2008).

Engagement with the stakeholder group was facilitated through creation of an interactive and online application, the Stream Classification and Priority Explorer (SCAPE, Figure S7, http://shiny.sccwrp.org/scape/, Beck 2018b). The SCAPE application can be used to select and visualize management priorities for all monitoring sites in the San Gabriel watershed (Figure S8) and was also critical for demonstrating how results from the statewide model could be used at a regional scale. Because of this application, the stakeholder group was able to explore the potential impacts of biointegrity policies 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. Additionally, the SCAPE application correctly identified sites where discrepancies between CSCI scores and other measures of stream condition had been previously observed. Without the landscape context provided by the model (i.e., Figure S5, right side), stakeholders had limited information to prioritize among sites (i.e., no context for scores, Figure S5, left side).

The SCAPE application also demonstrated core concepts of the model and allowed stakeholders to explore the key decision points that affected the model output. Specifically, drop-down menus and sliders allowed users to change certainties in the CSCI score predictions (e.g., 10th and 90th percentile predictions) and explore alternative thresholds for biological objectives (e.g., 10th percentile of reference scores that defined constraint classes). This functionality allowed the stakeholders to develop recommendations that were completely independent of the model, i.e., decisions were not hard-wired into the model nor SCAPE. Results in Figure 5 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 captured well by the model (e.g., Sierra Nevada, North Coast). Conversely, the chosen range for the lower and upper expectation of biological integrity is a tradeoff between which constraint classes are most appropriate for a region. Wider ranges force more stream segments into the “possible” constraint classes, whereas smaller ranges 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.

Alternative applications of the landscape model

Results from our analysis could be used for managing the biological integrity of streams under state or federal water quality mandates (e.g. “biological criteria” under the Clean Water Act). 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. The selection of appropriate management actions for streams requires the consideration of the physical and chemical condition of streams concurrent with biological monitoring results. 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. For example, for sites that meet biological objectives but where the models predict some degree of constraint, regulatory actions may be associated with protecting that condition and could be implemented in the short-term to prevent degradation. Moreover, additional actions could be recommended to determine why these sites score above the constrained expectations, such as causal assessments to identify site-specific characteristics contributing to biointegrity (e.g., intact physical habitat independent of landscape development). This flexibility is not intended to exclude sites from consideration that are less likely to achieve biological objectives, but rather to facilitate the decision-making process through a more transparent application of the model in a regulatory application. 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.

Non-regulatory applications of the landscape model are also possible by identifying where additional restoration, monitoring, or protection may have the most benefit. For example, landscape models could be used to support conservation planning, particularly at the watershed scale where land use practices can be a critical factor for decision-making. Ongoing work in California has focused on setting priorities for managing biodiversity that focus on watersheds within a conservation network (Howard et al. 2018). Results from the landscape model could be used to enhance this network by providing supporting information on constraints in an assessment framework. More generally, these applications could represent a novel use of bioassessment data beyond the pass/fail paradigm in the regulatory sense, for example, as tools for land use planning (Bailey et al. 2007). In many cases, including California, bioassessment indices have been sufficiently developed to allow large-scale condition assessment across regions, yet they are rarely used as planning tools to guide decisions on where resources should be focused (Nel et al. 2009). Our landscape model makes bioassessment data in California more accessible and identifies an appropriate expectation for the information, enabling the potential for both regulatory and non-regulatory applications.

Several states have implemented alternative use designations for applying bioassessment criteria in modified channels (FDEP 2011; USEPA 2013; MBI 2016). Although our results generally support the link between impacted biology and channel modification, a regulatory framework based on direct channel modification may be insufficient because constraints are more accurately defined relative to landscape development. As defined for the model, a constrained channel may or may not be engineered (see supplement for Tecolote Creek example, Figure S9), but an engineered channel in a developed landscape will typically be constrained. Furthermore, channel modification does not always result in biological degradation, particularly if the contributing watershed is largely undeveloped. For example, Stein et al. (2013) observed reference-like bioassessment index scores in armored reaches within national forest lands in southern California. A classification framework for biological constraints using only channel modification would provide incomplete and potentially misleading information on streams. Ideally, context for evaluating biological condition from a landscape model, in conjunction with reach-specific data on channel modification, should be used.

Our approach to assessing constrained streams is readily transferable outside of California. The landscape model could be applied to other bioassessment methods, such as a multi-metric index (the most common bioassessment approach within the US; Buss et al. 2014), O/E assessments (Moss et al. 1987), biological condition gradients (Davies and Jackson 2006), or with other biological endpoints (e.g., fish or diatoms). More importantly, our use of national geospatial datasets (i.e., NHDPlus, McKay et al. 2012; StreamCat, Hill et al. 2016) means that these methods could be applied across the United States. National bioassessment indices have been developed and the landscape model could be developed as a national-scale product of constraints on biological condition to complement recent work that predicted probable biological conditions with the National Rivers and Streams Assessment (Hill et al. 2017). Global geospatial datasets of freshwater-specific environmental variables are also available and could be used to develop similar models outside of the United States (Domisch et al. 2015).

Extension of the landscape models beyond California should also consider landscape stressors that are predictive of biotic condition in other regions. For example, urban and agricultural gradients were sufficient to characterize constraints in many regions of California, whereas Hill et al. (2017) found that the volume of water stored by dams was an important predictor of biological condition in the Northern Appalachian and Northern Plains regions of the US. In their paper, Hill et al. (2017) provided an example of how predictive models could be used to identify potential sites for restoration or conservation, however, their illustration did not explicitly identify sites that were over- or under-scoring relative to a biological endpoint. 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.

Model assumptions and limitations

There are several characteristics of the landscape model that could affect its performance when applied outside of urban and agricultural settings. First, the model was developed with a focus on the needs of managers that apply bioassessment tools in developed landscapes where conditions are presumably constrained. As such, landscape variables were chosen to capture the effects of development on CSCI scores in these areas (Table 1). Application of the model in regions where different stressors have strong impacts on stream condition should consider the relevance of urban and agricultural stressors and if an alternative model that better captures other stressor gradients is needed. For example, our results suggest that streams in the North Coast and Sierra Nevada regions are largely unconstrained, but the landscape model was a poor predictor of CSCI scores in these areas. The dominant stressors likely to affect stream condition in these regions originate from sources that are less common in developed landscapes, such as silviculture and cannabis cultivation. The current landscape model does not adequately capture these impacts outside of urban and agricultural environments. Moreover, poor model predictions are compounded by low sensitivity of the CSCI to relevant stressor gradients in these regions (Mazor et al. 2016). 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.

An additional assumption is that the landscape model can adequately discriminate between intractable constraints on biology that are spatially and temporally pervasive relative to more manageable constraints. That is, we assumed that the impacts of stressors included in the model, such as urbanization, require long-term extensive mitigation planning, whereas stressors associated with deviations from model predictions can be mitigated in the short-term using focused actions. These assumptions are not unique to our model and have been used in other applications that have evaluated biological potential (Paul et al. 2008; Chessman 2014; Waite et al. 2014). However, many stressors excluded from the model can have long-lasting impacts, leading to management scenarios where long-term recovery may only be possible with sustained and costly application of resources. For example, logging activities can impact benthic macroinvertebrate communities for a decade or more after harvesting activities have stopped (Stone and Wallace 1998; Quinn and Wright-Stow 2008). In urban areas, pervasive and profound alteration to groundwater and hydrology is common and stream communities in groundwater fed systems may require substantial time and resources for restoration. The potential legacy impacts of large-scale alterations of the natural environment are not well-captured by the current model, neither from a spatial nor temporal perspective. A more refined application of the landscape model would be necessary to evaluate different scales of impact, which could include developing separate models for each region, as well as more careful selection of model inputs to capture scales of interest for potential impacts on stream condition.

The landscape model describes constraints at scales larger than instream characteristics as a necessary approach to accurately predict bioassessment scores. Additional analyses that evaluate how different predictors influence model performance at different quantiles could provide insight into how landscape factors relate to constraints (e.g., Koenker and Machado 1999). Further, a distinction between constraints on biological condition and channel modification is implicit such that indication of the former by the model does not explicitly indicate presence of the latter. As noted above, our results consistently indicated that engineered channels are biologically constrained, but the model is based on an a priori selection of land use variables to predict biotic integrity. A correspondence between habitat limitations and channel modification is likely in many cases, but data are insufficient to evaluate biological effects statewide relative to land use constraints. Moreover, bioassessment scores can be similar in modified channels compared to natural streams independent of watershed land use, i.e., concordance between degraded stream condition and channel modification may not always be observed (Stein et al. 2013). More comprehensive assessments at individual sites may be needed to diagnose the immediate causes of degraded condition.

An additional consideration in using the landscape model is the meaning of biologically constrained relative to whole stream communities. Biologically constrained sites were considered those where present landscapes were likely to limit CSCI scores that describe macroinvertebrate condition. In many cases, poor biotic condition of the macroinvertebrate community translates to poor stream condition. However, a constrained macroinvertebrate community does not always mean other biological attributes of stream condition (e.g., fish assemblages) are also constrained. Urban streams sometimes support diverse algal assemblages such that algal-based measures of biotic condition may alternatively suggest good biotic condition relative to macroinvertebrate-based indices (Brown et al. 2009; Mazor et al. 2018). Broadening the landscape model to include multiple taxonomic assemblages or endpoints would allow a more complete assessment of how condition relates to landscape alteration.

Finally, there are a few concerns applying a landscape modelling approach for bioassessment using the NHD-Plus flowlines as a base layer. We applied our model to the entire network of the NHD-Plus represented in StreamCat, which included a large number of intermittent or ephemeral streams, as well as non-wadeable rivers. Therefore, the application of model results in these stream-types is open to question, valid only to the degree that the CSCI and its response to landscape disturbance can represent more relevant measures of biological integrity. In regions where ephemeral streams are particularly common (e.g., the inland deserts or the South Coast region), estimates of the extent of constrained or unconstrained streams may be inaccurate.

Summary

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 any of the above suggestions to better understand the utility of these tools in applied contexts.

The landscape model can be used to characterize the extent of biologically constrained channels in developed landscapes and provides a tool to determine how managers can best prioritize resources for stream management by understanding landscape factors that might constrain each segment. Our application to the San Gabriel watershed demonstrated how the statewide results can be used at a spatial scale where many management decisions are implemented through close interaction with a regional stakeholder group with direct interests in the local resources. The approach leverages information from multiple sources to develop a context for biological assessment that provides an expectation of what is likely to be achieved based on current land use development. This can facilitate more targeted management actions that vary depending on the landscape context and can also inform decisions on extent and effort for future monitoring locations.

Supplement

Geospatial data of model results mapped to stream reaches in California is provided at Beck (2018a). The SCAPE model application website is available at http://shiny.sccwrp.org/scape/, full source code accessible at Beck (2018b). Additional content for the case study, figures, and tables are available in the supplement.

Author contributions

MB, RM, SJ, KW, JW, PO, RH, CL, MS, and ES performed the research and analyzed the data. MB, RM, SJ, JW, PO, RH, and CL wrote the paper. RM, SJ, KW, and PO provided data. All authors discussed the methods and results and contributed to the development of the manuscript.

Acknowledgments

The authors acknowledge support from the San Gabriel River Regional Monitoring Program and the California State Water Resources Control Board. We thank Phil Markle and Lester Yuan for reviewing an earlier draft of the manuscript. The views expressed in this article are those of the

authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency. Any mention of trade names, products, or services does not imply an endorsement by the U.S. Government or the U.S. Environmental Protection Agency. The EPA does not endorse any commercial products, services, or enterprises.

Figure captions

Figure 1 Urban and agricultural land use (a) and distribution of observed stream CSCI scores (b) in California. Cover of urban and agricultural land use in stream watersheds was used to develop a landscape model for stream segment expectations of bioassessment scores. Breakpoints for CSCI scores are the 1st, 10th, and 30th percentile of scores at least-disturbed, reference sites throughout the state. Altered and intact refers to biological condition (Mazor et al. 2016). Grey lines are major environmental regions in California defined by ecoregional and watershed boundaries, CV: Central Valley, CH: Chaparral, DM: Deserts and Modoc Plateau, NC: North Coast, SN: Sierra Nevada, SC: South Coast.

Figure 2 Conceptualized response and management pathways captured by the landscape model under the Driver-Pressure-Stress-Impact-Response (DPSIR) framework (Smeets and Weterings 1999). Landscape predictors provided in StreamCat (Hill et al. 2017) were used to describe pressures from urban and agricultural development that could impact the macroinvertebrate community in streams by altering physical and chemical habitat. Biological response was measured using the CSCI (Mazor et al. 2016) as an impact indicator and then evaluated relative to ranges of CSCI scores that were expected at each site provided by the landscape model. Observed CSCI scores and context from the landscape model provide a basis for informing management actions that could address environmental impacts at different points in the response pathway, where the management pathway could address causes at different scales and efficiencies.

Figure 3 Application of the landscape model to identify site expectations and bioassessment performance for sixteen example stream segments. A range of CSCI scores is predicted from the model (a) and the lower and upper limits of the expectations are cut to define a certainty range

for the predictions (b). Overlap of the certainty range at each segment with a chosen CSCI threshold (c) defines the stream segment classification as likely unconstrained, possibly unconstrained, possibly constrained, and likely constrained. The observed bioassessment scores are described relative to the classification as over scoring (above the certainty threshold), expected (within), and under scoring (below) for each of four stream classes (d).

Figure 4 Statewide application of the landscape model showing the stream segment classifications. Major regional boundaries are also shown (see Figure 1).

Figure 5 Changes in stream segment classes by region and statewide for different scenarios used to define biological constraints. Twenty-seven scenarios were tested that evaluated different combinations of certainty in the CSCI predictions (nine scenarios from wide to narrow prediction ranges as identified by the tail cutoff for the expected quantiles) and potential CSCI thresholds (three scenarios from low to high). The percentage of total stream length for likely unconstrained and likely constrained is shown for each scenario. Stream classifications as possibly unconstrained or possibly constrained are not shown but can be inferred form the area of white space above or below each bar. The solid black line indicates the percentage division between unconstrained and constrained classifications. CV: Central Valley, CH: Chaparral, DM: Deserts and Modoc Plateau, NC: North Coast, SN: Sierra Nevada, SC: South Coast.

Tables

Table 1 Land use variables used to develop the landscape model of stream bioassessment scores. All variables were obtained from StreamCat (Hill et al. 2016) and applied to stream segments in the National Hydrography Dataset Plus (NHD-plus, McKay et al. 2012). The measurement scales for each variable are at the riparian (100 m buffer), catchment, and/or watershed, scale relative to a stream segment. Combined scales for riparian measurements (e.g., riparian + catchment, riparian + watershed) are riparian estimates for the entire catchment or watershed area upstream, as compared to only the individual segment. Total urban and agriculture land use variables were based on sums of individual variables in StreamCat as noted in the description. Rp100: riparian, Cat: catchment, Ws: watershed

Name	Scale	Description	Unit
CanalDens	Cat, Ws	Density of NHDPlus line features classified as canal, ditch, or pipeline	km/sq km
PctImp2006	Cat, Ws, Cat + Rp100, Ws + Rp100	Mean imperviousness of anthropogenic surfaces (NLCD 2006)	%
TotUrb2011	Cat, Ws, Cat + Rp100, Ws + Rp100	Total urban land use as sum of developed open, low, medium, and high intensity (NLCD 2011)	%
TotAg2011	Cat, Ws, Cat + Rp100, Ws + Rp100	Total agricultural land use as sum of hay and crops (NLCD 2011)	%
RdDens	Cat, Ws, Cat + Rp100, Ws + Rp100	Density of roads (2010 Census Tiger Lines)	km/sq km
RdCrs	Cat, Ws	Density of roads-stream intersections (2010 Census Tiger Lines-NHD stream lines)	crossings/sq km

Table 2 Stream class definitions describing potential biological constraints. Classes are based on the overlap of the range of likely bioassessment scores with a potential threshold for a biological objective. Identifying stream classes requires selecting the cutoff range of likely scores from the landscape model and a chosen threshold for the objective.

Class	Definition	Example
Likely unconstrained	Lower bound of prediction range is above threshold	10th percentile > 0.79
Possibly unconstrained	Lower bound of prediction range is below threshold, but median prediction is above	50th percentile > 0.79
Possibly constrained	Upper bound of prediction range is above threshold, but median prediction is below	50th percentile < 0.79
Likely constrained	Upper bound of prediction range is below threshold	90th percentile < 0.79

Table 3 Performance of the landscape model by calibration (Cal) and validation (Val) datasets in predicting CSCI scores. The statewide dataset (Figure 4) and individual regions of California (Figure 1) are evaluated. Averages and standard deviations (in parentheses) for observed and predicted CSCI values of each dataset are shown. Pearson correlations (r), root mean squared errors (RMSE), intercept, and slopes are for comparisons of predicted and observed values to evaluate model performance. All correlations, intercepts, and slopes are significant at alpha = 0.05. CV: Central Valley, CH: Chaparral, DM: Deserts and Modoc Plateau, NC: North Coast, SN: Sierra Nevada, SC: South Coast.

Dataset	Location	n	Observed	Predicted	r	RMSE	Intercept	Slope
Cal	Statewide	1965	0.82 (0.26)	0.83 (0.20)	0.75	0.17	0.04	0.93
	CH	512	0.76 (0.27)	0.79 (0.21)	0.71	0.19	0.03	0.92
	CV	116	0.51 (0.18)	0.57 (0.15)	0.66	0.15	0.05	0.81
	DM	86	0.87 (0.22)	0.91 (0.14)	0.50	0.20	0.15	0.79
	NC	208	0.92 (0.20)	0.94 (0.13)	0.55	0.17	0.12	0.86
	SC	631	0.79 (0.24)	0.78 (0.21)	0.75	0.16	0.11	0.87
	SN	412	0.98 (0.18)	0.98 (0.09)	0.45	0.16	0.12	0.88
Val	Statewide	655	0.82 (0.25)	0.84 (0.20)	0.72	0.18	0.07	0.90
	CH	172	0.76 (0.27)	0.81 (0.21)	0.74	0.19	-0.04	0.98
	CV	40	0.52 (0.19)	0.59 (0.16)	0.49	0.19	0.16	0.60
	DM	28	0.84 (0.17)	0.93 (0.11)	0.55	0.17	0.07	0.83
	NC	71	0.94 (0.19)	0.96 (0.11)	0.55	0.16	0.00	0.98
	SC	208	0.80 (0.24)	0.78 (0.21)	0.72	0.17	0.17	0.81
	SN	136	0.97 (0.17)	0.98 (0.09)	0.21	0.17	0.57	0.41

Table 4: Summary of stream length for each stream class statewide and major regions of California (Figures 1, 4). Lengths are in kilometers with the percentage of the total length in a region in parentheses. All lengths are based on a CSCI threshold of 0.79 and the 10th to 90th percentile of expected scores from the landscape model. CV: Central Valley, CH: Chaparral, DM: Deserts and Modoc Plateau, NC: North Coast, SN: Sierra Nevada, SC: South Coast.

	constrained		unconstrained	
Region	likely	possibly	possibly	likely
Statewide	8150 (4)	24735 (11)	101591 (46)	85317 (39)
CV	3356 (22)	8010 (52)	3202 (21)	951 (6)
CH	1642 (3)	7840 (13)	30693 (50)	21206 (35)
DM	255 (0)	3395 (6)	27194 (47)	26479 (46)
NC	108 (0)	1442 (5)	14152 (49)	13286 (46)
SN	20 (0)	1067 (3)	18228 (48)	19032 (50)
SC	2770 (15)	2981 (16)	8122 (45)	4363 (24)

Table 5: Summary of CSCI scores by relative expectations for each stream class statewide and in each major region of California (Figures 1, 4). Average CSCI scores (standard deviation) and counts (percent) of the number of monitoring stations in each relative score category and region are shown. Sites are over-scoring if the observed scores are above the range of expectations at a segment, expected if within the range, or under-scoring if below the range. CV: Central Valley, CH: Chaparral, DM: Deserts and Modoc Plateau, NC: North Coast, SN: Sierra Nevada, SC: South Coast.

	under-scoring		expected		over-scoring	
Region	CSCI	n (%)	CSCI	n (%)	CSCI	n (%)
Statewide	0.54 (0.21)	267 (10)	0.83 (0.23)	2041 (80)	1.08 (0.17)	242 (9)
CH	0.47 (0.18)	89 (13)	0.79 (0.24)	535 (80)	1.08 (0.17)	45 (7)
CV	0.34 (0.12)	25 (17)	0.54 (0.17)	118 (81)	0.63 (0.25)	2 (1)
DM	0.6 (0.17)	15 (14)	0.9 (0.17)	89 (80)	1.15 (0.08)	7 (6)
NC	0.66 (0.17)	28 (10)	0.93 (0.16)	228 (82)	1.15 (0.08)	22 (8)
SC	0.54 (0.22)	56 (7)	0.78 (0.22)	656 (81)	1.02 (0.2)	97 (12)
SN	0.67 (0.16)	54 (10)	0.99 (0.11)	415 (77)	1.16 (0.06)	69 (13)

References

Allan, D., D. Erickson, and J. Fay. 1997. The Influence of Catchment Land Use on Stream Integrity Across Multiple Spatial Scales. *Freshwater Biology* 37 (1): 149–61. https://doi.org/10.1046/j.1365-2427.1997.d01-546.x.

Bailey, R. C., T. B. Reynoldson, A. G. Yates, J. Bailey, and S. Linke. 2007. Integrating Stream Bioassessment and Landscape Ecology as a Tool for Land Use Planning. *Freshwater Biology* 52 (5): 908–17. https://doi.org/10.1111/j.1365-2427.2006.01685.x.

Beck, M. W. 2018a. Constrained streams for biological integrity in California. Knowledge Network for Biocomplexity. urn:uuid:75411f50-32ed-42a5-bbfd-26833c7a441f.

———. 2018b. SCCWRP/SCAPE: v1.0 (Version 1.0). Zenodo, http://doi.org/10.5281/zenodo.1218121.

Beechie, T., G. Pess, P. Roni, and G. Giannico. 2007. Setting River Restoration Priorities: A Review of Approaches and General Protocol for Identifying and Prioritizing Actions. *North American Journal of Fisheries Management* 28 (3): 891–905. https://doi.org/10.1577/M06-174.1.

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.

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

Brody, S. D. 2003. Measuring the Effects of Stakeholder Participation on the Quality of Local Plans Based on the Principles of Collaborative Ecosystem Management. *Journal of Planning Education and Research* 22 (4): 407–19. https://doi.org/10.1177/0739456X03022004007.

Brown, L. R., T. F. Cuffney, J. F. Coles, F. Fitzpatrick, G. McMahon, J. Steuer, A. H. Bell, and J. T. May. 2009. Urban Streams Across the USA: Lessons Learned from Studies in Nine Metropolitan Areas. *Journal of the North American Benthological Society* 28 (4): 1051–69. https://doi.org/10.1899/08-153.1.

Brown, L. R., J. T. May, A. C. Rehn, P. R. Ode, I. R. Waite, and J. G. Kennen. 2012. Predicting Biological Condition in Southern California Streams. *Landscape and Urban Planning* 108 (1): 17–27. https://doi.org/10.1016/j.landurbplan.2012.07.009.

Buss, D. F., D. M. Carlisle, T. -S. Chon, J. Culp, J. s. Harding, H. E. Keizer-Vlek, W. A. Robinson, S. Strachan, C. Thirion, and R. M. Hughes. 2014. Stream Biomonitoring Using Macroinvertebrates Around the Globe: A Comparison of Large-Scale Programs. *Environmental Monitoring and Assessment* 187: 4132. https://doi.org/10.1007/s10661-014-4132-8.

Cade, B. S., and B. R. Noon. 2003. A Gentle Introduction to Quantile Regression for Ecologists. *Frontiers in Ecology and the Environment* 1 (8): 412–20.

Cao, Y., C. P. Hawkins, J. Olson, and M. A. Kosterman. 2007. Modeling Natural Environmental Gradients Improves the Accuracy and Precision of Diatom-Based Indicators. *Journal of the North American Benthological Society* 26 (3): 566–85. https://doi.org/10.1899/06-078.1.

Carline, R. F., and M. C. Walsh. 2007. Responses to Riparian Restoration in the Spring Creek Watershed, Central Pennsylvania. *Restoration Ecology* 15 (4): 731–42. https://doi.org/10.1111/j.1526-100X.2007.00285.x.

Carlisle, D. M., J. Falcone, and M. R. Meador. 2009. Predicting the Biological Condition of Streams: Use of Geospatial Indicators of Natural and Anthropogenic Characteristics of Watersheds. *Environmental Monitoring and Assessment* 151 (1-4): 143–60. https://doi.org/10.1007/s10661-008-0256-z.

Chen, K., R. M. Hughes, S. Xu, J. Zhang, D. Cai, and B. Wang. 2014. Evaluating Performance of Macroinvertebrate-Based Adjusted and Unadjusted Multi-Metric Indices (MMI) Using Multi-Season and Multi-Year Samples. *Ecological Indicators* 36: 142–51. https://doi.org/10.1016/j.ecolind.2013.07.006.

Chessman, B. C. 2014. Predicting Reference Assemblages for Freshwater Bioassessment with Limiting Environmental Difference Analysis. *Freshwater Science* 33 (4): 1261–71. https://doi.org/10.1086/678701.

Chessman, B. C., M. Muschal, and M. J. Royal. 2008. Comparing Apples with Apples: Use of Limiting Environmental Differences to Match Reference and Stressor-Exposure Sites for Bioassessment of Streams. *River Research and Applications* 24 (1): 103–17. https://doi.org/10.1002/rra.1053.

Chessman, B. C., and M. J. Royal. 2004. Bioassessment Without Reference Sites: Use of Environmental Filters to Predict Natural Assemblages of River Macroinvertebrates. *Journal of the North American Benthological Society* 23 (3): 599–615. https://doi.org/10.1899/0887-3593(2004)023%3C0599:BWRSUO%3E2.0.CO;2.

Clapcott, J. E., K. J. Collier, R. G. Death, E. O. Goodwin, J. S. Harding, D. Kelly, J. R. Leathwick, and R. G. Young. 2011. Quantifying Relationships Between Land-Use Gradients and Structural and Functional Indicators of Stream Ecological Integrity. *Freshwater Biology* 57 (1): 74–90. https://doi.org/10.1111/j.1365-2427.2011.02696.x.

Cuffney, T. F., S. S. Qian, R. A. Brightbill, J. T. May, and I. R. Waite. 2011. Response to King and Baker: Limitations on Threshold Detection and Characterization of Community Thresholds. *Ecological Applications* 21 (7): 2840–5. https://doi.org/10.2307/41416699.

Davies, S. P., and S. K. Jackson. 2006. The Biological Condition Gradient: A Descriptive Model for Interpreting Change in Aquatic Ecosystems. *Ecological Applications* 16 (4): 1251–66.

Domisch, S., G. Amatulli, and W. Jetz. 2015. Near-Global Freshwater-Specific Environmental Variables for Biodiversity Analyses in 1 Km Resolution. *Scientific Data* 2: 150073. https://doi.org/10.1038/sdata.2015.73.

ESRI (Environmental Systems Research Institute). 2016. ArcGIS v10.5. Redlands, California.

Finkenbine, J. K., J. W. Atwater, and D. S. Mavinic. 2000. Stream Health After Urbanization. *Journal of the American Water Resources Association* 36 (5): 1149–60. https://doi.org/10.1111/j.1752-1688.2000.tb05717.x.

FDEP (Florida Department of Environmental Protection). 2011. Development of aquatic life use support attainment thresholds for Florida’s Stream Condition Index and Lake Vegetation Index. DEP-SAS-003/11. Tallahassee, Florida: FDEP Standards; Assessment Section, Bureau of Assessment; Restoration Support.

Fox, E. W., R. A. Hill, S. G. Leibowitz, A. R. Olsen, D. J. Thornbrugh, and M. H. Weber. 2017. Assessing the Accuracy and Stability of Variable Selection Methods for Random Forest Modeling in Ecology. *Environmental Monitoring and Assessment* 189: 316. https://doi.org/10.1007/s10661-017-6025-0.

Fry, J., G. Xian, S. Jin, J. Dewitz, C. Homer, L. Yang, C. Barnes, N. Herold, and J. Wickham. 2011. Completion of the 2006 National Land Cover Database for the Conterminous United States. *Photogrammetric Engineering and Remote Sensing* 77 (9): 858–64.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Hill, R. A., E. W. Fox, S. G. Leibowitz, A. R. Olsen, D. J. Thornbrugh, and M. H. Weber. 2017. Predictive Mapping of the Biotic Condition of Conterminous U.S. Rivers and Streams. *Ecological Applications* 27 (8): 2397–2415. https://doi.org/10.1002/eap.1617.

Hill, R. A., M. H. Weber, S. G. Leibowitz, A. R. Olsen, and D. J. Thornbrugh. 2016. The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. *Journal of the American Water Resources Association* 52: 120–28. https://doi.org/10.1111/1752-1688.12372.

Howard, J. K., K. A. Fesenmyer, T. E Grantham, J. H. Viers, P. R. Ode, P. B. Moyle, S. J. Kupferburg, et al. 2018. A Freshwater Conservation Blueprint for California: Prioritizing Watersheds for Freshwater Biodiversity. *Freshwater Science* 37 (2): 417–31. https://doi.org/10.1086/697996.

Hynes, H. B. N. 1975. The Stream and Its Valley. *SIL Proceedings, 1922-2010* 19 (1): 1–15. https://doi.org/10.1080/03680770.1974.11896033.

Johnson, L., C. Richards, G. Host, and J. Arthur. 1997. Landscape Influences on Water Chemistry in Midwestern Stream Ecosystems. *Freshwater Biology* 37 (1): 193–208. https://doi.org/10.1046/j.1365-2427.1997.d01-539.x.

Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant, and I. J. Schlosser. 1986. Assessing Biological Integrity in Running Waters: A Method and Its Rationale. Special Publication 5. Champaign, Illinois: Illinois Natural History Survey.

Kenney, M. A., P. R. Wilcock, B. F. Hobbs, N. E. Flores, and D. C. Martínez. 2012. Is Urban Stream Restoration Worth It? *Journal of the American Water Resources Association* 48 (3): 603–15. https://doi.org/10.1111/j.1752-1688.2011.00635.x.

Koenker, R., and J. A. F. Machado. 1999. Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association* 94 (448): 1296–1310.

Lester, R. E., and A. J. Boulton. 2008. Rehabilitating Agricultural Streams in Australia with Wood: A Review. *Environmental Management* 42 (2): 310–26. https://doi.org/10.1007%2Fs00267-008-9151-1.

Loflen, C., H. Hettesheimer, L. B. Busse, K. Watanabe, R. M. Gersberg, and V. Lüderitz. 2016. Inadequate Monitoring and Inappropriate Project Goals: A Case Study on the Determination of Success for the Forester Creek Improvement Project. *Ecological Restoration* 34 (2): 124–34. https://doi.org/10.3368/er.34.2.124.

Mazor, R. D. 2015. Bioassessment of Perennial Streams in Southern California: A Report on the First Five Years of the Stormwater Monitoring Coalition’s Regional Stream Survey. 844. Costa Mesa, California: Southern California Coastal Water Research Project.

Mazor, R. D., M. W. Beck, and J. Brown. 2018. 2017 Report on the SMC Regional Stream Survey. 1029. Costa Mesa, California: Southern California Coastal Water Research Project.

Mazor, R. D., A. C. Rehn, P. R. Ode, M. Engeln, K. C. Schiff, E. D. Stein, D. J. Gillett, D. B. Herbst, and C. P. Hawkins. 2016. Bioassessment in Complex Environments: Designing an Index for Consistent Meaning in Different Settings. *Freshwater Science* 35 (1): 249–71.

Mazor, R. D., E. D. Stein, P. R. Ode, and K. Schiff. 2014. Integrating Intermittent Streams into Watershed Assessments: Applicability of an Index of Biotic Integrity. *Freshwater Science* 35 (2): 459–74. https://doi.org/10.1086/675683.

MBI (Midwest Biodiversity Institute). 2016. Identification of predictive habitat attributes for Minnesota streams to support tiered aquatic life uses. MBI Technical Report MBI/OHPAN1518840. Columbus, Ohio: Midwest Biodiversity Institute, prepared on behalf of the Minnesota Pollution Control Agency.

McKay, L., T. Bondelid, T. Dewald, J. Johnston, R. Moore, and A. Reah. 2012. NHDPlus Version 2: User Guide.

Meinshausen, N. 2006. Quantile Regression Forests. *Journal of Machine Learning Research* 7: 983–99.

Meinshausen, N. 2017. *QuantregForest: Quantile Regression Forests*. https://CRAN.R-project.org/package=quantregForest.

Morgan, R. P., and S. E. Cushman. 2005. Urbanization Effects on Stream Fish Assemblages in Maryland, USA. *Journal of the North American Benthological Society* 24 (3): 643–55.

Moss, D., M. T. Furse, J. F. Wright, and P. D. Armitage. 1987. The Prediction of the Macro-Invertebrate Fauna of Unpolluted Running-Water Sites in Great Britain Using Environmental Data. *Freshwater Biology* 17 (1): 41–52. https://doi.org/10.1111/j.1365-2427.1987.tb01027.x.

Nel, J. L., D. J. Roux, R. Abell, P. J. Ashton, R. M. Cowling, J. V. Higgins, M. Thieme, and J. H. Viers. 2009. Progress and Challenges in Freshwater Conversation Planning. *Aquatic Conservation: Marine and Freshwater Ecosystems* 19 (4): 474–85. https://doi.org/10.1002/aqc.1010.

Ode, P. R., A. E. Fetscher, and L. B. Busse. 2016. Standard Operating Procedures (SOP) for the Collection of Field Data for Bioassessments of California Wadeable Streams: Benthic Macroinvertebrates, Algae, and Physical Habitat. SWAMP-SOP-SB-2016-0001. Sacramento, California: California State Water Resources Control Board Surface Water Ambient Monitoring Program. https://www.waterboards.ca.gov/water_issues/programs/swamp/bioassessment/docs/combined_sop_2016.pdf.

Ode, P. R., T. M. Kincaid, T. Fleming, and A. C. Rehn. 2011. Ecological Condition Assessments of California's Perennial Wadeable Streams: Highlights from the Surface Water Ambient Monitoring Program's Perennial Streams Assessment (PSA). Sacramento, California: A collaboration between the State Water Resources Control Board's Non-Point Source Pollution Control Program (NPS Program), Surface Water Ambient Monitoring Program (SWAMP), California Department of Fish; Game Aquatic Bioassessment Laboratory; the U.S.

Environmental Protection Agency. https://www.waterboards.ca.gov/water_issues/programs/swamp/docs/reports/psa_smmry_rpt.pdf.

Ode, P. R., A. C. Rehn, R. D. Mazor, K. C. Schiff, E. D. Stein, J. T. May, L. R. Brown, et al. 2016. Evaluating the Adequacy of a Reference-Site Pool for Ecological Assessments in Environmentally Complex Regions. *Freshwater Science* 35 (1): 237–48.

Paul, M. J., D. W. Bressler, A. H. Purcell, M. T. Barbour, E. T. Rankin, and V. H. Resh. 2008. Assessment Tools for Urban Catchments: Defining Observable Biological Potential. *Journal of the American Water Resources Association* 45 (2): 320–30. https://doi.org/10.1111/j.1752-1688.2008.00280.x.

Quinn, J. M., and A. E. Wright-Stow. 2008. Stream Size Influences Stream Temperature Impacts and Recovery Rates After Clearfell Logging. *Forest Ecology and Management* 256 (12): 2101–9. https://doi.org/10.1016/j.foreco.2008.07.041.

RDCT (R Development Core Team). 2018. R: A language and environment for statistical computing, v3.5.1. R Foundation for Statistical Computing, Vienna, Austria.

Reed, M. S. 2008. Stakeholder Participation for Environmental Management: A Literature Review. *Biological Conservation* 141 (10): 2417–31. https://doi.org/10.1016/j.biocon.2008.07.014.

Rehn, A. C. 2015. The Perennial Streams Assessment (PSA): An Assessment of Biological Condition Using the New California Stream Condition Index (CSCI). SWAMP Management Memorandum, SWAMP-MM-2015-0001. Sacramento, California: California State Water

Resources Control Board Surface Water Ambient Monitoring Program. https://www.waterboards.ca.gov/water_issues/programs/swamp/bioassessment/docs/psa_memo_121015.pdf.

Reynoldson, T. B., R. H. Norris, V. H. Resh, K. E. Day, and D. M. Rosenberg. 1997. The Reference Condition: A Comparison of Multimetric and Multivariate Approaches to Assess Water-Quality Impairment Using Benthic Macroinvertebrates. *Journal of the North American Benthological Society* 16 (4): 833–52. https://doi.org/10.2307/1468175.

Richards, C., R. Haro, L. Johnson, and G. Host. 1997. Catchment and Reach-Scale Properties as Indicators of Macroinvertebrate Species Traits. *Freshwater Biology* 37 (1): 219–30. https://doi.org/10.1046/j.1365-2427.1997.d01-540.x.

Roni, P., and T. Beechi. 2012. *Stream and Watershed Restoration Guide: A Guide to Restoring Riverine Processes and Habitats*. First. Hoboken, New Jersey: John Wiley & Sons.

SDRWQB (San Diego Regional Water Quality Control Board). 2016. Clean Water Act Sections 305(B) and 303(D) Integrated Report for the San Diego Region. Sacramento, California: California Environmental Protection Agency. https://www.waterboards.ca.gov/sandiego/water_issues/programs/303d_list/docs/Staff_Report_101216.pdf.

Shoredits, A. S., and J. A. Clayton. 2013. Assessing the Practice and Challenges of Stream Restoration in Urbanized Environments of the USA. *Geography Compass* 7 (5): 358–72. https://doi.org/10.1111/gec3.12039.

Sleeter, B. M., T. S. Wilson, C. E. Soulard, and J. Liu. 2011. Estimation of the Late Twentieth Century Land-Cover Change in California. *Environmental Monitoring and Assessment* 173 (1-4): 251–66. https://doi.org/10.1007/s10661-010-1385-8.

Smeets, E., and R. Weterings. 1999. Environmental Indicators: Typology and Overview. No. 25. Copenhagen, Denmark: European Environmental Agency.

Stein, E. D., M. R. Cover, A. E. Fetscher, C. O'Reilly, R. Guardado, and C. W. Solek. 2013. Reach-Scale Geomorphic and Biological Effects of Localized Streambank Armoring. *Journal of the American Water Resources Association* 49 (4): 780–92. https://doi.org/10.1111/jawr.12035.

Stein, E. D., A. Sengupta, R. D. Mazor, K. McCune, B. P. Bledsoe, and S. Adams. 2017. Application of Regional Flow-Ecology Relationships to Inform Watershed Management Decisions: Applications of the ELOHA Framework in the San Diego River Watershed, California, USA. *Ecohydrology* 10 (7): e1869. https://doi.org/10.1002/eco.1869.

Stoddard, J. L., D. P. Larsen, C. P. Hawkins, R. K. Johnson, and R. H. Norris. 2006. Setting Expectations for the Ecological Condition of Streams: The Concept of Reference Condition. *Ecological Applications* 16 (4): 1267–76. https://doi.org/10.1890/1051-0761(2006)016[1267:SEFTEC]2.0.CO;2.

Stone, M. K., and J. B. Wallace. 1998. Long-Term Recovery of a Mountain Stream from Clear-Cut Logging: The Effects of Forest Succession on Benthic Invertebrate Community Structure. *Freshwater Biology* 39 (1): 151–69. https://doi.org/10.1046/j.1365-2427.1998.00272.x.

USEPA (US Environmental Protection Agency). 2013. Technical Support Document for EPA's Action on the State of Oregon's Revised Water Quality Standards for the West Division Main Canal. USEPA Region 10, Seattle, Washington.

USGS (US Geological Survey). 1999. The quality of our nation's waters: nutrients and pesticides. Reston, Virginia.

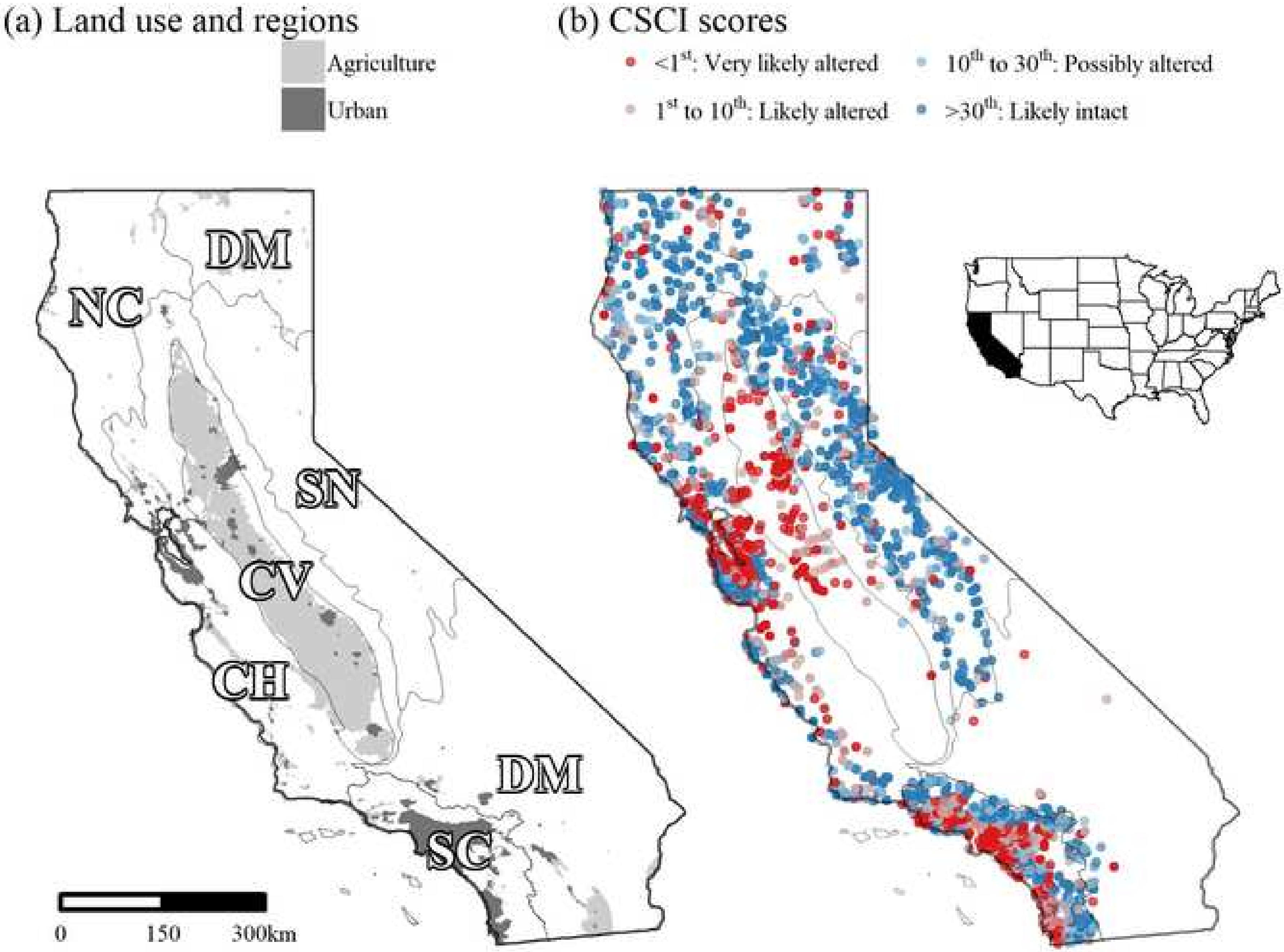
Vølstad, J. H., N. E. Roth, G. Mercurio, M. T. Southerland, and D. E. Strebel. 2004. Using Environmental Stressor Information to Predict the Ecological Status of Maryland Non-Tidal Streams as Measured by Biological Indicators. *Environmental Monitoring and Assessment* 84 (3): 219–42. https://doi.org/10.1023/A:1023374524254.

Waite, I. R. 2013. Development and Application of an Agricultural Intensity Index to Invertebrate and Algal Metrics from Streams at Two Scales. *Journal of the American Water Resources Association* 49 (2): 431–48. https://doi.org/10.1111/jawr.12032.

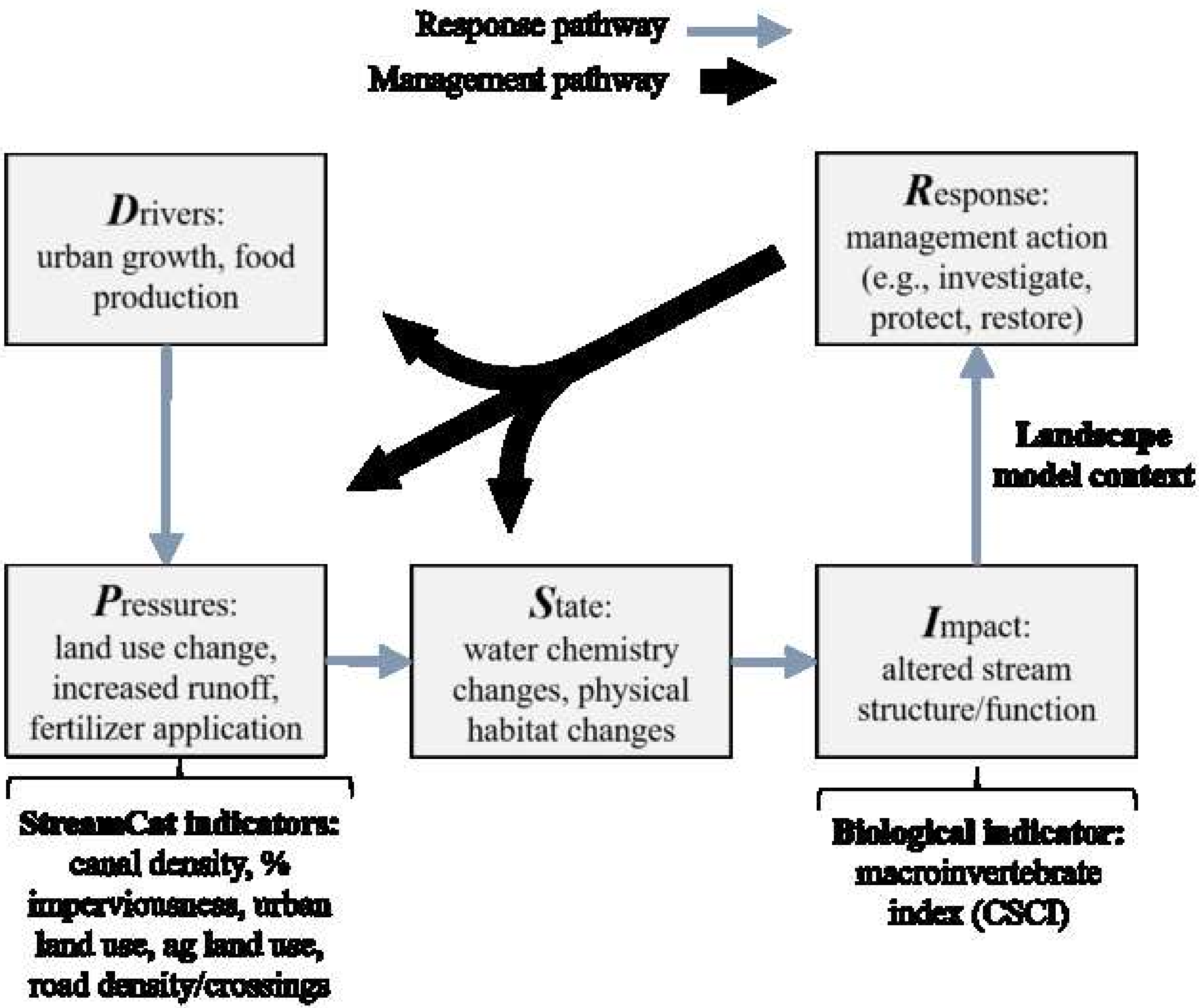
Waite, I. R., J. G. Kennen, J. T. May, L. R. Brown, T. F. Cuffney, K. A. Jones, and J. L. Orlando. 2014. Stream Macroinvertebrate Response Models for Bioassessment Metrics: Addressing the Issue of Spatial Scale. *PLOS ONE* 9 (3): e90944. https://doi.org/10.1371/journal.pone.0090944.

Walsh, C. J., A. H. Roy, J. w. Feminella, P. D. Cottingham, P. M. Groffman, and R. P. Morgan. 2005. The Urban Stream Syndrome: Current Knowledge and the Search for a Cure. *Journal of the North American Benthological Society* 24 (3): 706–23. https://doi.org/10.1899/04-028.1.

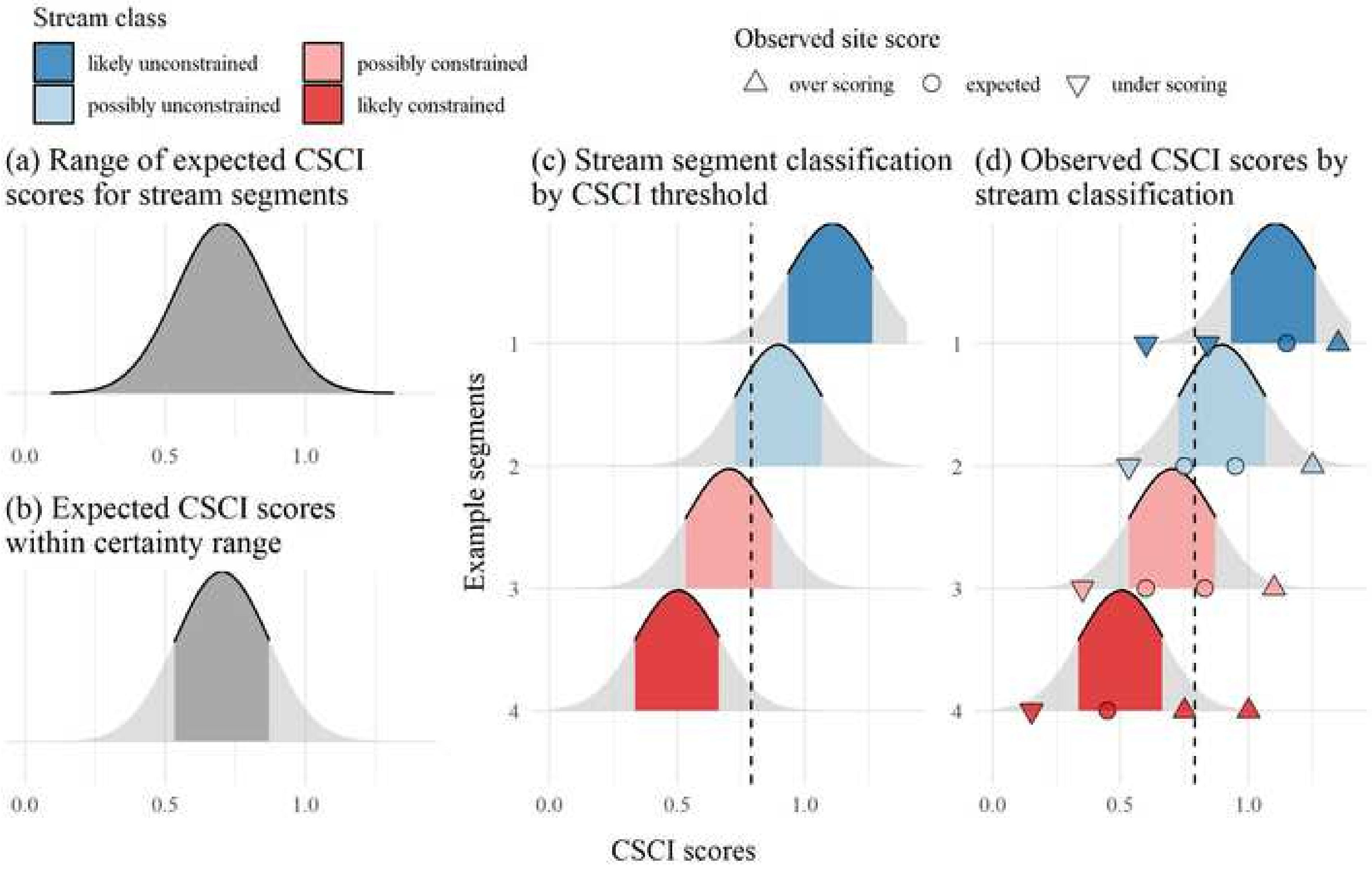
Wang, L., J. Lyons, P. Kanehl, and R. Gatti. 1997. Influences of Watershed Land Use on Habitat Quality and Biotic Integrity in Wisconsin Streams. *Fisheries* 22 (6): 6–12. https://doi.org/10.1577/1548-8446(1997)022%3C0006:IOWLUO%3E2.0.CO;2.

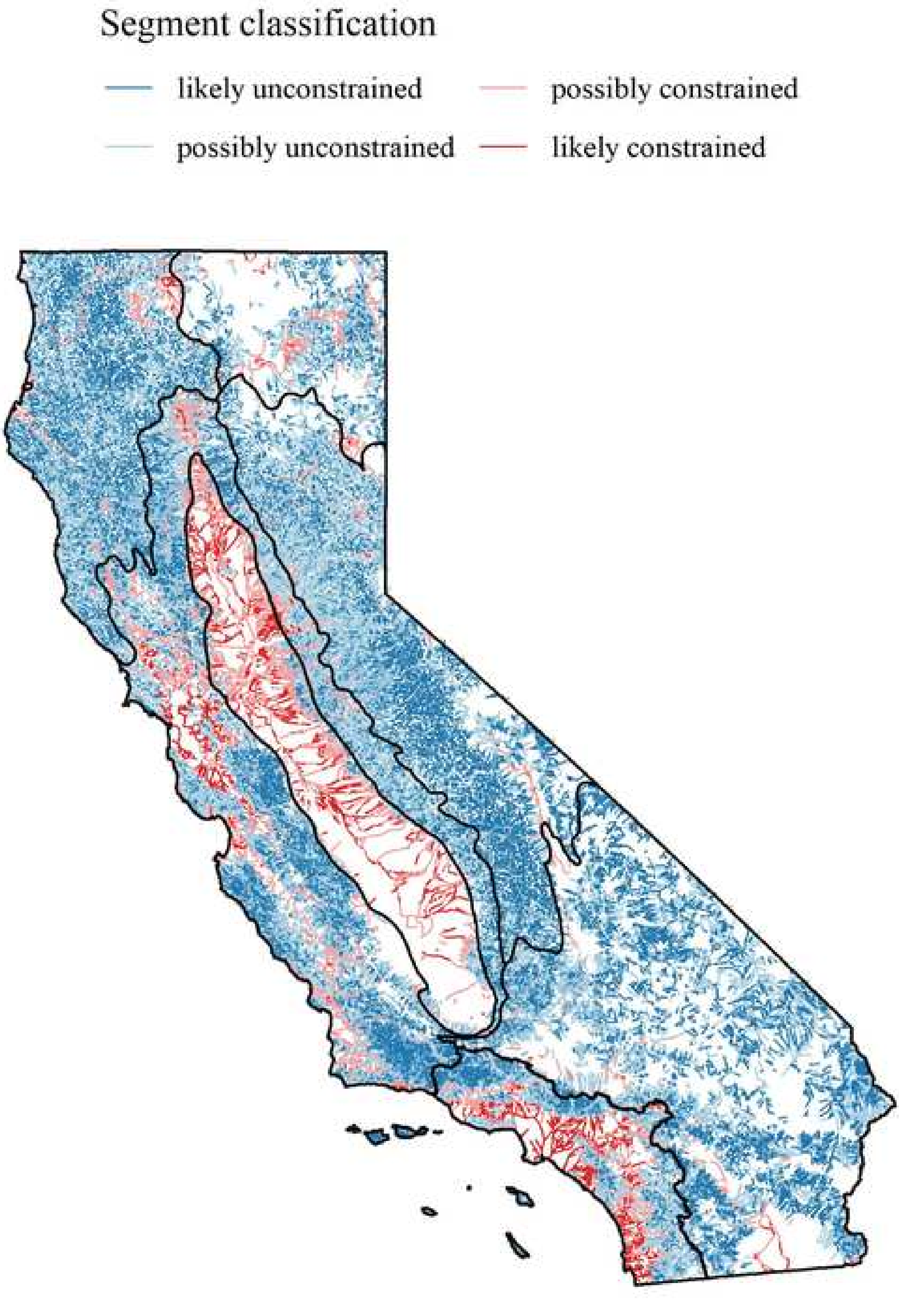

(a) Land use and regions
Agriculture
Urban
DM
NC
SN
CV
CH
DM
SC
0
150
300km
(b) CSCI scores
<1st: Very likely altered
10th to 30th: Possibly altered
1st to 10th: Likely altered
>30th: Likely intact

2

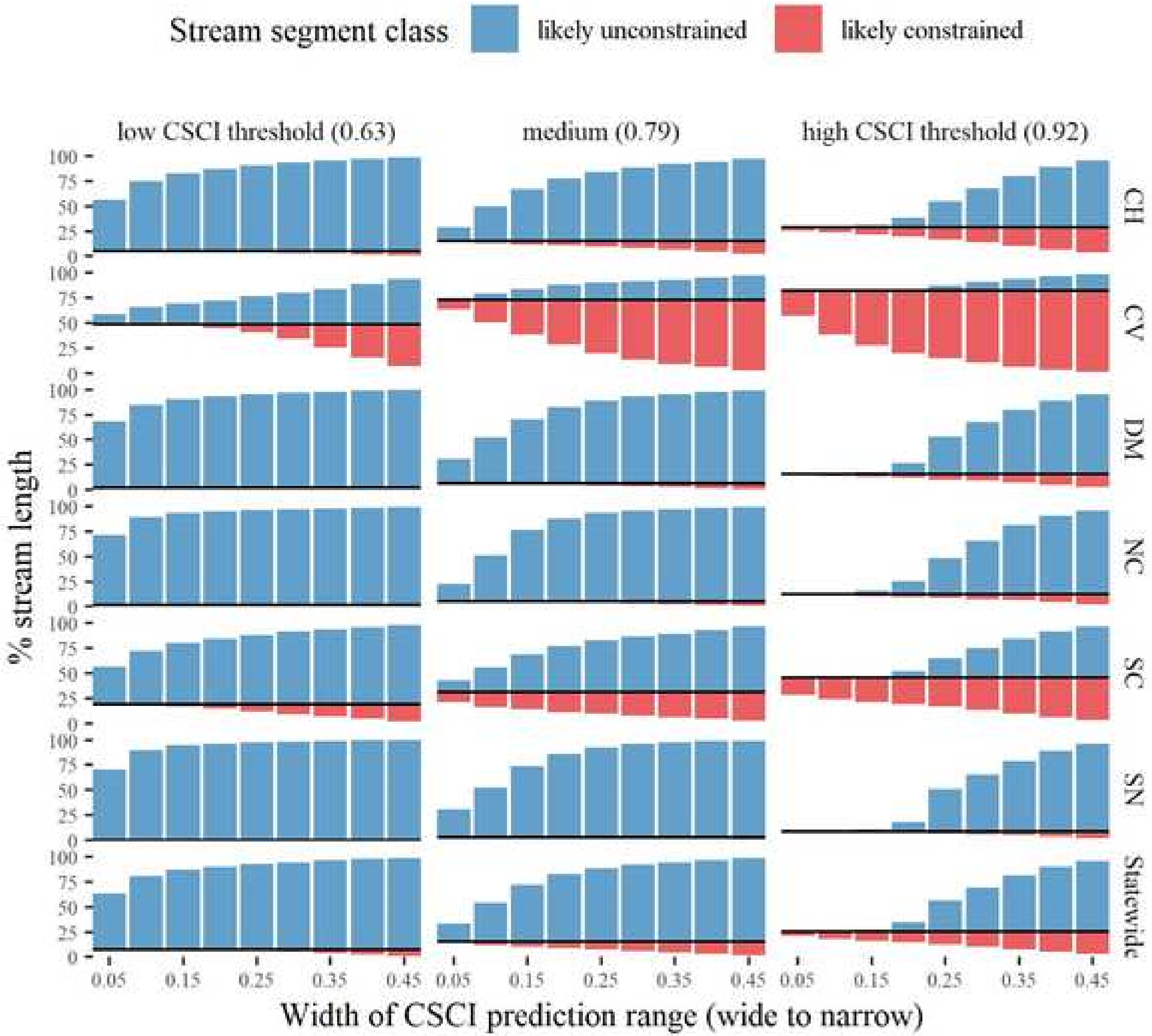

Response pathway
Management pathway
Drivers: urban growth, food production
Pressures: land use change, increased runoff, fertilizer application
State: water chemistry changes, physical habitat changes
Impact: altered stream structure/function
Response: management action (e.g., investigate, protect, restore)
Landscape model context
StreamCat indicators: canal density, % imperviousness, urban land use, ag land use, road density/crossings
Biological indicator: macroinvertebrate index (CSCI)

3


Stream class
likely unconstrained
possibly unconstrained
possibly constrained
likely constrained
Observed site score
△ over scoring
○ expected
▽ under scoring
(a) Range of expected CSCI scores for stream segments
0.0
0.5
1.0
(b) Expected CSCI scores within certainty range
0.0
0.5
1.0
(c) Stream segment classification by CSCI threshold
(d) Observed CSCI scores by stream classification
Example segments
1
2
3
4
0.0
0.5
1.0
CSCI scores


Segment classification
likely unconstrained
possibly constrained
possibly unconstrained
likely constrained

5



Stream segment class
likely unconstrained
likely constrained
low CSCI threshold (0.63)
medium (0.79)
high CSCI threshold (0.92)
CH
CV
DM
NC
SC
SN
Statewide
% stream length
100
75
50
25
0
0.05
0.15
0.25
0.35
0.45
Width of CSCI prediction range (wide to narrow)