CB TODO:

need to look at style guidelines and make changes for that.

Look at residual sites close to log ponds, what is different about them?

Do spell check on rmd docs.

RESPONSE

Thank you for the very constructive and thoughtful reviewer and editor comments. Much improved.

The editor recommended moderate revision on the basis of the manuscript requiring some additional validation of the new method we propose. We have now included a comparison to other multivariate methods and a cross-validation and updated the text accordingly. Details are below in response to Reviewer 2 and we thank the reviewers for suggesting this interesting addition. All other comments required only textual changes and most of these were minor. We have addressed each in turn below.

We have also archived the code on GitHub, as requested. We can provide the link in the manuscript or supplement once the need for double blind review has passed.

COMMENTS FROM HANDLING EDITOR

Line 59: do you mean ‘unobserved’ rather than ‘unobservable’?

Corrected

Line 81: need to provide some explication/explanation of the structural model depicted in Figure 1. (Note: the first reference to a figure in the text should be spelled out in full—e.g. ‘Figure 1’. Subsequent references to the figure can be abbreviated to Fig. X)  
Corrected

Line 92: something wrong with the notation here? (Kk is mentioned on lines 93 and 94, but on line 92 you have Ki?)

Corrected, it should have been Kappa[k,.] in the equation:

“where there are latent variables across the sites, is a row vector of coefficients and is a column vector of covariate values for each site. Latent variables for residual correlations have .”

Lines 128 on: At the introduction of the case study, it would be helpful to provide the reader with some basic details such as a location map, the size/area of Kia District, the areal extent of forests in the Kia District, and location or distribution of the 49 benthic survey sites. You could do this with reference to Figure 4. [Perhaps Figure 4 could be modified to indicate the boundary of the Kia District, some place names, and the distribution of forests in the Kia District?]

Corrected:

“We model change in cover of benthic habitats across a gradient of sedimentation in the Kia District of Isabel Province, Solomon Islands. Kia district has a human population of about 2000 people (Solomon Island Government 2009), and the historical extent of forests in our study region was approximately 79000 ha (Hansen et al. 2009), but logging in Kia district has removed 50% of primary forests over the last 15 years (Figure 4).”

We have added some place names to the map and an inset to show where our study region is located within the Solomon Islands. We have decided not to include a map of the forest, because creating an accurate map is beyond the scope of our study (it is not straightforward due to the potential of satellite images to confuse regrowth with primary forest). However, we have included the locations of log-ponds on our map, which are they key sources of sediment that have affected the reefs we have studied

Lines 151-154: I was initially confused by the description in these lines until I examined all the Appendices. I think it would help to explain that you started out with 31 habitat categories as per English et al. (1994), but also supplemented with two additional categories. But for analysis, these 31 categories were collapsed into 17 focal groups (see Appendix S3). [Alternatively, I would rename Appendix S3 as Appendix S1.]

Corrected: “Benthic habitats were categorised into 31 categories as per English et al. (1994), however we added categories for Acropora Branching Dead and Coral Branching Dead (Appendix S3).” Note we have retained this appendix as S3, because we now also cite S1 and S2 in the paragraph above.   
  
Line 175: in the text, you use the terms ‘high’ or ‘low’ flow, but in Appendix S2 you use the terms ‘strong’ and ‘mild’ in relation to flow. Please just apply one set of terms consistently across all the materials.

Corrected to strong and mild.   
  
Line 198: insert a comma after ‘surveys’

Corrected  
  
Line 224: a little elaboration on how you assessed bias would be helpful to the reader. [Alternatively, you could incorporate Table S2 from Appendix 5 into the text proper.]  
Corrected:

“We checked the nearest distance model for bias in its predictions of each habitat separately by comparing the observed versus expected values for habitat counts”

Line 271: I don’t think you need the ‘However’ at the beginning of the sentence.

Corrected: “Our findings are consistent with…”

Line 296: insert a comma after ‘particular’

Corrected  
  
Lines 333-334: Instead of the sentence you have here, can I suggest something like, “There are some technical challenges to further development of joint models and their use in conservation applications.”

Corrected: “There are some technical challenges to further development of joint models and their use for informing conservation decisions”  
  
Line 336 and 341: a priori (italicize, and no hyphen) instead of ‘a-priori’

Corrected  
  
Lines 361-362: Can I suggest instead, something like, “Our flexible joint modelling approach to estimating the impact of logging on lagoonal coral reef communities enabled prediction of community turnover…”

Corrected:

“Our flexible joint modelling approach to estimating the impact of logging on lagoonal coral reef communities enabled prediction of community turnover across a gradient of human impacts and of the extent of logging impacts to coral reefs”  
  
Line 368: Citations need to be brought in line with the Conservation Biology format. Please refer to the Author Style Guide when fixing up the references.

Citations have been corrected.   
  
Table 1 caption: A bit more detail would be useful to a reader. For instance, “Comparison of four candidate models using the Watanabe-Akaike information criterion (WAIC)…” Remind the reader what the Bayesian ordination model and what the constrained model includes so they don’t have to refer back to the text for that information.

Corrected: “Comparison of candidate models using the predictive ordinate and the Watanabe-Akaike Information Criterion (WAIC).”  
  
Figure 2 caption: add “credible intervals (CIs)”

Corrected  
  
COMMENTS FROM REVEWER 1

The ms is very well organized and well written except for some lapses in grammar and word use, mostly in the Methods and Results sections, that can be easily fixed.

Thank you for the helpful feedback. All the comments that follow are minor editorial changes, we have corrected them all in the manuscript:

1. Introduction, line 37:  Change “to” to “at”.  
   2. Line 63: Rephrase “Around Kia sediment run-off …” to “Sediment run-off around Kia …”  
   3. Methods, line 84: Need to carefully restructure the equation statement so that the “Equation 1” label and comma, followed by lower-case “w” in “where”, conforms with Biol Conserv formatting protocols. Ditto for Equation 2 (Line 92) and Equation 3 (line 167).  
   4. Line 100: Change “to” to “on”.  
   5. Line 101: Replace closing parenthesis with semicolon after “change”.  
   6. Line 132: Insert “and between “stored” and “then”.  
   7. Line 152: Insert semicolon after “(1994)” and a comma after “however”.  
   8. Line 153: Further specify “in deep water near” as “in deep lagoonal water adjacent to the site prior to commencing benthic surveys (Hamilton et al. 2017).”  
   9. Line 155: Restate “we were most interested in” as “with which we were most interested”.  
   10. Line 163: Insert “that” after “ensure”.  
   11. Line 171: After “identify” rephrase as “the best model, followed by visualisation of mean …”  
   12. Line 174: Change “dividing” to “divided”.  
   13. Line 182: Delete comma.  
   14. Line Line 183: Correct spelling of “visualised”.  
   15. Line 197: Replace “closer to” with “nearer”.  
   16. Line 202: Insert comma after “new” and strike “logging that has”.  
   17. Line Lines 203-204: Remove abbreviations of latitude and longitude from superscripts of degrees and change to regular font capitals that follow their respective degree values.

\* 18. Results Line 216: Figure 1 is not cited; logically, it should be referred to here, perhaps with brief mention that it provides a schematic summary of patterns attributable to the first and second latent variables.

Agreed. We think the reviewer is referring to Figure 2 (figure 1 just gives a schematic of the model’s structure and has no new results). We have not added any additional interpretation of the figure, because its interpretation is given below and in the figure legend.

All the comments that follow are minor editorial changes, we have corrected them all in the manuscript:

19. Line 216: Strike “However,” and rephrase as “We therefore proceeded”.  
20. Line 219: Delete “that” and the comma immediately preceding it.  
21. Line 223: Replace “further” with “farther” when referring to distance. Ditto re Fig. 2 caption.  
\* 22. Line 231: Fig. 4 is cited here along with Fig. 2a in support of assertion that the constrained latent variable represents a gradient from high to low benthic complexity. I agree that the Fig. 4 map of the probability field should be cited here, but it should be shifted forward in position and relabeled as Fig. 3.  
23. Line 234: Correct misspelled “also”.  
24. Line 235 (twice) and line 238: Relabel all mention of Fig. 3 as Fig. 4 and re-order as 4th in sequence of text figures.  
25. Lines 235 & 254: Replace comma with semicolon before “however” and insert a comma after it.  
26. Line 235: Replace “change increase closer” with “increase in cover closer”.  
27. Line 237: Replace “near to” with “nearer”.  
28. Line 239: Replace “”a positive association” with “positive associations”.  
29. Line 242: Replace “Mean estimates” with “Means estimated” and “at taken”.  
30. Lines 247-248: Insert “the” before “footprint”; delete the comma that follows the 5th word after it.  
31. Line 249: Replace “near to” with “nearer”.  
32. Line 252: Replace “close to” with “nearer”.  
33. Line 257: Replace “with” with “that” .

\* 34. Line 258: Insert upper limit of estimated impact area—“179 to ? hectares”.

Agreed. 179 is the upper limit, so we have added “or up to 179 hectares”.

\* 35. Discussion, lines 310-311: Suggest mentioning, in order to provide a greater perspective, the concept of “space-for-time substitution” (Pickett, S. T. A. 1989. Space-for-time substitution as an alternative to long-term studies. In Long-term Studies in Ecology G.E. Likens, ed.), pp. 110–135. New York: Springer-Verlag). This long-recognized approach continues to be used to identify and predict impact in a more timely and cost-effective manner (DeMartini et al. 2013).

Agreed. Added: “For instance, time-series data may provide more accurate, but more expensive estimation of logging impacts than the ‘space-for’time’ substitution we used here (Pickett 1989, DeMartini et al. 2013).”

All the comments that follow are minor editorial changes, we have corrected them all in the manuscript:

36. Line 331: Suggest replacing the first “precise” with “localized” and the second “precise” with “finer-scale”.  
37. Line 336: Replace “amount” with “magnitude”.  
38. Line 342: Shift lead “However” to between “model” and “is”.  
39. Line 355: Replace “their” with “there”.

\* 40. Lines 361-363: The first sentence is redundant with the lead sentence of the Discussion. Both sentences on lines 361-363 are out-of-place and should be used (re-worded) as the first several sentences of the Discussion. The concluding paragraph of the Discussion would then need a new lead topic sentence. I suggest something like: “Our case study demonstrates the successful application of a joint model to identify and predict impacts on coral reef habitat.” Continue with “More generally, joint models offer a useful tool for … additional reef and other ecosystems …”.

Agreed. This sentence has been corrected as per the comment from the Handling Editor.

All the comments that follow are minor editorial changes, we have corrected them all in the manuscript:

\* 41. Figure 2: Panel (a) on the far left lacks its label; ditto for panel (b) and for panel (c) on the far right.  
42. Figure 3 caption: Replace “Significant was” with “Significance”.  
43. Appendix S5 Model Evaluation, Table S2 caption: Correct typo “now” to “no"

COMMENTS FROM REVIEWER 2  
  
The main thing that is missing from the paper at the moment is some evaluation of the method proposed here – how do we know that it works?  Examples of what could be done are to include a simulation (studying how reliably it recovers the underlying pattern, and checking it converges to the “true” answer as sample size increases) or using training/test splits and studying predictive performance, to compare across different models.  Or a more arm-wavy comparison to something like GDM, to compare in terms of how informative the different tools are at informing conservation strategies.

Thank you for these helpful suggestions. They are all excellent suggestions for evaluating the method. We have provided an evaluation of the model’s predictive ability using leave one out cross validation and we also compare the model’s output to two other multivariate methods that are commonly used to analyse how community structure varies across environmental gradients (MDS and CCA). We have not include a simulation study at this stage, due to technical challenges in the automatic evaluation of 1000s of model fits (we explain why below). We have added to the text (Methods):

“To confirm the Bayesian approach was consistent with existing methods, we also compare its ordination of habitats and sites to those obtained from two widely used approaches to multivariate analysis (Legendre & Legendre 2012). The first was non-metric multidimensional scaling which performs an unconstrained ordination of sites that we relate to the environmental covariates post-hoc. The second was constrained correspondence analysis, where the ordination of communities at sites was constrained by distance to the nearest log pond and flow conditions.”

And to the Results:

“Results for the site ordinations and habitat ordinations from the Bayesian model were comparable to CCA and MDS, with the exception of Halimeda algae (Appendix S6). Halimeda algae was found to make a stronger contribution to the water quality gradient in the Bayesian model than the CCA (Figure S5). There was generally higher cover of Halimeda algae closer to log-ponds, but counts of this habitat were highly variable and included many zeros. The Poisson error structure used in the Bayesian model may have been better able to model the noisy response of Halimeda algae log ponds than the CCA. “

And to the Discussion:

“We also found the Bayesian model was better able to attribute noisy gradients in habitat cover to pollution than traditional methods like CCA. Improved statistical power in the Bayesian models over traditional methods may be because they explicitly model the mean and variance of abundance counts (Hui et al. 2015).”

We have added the complete comparison as Appendix S6.

As suggested we have also used training/test splits to evaluate the model and have performed leave-one-out cross validation, adding the full results and technical details of the method to Appendix S6. IN the methods we state:

“Additionally, we used leave-one-out cross validation to evaluate the model’s fit to each habitat type (Appendix S5).”

And in the results:

“The cross-validation also supported models with 1-2 unconstrained latent variables for those habitats that were related to the water quality variables (Figure S2).”

We were unable to include a simulation study because we had issues with automatically confirming convergence of the Bayesian model across 1000s of fits to simulated data. In brief, latent variable models are under-determined, an issue that has been previously documented in Hui et al 2015. This in itself is not an issue, however, the parameter estimates in the MCMC chains require careful checking to ensure that parameters do not switch signs part way through the MCMC draws, effectively giving parameter estimates that are the mirror image of earlier estimates. It is not valid to estimate the posterior distribution for a parameter across an MCMC chain that has switched. We manually checked the MCMC chains for the models we fit to the data to confirm that switching was not an issue. However, we have tried and have been unable to develop an efficient automated procedure to address the switching issue in simulated data where we need to evaluate 1000s of MCMC chains. Therefore at this time we have chosen not to include a simulation study as some technical aspects of automating model checking are required for it to provide reliable results. It was straightforward to manually confirm that switching had not occurred in the relatively few model runs we performed for the main analysis and in the cross-validation.

Line 111 attenuation

Corrected

Line 117: this is a little confusing as written, why not say you report predicted probabilities of occurrence (if you want more details, explaining they can be found directly from the fitted model, by inverse-probit-transforming the linear predictors).

Corrected

Line 157: rare categories do not necessarily imply zero-inflation, they imply that the mean is small! (A Poisson with mean of 0.1 is more than 90% zeros)

Corrected and removed reference to zero-inflated models:

“We also aggregated categories to avoid having numerous infrequently recorded categories that had many zeros and contain little information about environmental gradients. ”

Lines 181-187: another possible model to compare to is one where the predictors were applied to response rather than to latent variables.

We agree that it would be interesting to conduct a thorough comparison of our approach and that of Hui 2016, however, we feel this is beyond the scope of this article. We have not included these additional simulations because such a model does not allow one to map the hidden gradient of benthic communities at different distances from logging, which was our main objective.

Combinations of constrained and unconstrained latent variables were used, with just one constrained variable.  Is there a reason why this was done, rather than specifying multiple constrained latent variables?  Because the constrained variables were defined as N(kX,1), can’t they be understood as a linear predictor operating on the latent (reduced rank) scale kX plus an unconstrained latent variable N(0,1)?  Multiple variables of this type could be added to the model and the data could sort out for itself the extent to which these should be constrained, through choice of parameters k.  
We agree that one could add multiple constrained latent variables and let the model sort out which gradients are the strongest drivers of community structure. We have added this suggestion to the discussion:

“Future studies could extend our joint model to include multiple environmental gradients as multiple different constrained latent variables. If there are hypothesized but unmeasured gradients, models of spatial autocorrelation can help obtain more accurate inference…”

Because we do not have a hypothesis for multiple gradients in our study region (as explained in discussion), we have not included multiple constrained latent variables at this point.

Yes, another way of interpreting the mean of the constrained latent variable is as a linear predictor on the reduced rank scale. We can include this explanation in the text if the reviewer requests, however we have not done so at the moment, to avoid technical terms that non-statistical readers may find hard to interpret.

Lines 339-341: why not use the data to choose the number of constrained and unconstrained LV’s?  e.g. using DIC?  This is an advantage of model-based approaches, that you can use standard model-based machinery to make analysis decisions, instead of imposing decisions a priori.

Thank you, this is a helpful suggestion that has helped us better explain one of the technical challenges we discuss in the Discussion. We chose to use the WAIC for model selection, which is similar to the DIC, but has recently been shown to have better theoretical support (Vehtari et al. 2017, cited in text).

The reason we chose to proceed with 2 unconstrained latent variables was because we suspected additional variables were overfitting the data, and weakening our ability to predict to unsampled sites. We have reworded the discussion to explain this issue:

“For instance, we found that the inclusion of greater numbers of unconstrained variables hid the effect of turbidity gradient. Inclusion of too many unconstrained variables may enhance the model’s fit to the data, but come at the cost of lower generality – because unconstrained variables cannot be predicted to unsampled sites. Therefore, we recommend the analyst select the number of constrained and unconstrained variables a priori (Hui et al. 2015), but further work is needed to test how accurate models with multiple unconstrained variables are when predicting to new regions.”

Figure 2 x-axis label has a typo

Corrected.

Figure 3 presents the modelling results but not the raw data – would it work to add some fine-grain spots, so we can see the original observations?

Corrected.

Figure 4 caption – this is actually the inverse of the probit transform.  Maybe just say back-transforming to predicted probabilities (or drop the transform reference altogether?  Fitted values are a pretty standard thing in GLM-type models)

Corrected, removed reference to transformed variables.

Line 373 there is a problem with authors for this reference.

Corrected