Responses to reviews of "How predictable is extinction? Forecasting species survival at million-year timescales"

Smits, Peter
psmits@berkeley.edu

Finnegan, Seth
sethf@berkeley.edu

Below are our responses to the editor and reviewer comments that we recieved on our original manuscrpt. We would like to thank the editor and the three reviewers for their thoughtful and insightful comments which have greatly improved the manuscript.

Editor and reviewer comments are presented in bold. Below these comments, we detail how we updated our manuscript in response to that comment.

1 Editor comments

Line 5 (abstract): comma after to answer this question

Line 7: period after parenthesis

Line 10: comma before the improvement in predictive power

Line 25: the suggestion or suggestions

We have updated the relevant sections of the text with these suggestions.

Line 31-33: I think this sentence needs to be clarified. Can you describe more clearly what is meant by how accurate or strong our predictions about the differences in extinction risk are, and how this differs from strong and general determinants of extinction risk?

In our general reorganization of the introduction section of this paper we tried to improve the clarity of our language. We thank the editor for pointing out this sentence as it was redundant and confusing.

Line 34: try to avoid ending a sentence with a preposition

Line 40: among taxa, rather than between?

Line 41: comma before and the relative risk of

Line 77: need space before Occurrences were filtered

Line 78: was no moreis that right?

We have updated the relevant sections of the text with these suggestions.

Line 83: what about taxa with age ranges that were more uncertain than 1 mys? Did this ever occur? [Ah, I read further and now see your call to the SI]

Thank you for pointing this issue out. We have tried to improve our references to the supplemental material by putting more references in the text instead of just a single reference at the end of the section.

Line 86: I echo Wolfgangs comment in wondering how congruent would be patterns if they would be measured using another range-based metric?

Thank you both for the suggestion to move to minimum spanning tree measure of geographic range. We have updated our analysis using this new range metric. The change in range metric did not change our interpretation of the results.

Line 115: insert in prior to Section S1.3

We have updated the relevant section of the text with this suggestion.

Line 114(ish): have you considered testing change in geographic range over multiple intervals? I.e., maybe a stronger signal would be observed if there was a persistent decline in range size over multiple 1 my time bins? Similarly, species may be buffered from extinction if their range increases over multiple time bins. It would be interesting to tease whether this is more important than absolute size

Thank you for this suggestion. We have update our analysis to include change in geographic range from the last three observations. This change adds two more covariates to all models that include our "historical covariates." This addition to our analysis did change our results, and we now consider our most complex model that includes historical covariates and variation in covariate effect over time to be overfit to our data. Our models that include either historical covariates or variation in effects over time, however, are found to have functionally equivalent out-of-sample performance.

Line 119: try to avoid ending a sentence with a preposition

Line 183: Comparing?

Line 195: I am probably being dense, but how can the in-sample have a poorer performance than the out-of-sample? Isnt this fairly counter-intuitive?

Line 214: forecast vs forecasting?

Line 224: an before extremely important

We have updated the relevant section of the text with this suggestion, in particular focusing on clarity.

2 Reviewer comments

2.1 Reviewer 1

Smits and Finnegan use the Neptune database to check which combination of potentially influencing variables has the best predictive power of extinction risk in the marine plankton. They use sophisticated logistic modelling of the time-binned raw data to suggest that considering changes of parameters only marginally improves prediction accuracy. Overall, this is a very well written, although difficult to comprehend, account, which is to the point and rich in insights.

While the main result is well founded by the sophisticated analyses (in fact too sophisticated to be understandable in all details by the reviewer), it is somewhat disappointing, as the relative importance of parameters in multivariate models remain vague. It is also disappointing to see that change of range had a negligible effect in the multivariate context, as this is one of the main arguments for using the fossil record in extinction risk assessments (besides taxon age and intrinsic turnover rates). I argue that, to really discard change of range, the authors would need to look at longer trajectories of range. Bin-to-bin changes at this temporal resolution (1 myr) are perhaps too noisy to inform models. I am not asking for additional analyses for this paper, but this point should be added to the discussion.

Thank you for this suggestion. We have update our analysis to include change in geographic range from the last three observations. This change adds two more covariates to all models that include our "historical covaraites." This addition to our analysis did change our results, and we now consider our most complex model that includes historical covariates and variation in covariate effect over time to be overfit to our data. Our models that include either historical covariates or variation in effects over time, however, are found to have functionaly equivalent out-of-sample performance.

We are confronted with four models and even the simplest one is already quite complex. Is there a chance to provide for each model the relative importance of each variable for the model performance? Alternatively, as standing geographic range is among the best predictors of extinction risk in paleo studies, can the authors add a very simple model of just event range? Id be curious to see if the AUC is then significantly lower than in the other models. Of course, in a frequentist world, I would start with a complex model and then perform model selection with a step function. Im not Bayesian enough to suggest a specific strategy for this paper, but somehow model performance needs to be balanced by model complexity. The authors seem to have compared just model performance.

Thank you for pointing out this issue in our text and we apologize for the lack of clarity. We have updated our methods section to further explain cross-validation and how we have implicitly dealt with model complexity because of the linked definitions of cross-validation

and AIC. From our updated text:

"Cross-validation is a procedure for estimating a models expected out-of-sample error. Information criteria such as AIC or WAIC are approximations of out-of-sample predictive error as estimated by cross-validation [1, 2]. Cross-validation implicitly takes into account model complexity because when a model is overfit to its data, out-of-sample predictions will be biased and inaccurate [1]. A high degree of similarity between out-of-sample and in-sample estimates indicates that the model is not overfit to the data (though it is not necessarily an adequate descriptor of the data). Cross-validation is preferable to simple metrics such as AIC because instead of a single value it produces, an entire posterior distribution of estimates."

The methods are very well explained (especially in the supplement), although uncertainties in parameter estimates merit some further prose. Using maximum great circle distance (GCD) as a measure of geographic range is ok but prone to sampling bias. Better sampled intervals will inevitably have greater mean ranges than more poorly sampled intervals. While the scaling applied by the authors takes different means and variance into account it may not be sufficient. If GCD is used, Id recommend normalizing to the maximum observed geographic range in an interval prior to scaling. Grid cell occupancy or minimum spanning trees are alternative measures, which are worthy of exploration.

Thank you both for the suggestion to move to minimum spanning tree measure of geographic range. We have updated our analysis using this new range metric. The change in range metric did not change our interpretation of the results.

We chose not to standardize geographic range to in-bin max because we believe this breaks "coherence" of the measure and the interpretation of the regression coefficients. With geographic range not standardized by max in-bin geographic range, the regression coefficients for the "change in range" covariates are the expected change in log-odds of extinction risk per standard deviation change in geographic range between that previous bin and the current bin. In contrast, if the variables are defined as percentages, the difference between bins gives a less obvious or interpretable regression coefficient. Additionally, max in-bin geographic range is nearly identical for the entire Cenozoic.

Looking at change of range is good, but the authors could also look at changes beyond two intervals. Although models incorporating multiple bins may result in overfitting, the argument is that long-term decline will increase extinction risk. Id argue that two bins (especially at the relatively fine temporal scale applied) can hardly measure long-term decline. Finally, the authors should provide a rationale why these and not other parameters were used in the study (see also below). Does latitude or skeletal mineralogy have little to add to model performance?

Thank you for this suggestion. We have update our analysis to include change in geographic range from the last three observations. This change adds two more covariates to all models that include our "historical covariates." This addition to our analysis did change our results, and we now consider our most complex model that includes historical covariates and variation in covariate effect over time to be overfit to our data. Our models that include either historical covariates or variation in effects over time, however, are found to have functionally equivalent

out-of-sample performance.

We have modified our discussion sections to further explain the rational behind our covariate choices. Simply, our goal is to provide a baseline for extinction prediction that can be compared to more specific models which include these more specific covariates. Additionally, a lot of the suggested covariates are either not available for a lot of the taxa throughout the Cenozoic or are nearly constant for the entire taxonomic group. In many ways, by allowing our parameter estimates vary by taxonomic group we have attempted to implictly deal with these types of ecological differences.

Presentation of results could be improved. Showing raw data of extinction rates in the four groups, and individual analyses how age and phyla actually influence extinction risk would be preferred over the wiggly lines in Figs. 2 and 3. The latter could be moved to the supplement. The text in the results section is very technical, describing too much detail.

Thank you for this suggestion. We have added a section to our supplemental material that includes figures presenting the parameter estimates for the group-level regression coefficients and the individual-level estimates for those regression coefficients over time.

The discussion is good, but may have to be partly rewritten in light of my comments.

Specific comments (line numbers refer to the smaller ones on the right):

- the relative risk of extinction exhibited by different taxonomic groups could be abbreviated as intrinsic extinction risk

- l. 68-69. Partial repetition of sentences before. Incredible should be omitted. If incredible why trust the age models?

- l. 79. Ecologically unusual should be omitted from reasoning. Not needed.

We have updated the relevant sections of the text with these suggestions.

- l. 88-91. Authors may want to add that Mg/Ca has its own issues, especially the steep rise of Mg/Ca in seawater through the Cenozoic. Cramer et al. have discussed and accounted for this trend, but with some assumptions, which result in considerable uncertainties.

Thank you for pointing this out. We have updated our Methods section with a discussion of the potential problems with Mg/Ca based temperature estimates.

- l. 132: paragraph starts with k-fold but is specified for 5-fold. I recommend rewriting this paragraph for the general application and then specify.

- l. 149-150: so we interpret values between 0.7 and 0.8 could then be considered so we consider values between 0.7 and 0.8 as

- Caption Fig. 1: Cut a from with a higher AUC values. It is not entirely clear what the error bars and bean plots represent. Are the AUC values derived from the 62 time bins or is there more to it? Please explain.

- l. 173: Dinoflagellates should be diatoms unless I have missed something. Same in labels of Figs. 2-4 and some incidences in the text. Please correct throughout.

- Xlab of Fig. 2 should be Mya or similar instead of CI

- l. 176-177. Repetition from methods.

- l. 183. Comparison Comparing

- l. 192-193. No single-sentence paragraphs

- l. 197. diatoms instead of Dinoflagellates

- l. 214-216. To my understanding AUC does not simply measure a rank-order analysis but model performance. That the AUCs of the eight models are statistically indistinguishable (are they?) does not necessarily imply a correct forecasting of rank-order extinction probability. Please rephrase!

We have updated the relevant sections of the text and our figures with these suggestions.

- l. 223-224. Variation of extinction intensity. Would be useful to either depict this variation or report some summary statistics (median, range, variance).

Thank you for this suggestion. We have added a section to our supplmental material that includes figures presenting the parameter estimates for the group-level regression coefficients and the individual-level estimates for those regression coefficients over time. These estimates include average log-odds of extinction over time.

- l. 233. What does scientifically significant mean is this context? I am aware that simple model selection is not applicable here but it would be good to have a comparable metric.

Thank you for pointing out this lack of clarity. We have removed this section and attempted to improve the clarity of our discussion of our results.

- l. 236-237. Kiessling and Kocsis and used more than just corals. Their coarser temporal resolution (geological stages) is an obvious suspect for the greater role of change of occupancy than reported here.

We have update our covarites to include additional lags. Please see the updated Methods, Results, and Discussion for how our results have changed.

- l. 244-245. Poor reasoning. For example skeletal mineralogy (silica versus calcium carbonate), trophic level (phyto- vs. zooplankton), latitudinal preference (e.g., median paleolatitude) or additional physico-chemical parameters (e.g., sea-level, Mg/Ca, pCO2) are intuitive variables, whereas taxon-age is less intuitive in light of the recent literature. Im not asking for additional analyses or more complex models but a sound rationale.

We have modified our discussion sections to further explain the rational behind our covariate choices. Simply, our goal is to provide a baseline for extinction prediction that can be compared to more specific models which include these more specific covariates. Additionally, a lot of the

suggested covariates are either not available for a lot of the taxa throughout the Cenozoic or are nearly constant for the entire taxonomic group. In many ways, by allowing our parameter estimates vary by taxonomic group we have attempted to implictly deal with these types of ecological differences.

- l. 247. Add for between not exist and all. In any case, a poor defense given the obvious traits to look at (see above).

- l. 254. Modern conservation determinations are not continuous but ordinal. The difference between critically endangered and endangered may be very different from the difference between endangered and vulnerable.

- l. 255-256. I argue that the authors results actually show that fossil data are not very relevant. If this impression is wrong, the authors should perhaps rephrase some parts of the text.

Also check use of My vs. Mya and cases in plankton groups. As a general rule these should be upper case in a formal context (e.g., Radiolaria) and lower case when used informally (e.g., radiolarians).

We have updated the relevant sections of the text with these suggestions.

The supplement is good but lacks sensitivity tests and a rationale of why the specific parameters and not others where chosen in the models.

Thank you for this suggestion. We have improved the language in our methods section surrounding in-sample verus out-of-sample forecasts. Our in-sample forecasts are a type of sensitivity test.

From the updated manuscript:

"In-sample forecasting is a posterior predictive check in that we are estimating our model's ability to correctly classify the data to which it was fit. Posterior predictive checks are a type of sensitivity analysis because we are checking the quality of model's fit to the data. If our models have poor in-sample forecasting performance, than our models are not adequate descriptors of the data and will most likely make poor out-of-sample predictions. In-sample forecasting measures, however, are not the same as understanding our models' ability to forecast future extinctions or if our models are overfit to our data and produce biased out-of-sample estimates [1]."

And:

"The in-sample model comparisons are useful for comparing the relative ability of our models to represent the data they were fit to, acting as quality control and sensitivity analysis. Comparison between the posterior distributions ..."

2.2 Reviewer 2

This manuscript uses a high-resolution record of marine microfossils to test the predictive power of fossil record extinctions. A lot of paleontological research has focused on determining the importance of particular biological and environmental risk factors, so this study is novel in testing whether past history can predict future extinction. The methodology is explained well and the main result is well-supported.

I only have two broader thoughts and two more specific notes:

I find it interesting that the predictability of extinction doesnt obviously differ between background intervals and intervals with more unusual environmental perturbations (such as the PETM or E-O transition). Perhaps thats because those events werent mass extinctions in the vein of the P/T or K/Pg. Theres also a fair amount of volatility in the AUC time-series. Nevertheless, does this consistent predictive ability regardless of differences in conditions say anything interesting about extinction predictions?

The caveat that human impacts may dramatically alter extinction risk seems like an important one and, as it stands, seems like it could undercut the (rather terse) final conclusion. Would conservation decisions really be bolstered by including fossil data? I would agree with that (and of course I would, as a paleobiologist), but it might be useful to expand a bit on this topic. Do you mean that a model such as yours, incorporating geographic range and other such parameters, would be helpful? In what way could it be used? Or do you mean using models based on past intervals with environmental changes inferred to be similar the PETM as an analogue for warming and acidification, for example? I wonder if it would be helpful to draw a stronger connection linking your finding to conservation biology, given the different stressors now.

WAITING FOR SETH

I found the description on lines 171-174 to be a bit confusing. The phrase this pattern (line 171) seems to refer to the posterior distribution less than/equal to 0.5. The sentence states that the pattern is absent in absent in VP model for forams and radiolarians, but to me the radiolarian pattern doesnt look too different from the V model (which also doesnt dip below 0.5). The next sentence argue that there are fewer period of low performance for calcareous nannos and dinoflagellates in the VP model, but the differences seem minimal. Perhaps Im not familiar with these graphs, but there only seems to be a difference in calcareous nannos around 50 Ma, and among dinoflagellates a small difference around 58 Ma and 30 Ma?

Thank you for pointing out this lack of clarity in the Results section. We have revised this section for clarity.

Figures 2-4. I guessed (and later confirmed) that the alternating gray and white

bars represented stages of the geological time scale, but it would be helpful to indicate that in the captions.

We have updated the relevant sections of the text with these suggestions.

2.3 Reviewer 3

This is an interesting paper on extinction prediction using microfossil plankton data from deep-sea cores.

Its very important that they found simple model with only a few parameters of geographic range, temperature etc can robustly predict extinction risk regardless of minor model difference.

I enjoyed my reading and hope the comments below help to improve the ms.

1. just in case, for foraminifera, benthic species are deleted from the database? Or they originally include planktonic foram only?

We have confirmed using MicroTax that all foram species included our data set our planktonic species only.

2. Neptune database has an internally consistent taxonomic identification strategy. Is this true? I guess its non-critical compilation of census data done by shipboard micropaleontologists over 50 years (ie taxonomic sense of different generation micropaleontologists are very different). David Lazarus (in charge person of Neptune) is radiolarian person, and so radiolarian may be better in taxonomy? Planktonic foram diversity is low, so may be better straightforward. Anyway, I recommend deeper reference search and reading to see how is Neptune taxonomy and explain it in the method section more in details.

We have added a statement in our introduction discussing that the Neptune database has an internal synonymization framework as detailed in Lazarus [3]. While the taxonomic identities may not be the most accepted by the community, they are consistent within the database.

3. Fig 1, better to explain what are C, V, CP, VP, AUC in the caption

4. Fig 2 and 3, What do gray bars mean? Gray bay = envelope?

5. It will help our understanding if you explain the main point/message of each fig in the contest of discussion in each caption

Thank you for these suggestions. We have updated our figure captions to be more descriptive. Additionally, we have clarified the link between our figure elements and our model description table.

6. I am not sure if its possible, but I am curious where geographic range or temperature are the key? Or both are essential? It is possible to try geographic range only model and temperature only model to see such?

Thank you for your suggestion. We chose to always included geographic range and temperature because these represent basic measurements of a species and its environmental context. Additionally, if we were to use models with one of these covariates and not the other, then we are assuming that the ignored covariate is known with 100% probability of having 0 effect on extinction risk. We are uncomfortable making this assumption and instead Instead we use weakly informative priors to limit spuriously strong parameter estimates and "shrink" effects towards 0 if that covariate has a weak outcome on the response.

We have also added a section to the supplement with the group-level and individual-level estimates for all of our regression coefficients to demonstrate the size of our covariate effects.

7. The authors use deep-water temperature (using geochemistry of benthic foram from deep-sea cores) data. And their biotic data is plankton. Their plankton can be mostly from ocean surface or photic zone (eg diatoms), but some have good bunch of deep-water species (radioralians).

The authors need some discussion and justification on this.

We have added a section to the Methods discussing this data limitation

From the updated manuscript:

"Mg/Ca based temperature estimates are measured from benthic forams, and are an estimate of deep water ocean temperature. The organisms in this study are all planktonic, Mg/Ca based temperature estimates do not describe the exact environment these organisms inhabit. Ideally, we would have detailed ocean surface water temperature estimates for the entire globe for the entire Cenozoic. Unfortunately, that type of data does not exist. So, we interpret our temperature estimates as reflecting the global climate state that an organisms experiences, and not as a descriptor of that taxon's local environmental ecology."

8. While geographical range consider spatial stuff, the authors use one temperature curve for all (I understand there will be no other way). Is there any difference in temperature sensitivity between narrow and wide distribution species; or low latitude and high latitude species? I understand this may be out of the scope. The authors may ignore.

We thank the reviewer for this comment. We do agree that surface temperature estimates, especially geographically localized ones, would be a major improvement to our data. However, detailed sea-surface estimates for the entire Cenozoic are not readily available and generating those would be outside the scope of this study.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. 2nd ed. New York, NY: Springer, 2009, pp. 1–694.

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. 3rd ed. Boca Raton, FL: Chapman and Hall, 2013, p. 675.

[3] D. Lazarus. Neptune: A marine micropaleontology database. *Mathematical Geology* 26 (7) (1994), 817–832.