How predictable is extinction? Forecasting species survival at million-year timescales

Smits, Peter
psmits@berkeley.edu

Finnegan, Seth
sethf@berkeley.edu

March 8, 2019

Abstract

One of the central tenants of conservation paleobiology is that by studying the past we can better predict the future. It has been suggested that risk assessments for extant species could potentially be improved by examining past extinction patterns and by using paleontological records to establish geographic range and abundance trajectories on geological timescales. Such suggestions prompt two key questions: (1) At a given timescale, to what extent are geographic range and extinction risk trajectories deterministic (past trends are likely to continue into the future) versus Markovian (the future depends only on the present state)? (2) At a given point in time, how accurate do extinction risk predictions based on instantaneous geographic ranges and knowledge of past extinction/survival patterns turn out to be? To address these questions we analyze the exceptionally well-sampled fossil record of Cenozoic planktonic microfossil taxa (foramanifera, radiolarians, diatoms, and calcareous nanoplankton). We examine how survival probability changes over time as a function of species age, time of observation, current geographic range, most recent change in geographic range, climate state, and the change in climate state. Our best performing model has an approximately 78% median probability of correctly ranking the relative extinction risk any two randomly selected species. Although the best-performing model includes time-varying effects and historical covariates (change in range and change in climate state) the improvement in predictive power over models with constant effects and no historical covariates is very minor. Our results imply that at million-year timescales geographic range trajectories are nearly Markovian, perhaps because the processes driving geographic range changes vary on substantially shorter timescales. However, models based on present geographic ranges and past range-risk relationships are likely to be modernly successful at predicting future extinction/survival outcomes. An important caveat is that human impacts may substantially disrupt range-risk dynamics so that the future will be less predictable than it has been in the past.

1 Introduction

Being able to predict which species are more likely than others to go extinct over a given time frame is crucial for planning decisions policies to ameliorate the current biodiversity crisis. Some species have been the focus of intense study to define short-term (e.g. decadal to centennial scale) abundance and geographic range trajectories.

We cannot know, however, we do not yet know which species are going to go extinct because this has not happened yet – it is unobservable. We approach this this problem by analyzing the past in order to predict the future. The fossil record preserves past extinction events, allowing us to develop a predictive model of species extinction based on this record and the properties of the observed species, both extinct and extant [1, 2]. By assessing the predictive performance of this model on unobserved data, we can quantify how precise our best estimates will be for future extinctions – we ask the probability that, given two random species, we correctly rank their relative risks of extinction.

Studying how species vary in their extinction risk over time means we can assess which species are at greater risk under unobserved conditions. We know that a species' risk of extinction varies over time in both intensity (average rate) and selectivity (difference in risk between taxa) [3–5]. Species, after all, can go extinct at any "moment" and the relative risk of extinction exhibited by different taxonomic groups and how that risk varies over time is an important dynamic which shapes the rate and structure of extinction. What has not been evaluated is that as extinction intensity and selectivity change over time, how accurate are our assessments based on past events likely to be when applied to the future? By specifically including and modeling the temporal variation in extinction risk, we are able to improve our overall predictions because we incorporate and explicitly model differences between observations from across a range extinction intensities and selectivities.

By analyzing extinction and survival data from the fossil record, the hope is this can aide in predicting the extinction risk of extant species – after all, the present must at some level be a function of the past. Past paleobiological studies of extinction have frequently focused on identifying and measuring the effect of various predictors on extinction risk [3, 5–10] or on how to identify or measure these effects [11–16]. This focus means that while we have a good understanding of which factors are strong and general determinates of extinction risk, we have less knowledge of how accurate or strong our predictions about the differences in extinction risk are. For example, while a predictor may be "significant" when comparing the odds of extinction risk between groups, the practical difference that predictor makes on prediction can be minimal [17]. By including the kinds of biological and abiotic predictors that have been shown to affect differences in extinction risk, we can quantify their actual, as opposed to relative, effects on predictive performance.

A related question is if the changes to biotic or abiotic predictors, and not just their values, are similarly important factors for predicting extinction. For example, we know that a species' global geographic range changes over its duration [18–21]. We also know that a species' geographic range size is a good predictor of differences extinction risk [3, 22–24]. This begs the question: how does a species' extinction risk change over its duration? While the phenomenon

of species' geographic range change over time has been studied [18–21], the potential predictive impact of this change has been under-evaluated (but see Kiessling and Kocsis [21]). Here, we explicitly estimate the effects of changing geographic range as well as differences in global climate on how well we can predict species extinction. Importantly, the inclusion of these "historical" predictors allows us to more fully evaluate the question of how much information about a species' past is necessary or useful when predicting a species' risk of extinction.

For this exercise we chose to analyze one of the best-sampled and studied fossil records – the Cenozoic record of skeletonized marine planktonic microorganisms F(oraminifera, Radiolaria, Diatoms, and Coccolithophores). These data are readily available through the Neptune database, an online repository of species occurrences obtained through the Deep Sea Drilling Program and the Ocean Drilling Project [25, 26]. This database provides abundant samples in space and time, a high degree of temporal resolution for the entirety of the Cenozoic, and has an internally consistent taxonomic identification strategy – as close to ideal data for this analysis as possible. The Neptune database records multiple phyla-scale taxonomic groups for over 60 million years, with incredible temporal resolution supported by the various age-models of the deep-sea cores from which the occurrences are recorded. Analyzing patterns of extinction and global occurrence at fine temporal scales means we can better elucidate how well we can predict species extinction at human-relevant scales.

2 Materials and Methods

2.1 Data Specifications

We analyzed microfossil occurrence information which was downloaded from the Neptune Database http://www.nsb-mfn-berlin.de/nannotax [25, 26]. All occurrence information was downloaded for calcareous nannofossils, diatoms, foraminifera, and radiolarians.This dataset of occurrences was then filtered to just those species which have their first occurrence at most 63 Mya. This choice means that our analysis avoids those taxa which survived the K/Pg boundary, those taxa which arose just after the K/Pg boundary, and means that our occurrence histories line up with the temperature time-series which was used as a predictor of extinction as discussed below.

All fossil occurrences were assigned to 1 My bins based on the estimated age of the fossil occurrence. After binning, each taxon's geographic range was calculated for each of the 1 My bins in which it occurred. Geographic range was calculated as the maximum great circle distance on an ellipsoid (i.e. the Earth) between any two occurrences of that species; this distance was measured in kilometers.

Temperature data used as covariates in this analysis are based on Magnesium/Calcium isotope ratios sourced from Cramer et al. [27]. These elemental ratios are considered more accurate estimates of past global temperature when compared to the frequently used Oxygen isotope-based estimates; this is because Mg/Ca based estimates are not effected by ice-volume and fresh-water input (e.g. meteoric water) which can alter Oxygen isotope ratios without

reflecting changes to the climate itself. This property is of particular importance for this analysis as polar ice-caps develop midway through the Cenozoic. Our data source, Cramer et al. [27], provides temperature estimates for every 0.1 My from 0 to 63 Mya. We binned these estimates into 1 My intervals as we did with the fossil occurrences. The temperature estimate for each 1 My interval was calculated as the mean of all estimates within that interval.

See Section S1.1 for a further explanation on how observations were temporally binned and how our covariates were standardized and transformed prior to analysis.

2.2   Model Specifications

We developed a discrete-time survival model to analyze our data in order to answer our question of how wee can we predict extinction risk at million year time scales. We considered four model variations: covariate effects are constant over time and none of our historical covariates are included (Model C), covariate effects are allowed to vary over time but we include none of our historical covariates (Model V), covariate effects are constant over time and our historical covariates are included (Model CP), covariate effects are allowed to vary over time and we include our historical covariates (Model VP). See Table 1 for further explanation of how these models differ from each other. For a complete description of the statistical model used in this analysis, please see Section S1.2. For a description of how our models were implemented, please see Section S1.3.

Importantly, we include species geologic age at time of observation as a non-nested varying intercept term. This factor may or may not contribute to differences in species extinction risk over time [7, 28–32], but its inclusion in our model is critical to its nature as a survival model [33]. Again, see Section S1.2 for how this term was specified as part of our model.

2.3   Model adequacy

We are interested in model adequacy and performance into two contexts: in-sample and out-of-sample predictive performance. "In-sample" means we are estimating how well our model predicts our observed data given that the model was fit to the entire dataset; this is a posterior predictive check in that we are comparing the posterior predictive distribution to our observed data. "Out-of-sample" is defined below.

Relative and absolute model adequacy of the four variant models was compared using the area under the receiver operating characteristic curve or AUC [34, 35]. This measure is commonly used in classification problems as it has the desirable characteristic of comparing the model's true positive rate with its false positive rate, as opposed to accuracy which only considers the count of true positives. AUC ranges between 0.5 and 1, with 0.5 indicating no improvement in performance from random and 1 indicating perfect performance. AUC can be interpreted as the probability that our model correctly ranks the relative extinction risks of any two randomly selected species [34, 35]. AUC values of approximately 0.8 or greater

Table 1: Models and their definitions

Code	Description	Covariates	R Formula Syntax[a]
C	Constant effects, no historical cov.	Geographic range, temperature	event[b] ~ range[c] + temp[d] + (1 | age[e]/phylum[f])
V	Varying effects, no historical cov.	Geographic range, temperature	event ~ range + temp + (1 + range + temp | phylum) + (1 | age/phylum)
CP	Constant effects, historical cov.	Geographic range, change in geographic range, temperature, previous temperature	event ~ range + range_diff[g] + temp + temp_lag[h] + (1 | age/phylum)
VP	Varying effects, historical cov.	Geographic range, change in geographic range, temperature, previous temperature	event ~ range + range_diff + temp + temp_lag + (1 + range + range_diff + temp + temp_lag | phylum) + (1 | age/phylum)

[a] See Equation S2 for full statistical model definition.
[b] Species observation where 1 if time of last observation, otherwise 0.
[c] Species geographic range in log km^2. Mean centered, scaled to sd = 1.
[d] Global temperature in degrees C. Mean centered, scaled to sd = 1.
[e] Species are at observation in millions of years.
[f] Taxonomic group of species (i.e. Foraminifera, Diatoms, Radiolarians, Calcaeous nannoplankton).
[g] Change in geographic range since last observation.
[h] Temperature at previous observation.

144 can be considered "good" [36], which means that values between 0.7 and 0.8 could then be considered "fair."

We are particularly interested in understanding how well our model predicts species extinction
147 given new, future data (out-of-sample data). To do this, we estimated average out-of-sample predictive error using 5-fold time-series cross-validation. For time-series data, the folds (data partitions) are approximately equal segments of time. The model is fit to the first fold and
150 the posterior estimates are used to predict the states of the observations in the second fold, then the model is fit to the first and second fold and the posterior states are used to estimate the states from the third fold, and so on with increasingly large numbers of folds used for
153 fitting a model to predict the states from the subsequent fold. With 63 time points, each of the five folds represents approximately 13 time points. Keep in mind, however, that each time point corresponds to many (100-1000) individual observations.

156 See our code repository https://github.com/psmits/trident for full code details. Our code uses "tidyverse" tools such as `dplyr` [37], `purr` [38], and `tidybayes` [39], thus some familiarity with that package ecosystem is necessary to fully comprehend how we've processed our data

and results.

3 Results

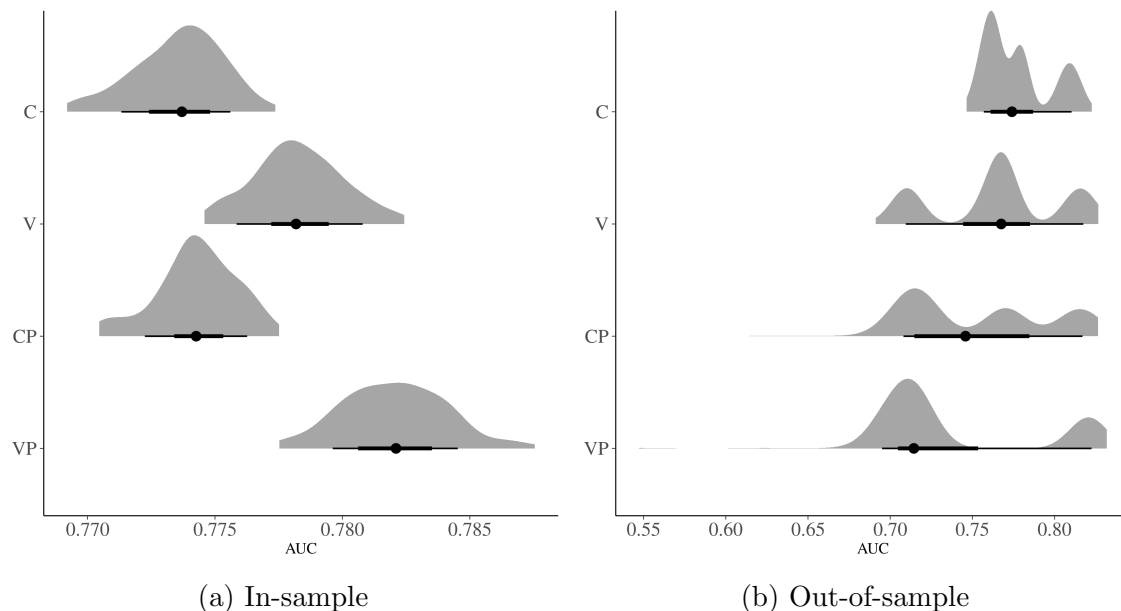(a) In-sample   (b) Out-of-sample

Figure 1: In-sample (1a) and out-of-sample (1b) AUC estimates for each of our four models. These estimates are calculated from the models posterior predictive distribution or from predictions made to new data, respectively. Models with a higher AUC values indicate better performance over models with lower AUC values. AUC is bounded between 0.5 and 1. See Table 1 for a description of each of the four models.

3.1 Model adequacy

The in-sample model comparisons are for determining their relative adequacy, or a model's ability to represent the data it was fit to. Comparison between the posterior predictive estimates of in-sample AUC for each of the four models demonstrates that, overall, all of the models have approximately equal in-sample performance (Fig. 1a). The parameter rich model VP has the greatest median in-sample AUC when compared to the other three models, but there is substantial overlap in their posterior distributions. Additionally, while our parameter rich model VP is possibly the most adequately performing model, the difference or improvement to performance is minimal at best – all four models have approximately equal in-sample AUC posterior distributions. All of the in-sample AUC estimates from our models are concentrated around an AUC of 0.77. It is therefore hard to conclude that there is one "best" model which we can rely upon as they are all nearly functional equivalent.

For many taxon/model combinations there are one or more time periods where posterior predictive in-sample AUC has a median value less than or equal to 0.5 – AUC value of 0.5

6

indicates that the model's predictions are no better than random (Fig. 2). However, this pattern is absent for the posterior predictive distribution of Foraminfera and Radiolaria for the VP model. Additionally, these periods of low model performance are rarer for the posterior predictive distribution of the VP model for calcareous nannoplankton and Dinoflagellates when compared to the other three models.
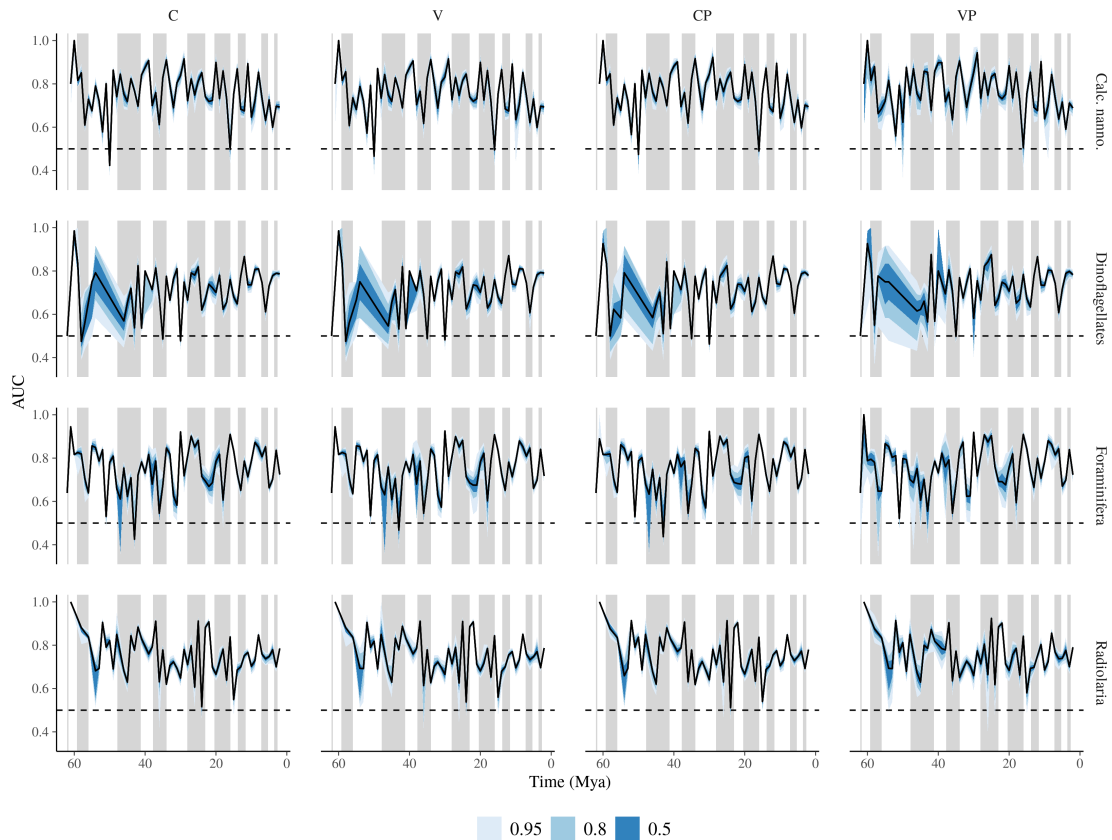

Figure 2: Comparison of posterior predictive AUC estimates for each of the four models, arranged over time and by taxonomic group. These estimates reflect each model's fit to the various taxonomic groups over time. The black line corresponds to the median AUC value, while the envelopes correspond to multiple credible intervals as indicated in the legend. In all cases, higher AUC values indicate greater predictive performance versus lower AUC values. See Table 1 for a description of each of the four models.

3.2 Cross-validation

Expected out-of-sample predictive performance was estimated using five-fold cross-validation [40]. This procedure yields four posterior (predictive) distributions, each corresponding to AUC values calculated from model-based predictions compared to the extinction state of the hold-out data. The resulting distribution when all folds are combined is very multimodal by their very nature of being fit to and estimated from different data sets and amounts of data

7

[40]. We find the visual degree of this multimodality increases with model complexity (Fig. 1b), most likely because the complex models allow for predictor effects to vary with time which allows for a greater range in possible parameter values which in turn yield a greater range of posterior predictions.

Comparison between the posterior predictive distributions of expected out-of-sample AUC (Fig. 1b) reveals a similar range in plausible values for all models as the in-sample AUC posterior predictive distributions (Fig. 1a). Interestingly, the differences between the posterior predictive distributions for the models have decreased. For example, model VP not clearly better than either models V or CP (Fig. 1b), which were shown earlier to be obviously worse-performing models based on in-sample performance (Fig. 1a). These differences means that the rank order of median out-of-sample AUC is different from the rank order of median in-sample AUC. However, the shapes of the posterior distributions means interepting from the median values is incorrect – the models are effectively indistinguishable in their expected out-of-sample AUC values.

We can also compare the posterior predictive distribution of expected out-of-sample AUC over time and taxonomic group for each of the four models (Fig. 3). For each taxonomic group, the time-series of posterior predictive values for each model are broadly congruent.

In the analysis of the posterior predictive distributions of the in-sample AUC values for the four models, we noted that there were time intervals where the models' predictions were no better than random (Fig. 2). This occurrence is generally much rarer for the posterior predictive distribution of out-of-sample AUC values – the major exception to this is Dinoflagellates, which for all four models has at least one time interval where the median the AUC of out-of-sample data were no better random. In contrast, the only other group for which median posterior predictive estimate of out-of-sample AUC reaches 0.5 is calcareous nannoplankton, and then only with model V.

4 Discussion

The results of this paper set out our baseline ability to predict relative differences in extinction risk. We find that all of our model have an approximate 77% to 79% probability of correctly rank the extinction risk of two randomly selected in-sample observations. Similarly, these models are expected to correctly rank the extinction risk of two randomly selected out-of-sample observations approximately 70% to 80% of the time. A slight decrease in performance when dealing with out-of-sample observations makes sense: each of the models fit during cross-validation is based on less data than the model fit on the full data (between 1/5th to 4/5ths of the original). The similarly between the in-sample and out-of-sample results indicates that our model is fairly robust to how extinction intensity has changed over the Cenozoic.

One of the most striking results of this analysis is that the in-sample and out-of-sample comparisons between our models demonstrate that while models where the historical coarviates
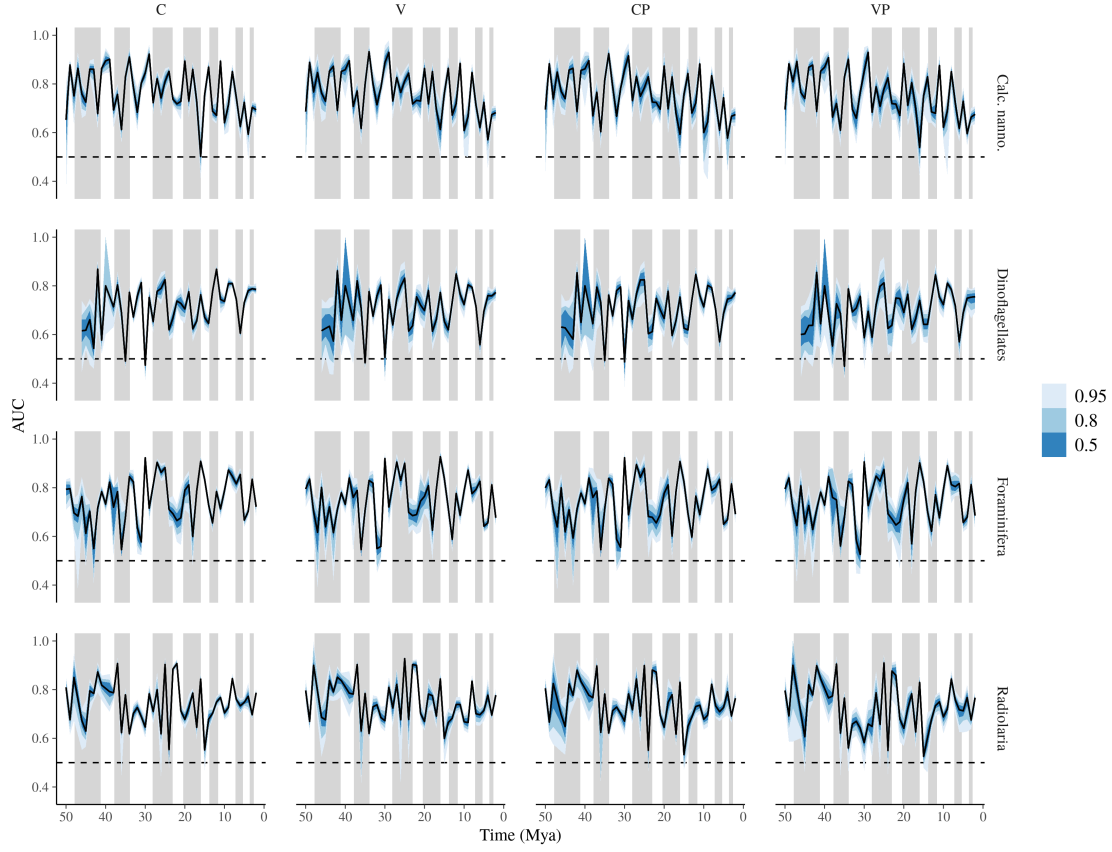
Figure 3: Comparison of out-of-sample AUC values over time as aggregated by taxonomic group for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogentity in performance within and between folds. This presentation decomposes each of the 12-million year folds by each of the taxonomic groups into the predictions made for each of the million-year intervals. The black line corresponds to the median AUC estimate, with the envelopes corresponding to multiple credible intervals as indicated in the legend. See Table 1 for a description of each of the four models.

are predictors of species extinction have better in-sample performance (Fig. 1a, 2), all of them generalize to out-of-sample data with a similar degree of success (Fig. 1b, 3).

Other noteworthy aspects of our in-sample results are that AUC estimates for our models models differ significantly between models and that all of these estimates are in a narrow range of possible AUC values (Fig. 1a). For our in-sample results, while the statistically best model does include the historical covariates and allows all covariate effects to vary over time, its practical difference in performance compared to the other models is virtually negligible (Fig. 1b). This means that even though model VP has a *statistically* greater AUC than the other three models, this result is not practically or *scientifically* significant.

The success of our models is partially driven by the size of our dataset and the hierarchical structure of our model. Our estimates are based on rather limited information about the taxa themselves, and our model only takes into account some aspects of species geographic

9

range and their rough taxonomic grouping. Instead of relying on large amounts of ecological information to shape individual differences, our model leverages most of the Neptune dataset through a multilevel model to constrain and improve our parameter estimates by sharing information about those parameters across taxa and time.

The principal reason we were not able to include more biological information in our models is that we lack most any life history or ecological information on most marine micro- and nano-plankton. Forams are the exception to this problem – there is life history, ecological, and physiological information for a selection of foram species [5]. However, not this information does not exist all foram species. If we want to include this type of information in a predictive model of extinction risk, we would be able to analyze only a single taxonomic subset of the fossil occurrences present the Neptune Database and then only a limited selection of those species. This means ignoring the majority of occurrence information present in the Neptune database.

In conclusion, our results provide a promising picture of our ability to predict the relative extinction risk of two randomly selected species. Considering that conservation decisions are made based on a continuum of risk, from most to least, this means that our results are in the same language as how conservation resources are allocated.

References

[1] P. G. Harnik et al. "Extinctions in ancient and modern seas". In: *Trends in Ecology and Evolution* 27.11 (2012), pp. 608–617.

[2] S. Finnegan et al. "Paleontological baselines for evaluating extinction risk in the modern oceans". In: *Science* 348.6234 (2015), pp. 567–570.

[3] J. L. Payne and S. Finnegan. "The effect of geographic range on extinction risk during background and mass extinction." In: *Proceedings of the National Academy of Sciences* 104.25 (2007), pp. 10506–10511.

[4] J. L. Payne et al. "Extinction intensity, selectivity and their combined macroevolutionary influence in the fossil record". In: *Biology Letters* 12.10 (2016), p. 20160202.

[5] T. H. G. Ezard et al. "Interplay Between Changing Climate and Species' Ecology Drives Macroevolutionary Dynamics". In: *Science* 332.6027 (2011), pp. 349–351.

[6] P. G. Harnik. "Direct and indirect effects of biological factors on extinction risk in fossil bivalves". In: *Proceedings of the National Academy of Science* 108.33 (2011), pp. 13594–13599.

[7] P. D. Smits. "Expected time-invariant effects of biological traits on mammal species duration". In: *Proceedings of the National Academy of Sciences* 112.42 (2015), pp. 13015–13020.

[8] S. E. Peters. "Environmental determinants of extinction selectivity in the fossil record". In: *Nature* 454.7204 (2008), pp. 626–629.

[9] P. G. Harnik, C. Simpson, and J. L. Payne. "Long-term differences in extinction risk among the seven forms of rarity". In: *Proceedings of the Royal Society B: Biological Sciences* 279.1749 (2012), pp. 4969–4976.

[10] M. Foote. "Substrate affinity and diversity dynamics of Paleozoic marine animals". In: *Paleobiology* 32.3 (2006), pp. 345–366.

[11] J. Alroy. "The Shifting Balance of Diversity Among Major Marine Animal Groups". In: *Science* 329.5996 (2010), pp. 1191–1194. arXiv: `arXiv:1011.1669v3`.

[12] J. Alroy. "Accurate and precise estimates of origination and extinction rates". In: *Paleobiology* 40.03 (2014), pp. 374–397.

[13] J. Alroy et al. "Effects of sampling standardization on estimates of Phanerozoic marine diversification". In: *Proceedings of the National Academy of Sciences* 98.11 (2001), pp. 6261–6266.

[14] J. Alroy, P. L. Koch, and J. C. Zachos. "Global Climate Change and North American Mammalian Evolution". In: *Paleotological Society* 26.4 (2000), pp. 259–288.

[15] J. Alroy. "New methods for quantifying macroevolutionary patterns and processes". In: *Paleobiology* 26.4 (2000), pp. 707–733.

[16] M. Foote. "Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis". In: *Paleobiology* 27.4 (2001), pp. 602–630.

[17] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press, 2006.

[18] M. Foote. "Symmetric waxing and waning of marine invertebrate genera". In: *Palaeobiology* 33.4 (2007), pp. 517–529.

[19] L. H. Liow et al. "Global occurrence trajectories of microfossils: environmental volatility and the rise and fall of individual species". In: *Paleobiology* 36.2 (2010), pp. 224–252.

[20] L. H. Liow and N. C. Stenseth. "The rise and fall of species: implications for macroevolutionary and macroecological studies". In: *Proceedings of the Royal Society B: Biological Sciences* 274.1626 (2007), pp. 2745–2752.

[21] W. Kiessling and Á. T. Kocsis. "Adding fossil occupancy trajectories to the assessment of modern extinction risk". In: *Biology Letters* 12.10 (2016), p. 20150813.

[22] D. Jablonski and K. Roy. "Geographical range and speciation in fossil and living molluscs". In: *Proceedings of the Royal Society B: Biological Sciences* 270.1513 (2003), pp. 401–406.

[23] D. Jablonski. "Species Selection: Theory and Data". In: *Annual Review of Ecology, Evolution, and Systematics* 39.1 (2008), pp. 501–524.

[24] D. Jablonski and G. Hunt. "Larval Ecology, Geographic Range, and Species Survivorship in Cretaceous Mollusks: Organismic versus Species-Level Explanations". In: *The American Naturalist* 168.4 (2006), pp. 556–564.

[25] D. Lazarus. "Neptune: A marine micropaleontology database". In: *Mathematical Geology* 26.7 (1994), pp. 817–832.

[26] C. Spencer-Cervato. "The Cenozoic deep sea microfossil record: explorations of the DSDP/ODP sample set using the Neptune database". In: *Palaeontologia Electronica* 2.2 (1999), pp. 4–286.

[27] B. S. Cramer et al. "Late Cretaceous-Neogene trends in deep ocean temperature and continental ice volume: Reconciling records of benthic foraminiferal geochemistry (δ18O and Mg/Ca) with sea level history". In: *Journal of Geophysical Research: Oceans* 116.12 (2011), pp. 1–23.

[28] S. Finnegan, J. L. Payne, and S. C. Wang. "The Red Queen revisited: reevaluating the age selectivity of Phanerozoic marine genus extinctions". In: *Paleobiology* 34.3 (2008), pp. 318–341.

[29] T. H. G. Ezard et al. "The meaning of birth and death (in macroevolutionary birth-death models)". In: *Biology Letters* 8.1 (2012), pp. 139–142.

[30] L. Van Valen. "A new evolutionary law". In: *Evolutionary Theory* 1 (1973), pp. 1–30.

[31] L. H. Liow et al. "Pioneering paradigms and magnificent manifestos–Leigh Van Valen's priceless contributions to evolutionary biology." In: *Evolution; international journal of organic evolution* 65.4 (2011), pp. 917–922.

[32] J. S. Crampton et al. "Greenhouse–icehouse transition in the Late Ordovician marks a step change in extinction regime in the marine plankton". In: *Proceedings of the National Academy of Sciences* 113.6 (2016), pp. 1498–1503.

[33] G. Tutz and M. Schmid. *Modeling discrete time-to-event data.* Springer International Publishing, 2016.

[34] T. Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. arXiv: `/dx.doi.org/10.1016/j.patrec.200` [`http:`].

[35] S. J. Mason and N. Graham. "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical signicance and interpretation". In: *Quarterly Journal of the Royal Metorological Society* 128 (2002), pp. 2145–2166.

[36] W. Tang, H. He, and X. M. Tu. *Applied categorical and count data analysis.* Boca Raton, FL: CRC Press, 2012.

[37] H. Wickham et al. *dplyr: A Grammar of Data Manipulation.* R package version 0.7.8. 2018.

[38] L. Henry and H. Wickham. *purrr: Functional Programming Tools.* R package version 0.2.5. 2018.

[39] M. Kay. *tidybayes: Tidy Data and Geoms for Bayesian Models.* R package version 1.0.3. 2018.

[40] T. Hastie, R. Tibshirani, and J. Friedman. "The Elements of Statistical Learning". In: *Bayesian Forecasting and Dynamic Models* 1 (2009), pp. 1–694. arXiv: `arXiv:1011.1669v3`.

S1 Supplementary Material

S1.1 Data Specifications

S1.1.1 Binning fossil occurrences

Because the estimated ages of each occurrence is a product of core-specific age-models and can be overly precise, the hope is that by binning the data this smooths over the between-core heterogeneity and thus homogenizes our disparate data sources. The occurrence histories of each species were then given binary codes used to model the presence or extinction of those species. For every occurrence of a species, except the last, that species is considered to have survived and was marked with a 0. The last occurrence of that species is considered the

bin in which the taxon has gone extinct – and is assigned a 1. This protocol means that we are reading the fossil record "as written," a practice that is potentially dangerous as it is a overconfident statement of preservation and may be shortening the actual duration of that species [1–10]. However, this practice is common with marine microfossil data due to their exceptional preservation rate [11–14]. In fact, with marine microfossils collected from cores a bigger problem may be over extending the duration of a species due to mixing and smearing within the cores CITATIONS.

S1.1.2 Covariate transforms and standardizations

Prior to analysis, geographic range was then log-plus-one transformed and standardized by mean-centering the data and then dividing by the standard deviation of the distribution of geographic ranges. This standardization means that a regression coefficient associated with this covariate describes the change in extinction probability per change in standard deviation of geographic range, that coefficients associated with similarly standardized covariates will be directly comparable in magnitude, and that the intercept term corresponds to the expected value of the outcome at when geographic range is its average value [15]. Change in geographic range between observations was measured from the standardized geographic range values and was not standardized separately.

Temperature was also transformed and standardized the in the same manner as geographic range. Difference in temperature between between observations was measured from the standardized temperature values and was not standardized separately.

S1.2 Model Specifications

In survival analysis, the hazard function describes the instantaneous rate of extinction of a species given its age and relevant covariates. The hazard function is defined as the conditional probability of a species going extinct by the end of the t-th interval given that it survived up until t and the relevant covariate information X for all k 1 My intervals [16]. For the discrete time intervals $T = 1, \cdots, k$, extinction is defined as $T = t$. The discrete time hazard function is defined as

$$\lambda(t|X) = P(T = t|T \geq t, X), \quad t = 1, \cdots, k. \tag{S1}$$

The hazard function (Eq. S1) is easily reparameterized as a logistic regression by defining that $\lambda(t|X) = h(\Theta)$ where $h(.)$ is a logit inverse-link function and Θ is the probability of a taxon going extinction during interval t [16]. $h(\Theta)$ is then modeled as with any regression as it is defined for all real-values. In this case, we opted for a hierarchical/mixed-effects model with multiple non-nested varying intercepts and slopes [15].

Our covariates matrix X is a $N \times D$ matrix where N is the total number of observations and D is the total number of covariates. The first column of X is entirely 1's as it corresponds to the intercept term in the regression model. The next two columns of X are two aspects of geographic range as continuous covariates: geographic range r during interval t, and the

difference d between the geographic range at $t-1$ and t. The difference in geographic range was calculated from the transformed and standardized geographic range values; this means that change in geographic range is in units of changes in standard deviations. The final two columns are two aspects of global temperature: mean temperature during interval t, and the lag of mean temperature (i.e. mean temperature during interval $t-1$.) As with change to geographic range, the lag of temperature is based on the transformed and standardized temperature estimates.

The matrix of time and phylum varying regression coefficients describing the effects of the covariates on a species' risk of extinction is called B – a w by p matrix, were w is the number of time temporal intervals and p is the number of phyla. The elements of this matrix, the regression coefficients, are themselves modeled as being multivariate normally distributed with vector of means α describing the average intercept and regression coefficient estimates of each coefficients for each phylum p. These phylum averages are themselves modeled as multivariate normally distributed with mean vector μ describing the overall average regression coefficients, including the intercept. μ has length D and is ordered intercept, range coefficient, change in range coefficient, temperature coefficient, temperature lag coefficient.

The effect of species age on the log-odds of species extinction is modeled as a non-nested random intercept A [16]. This term describes how the log-odds of extinction varies along a species duration, and how this effect can differ between the phyla. A is a l by p matrix, where l is the age at observation of a species and p is its phylum. A is modeled as following a multivariate normal distribution with phylum means being vector δ and covariance matrix Σ_A. The covariation between the elements of vector δ are modeled as a multivariate normal distribution with a mean vector of all 0s and covariance matrix Σ_δ.

To complete the generative model, we need to assign final priors to the "top level" parameters. In general we favored weakly informative priors which help regularize our estimates. In the case of a regression coefficient, this means a Normal distribution with mean 0 and a standard deviation of 3. For our scale parameters (e.g. standard deviations), we used half-Cauchy distributed priors with heavy tails but the majority of probability density near 0.

Our top-level intercept was given a more diffuse prior than our regression coefficients, which reflects our greater degree of uncertainty about its value. Our top-level regression coefficient for the effect of geographic range was given an informative prior reflecting the overwhelming amount of evidence that species with a larger than average geographic range have a lower risk of extinction than species with an average or less than average geographic range. In the context of this analysis, this means that we are again using a weakly informative prior but instead of centering the density around -1 (i.e. larger than average geographic range decreases extinction risk).

Instead of assigning a prior distribution for each of the covariance matrices in the model, we instead decomposed the covariance matrices (e.g. Σ_B) which allows us to assign independent priors for the scale and correlation aspects of covariance. The scale parameters were assigned half-Cauchy priors as described above in the context of all other scale parameters. The correlation matrices were assigned LKJ priors each with shape parameter set to 1. This choice of shape parameter produces a uniform distribution over possible correlation matrices.

These priors are also slightly more interpretable than other common prior distributions for covariance matrices such as the inverse-Wishart distribution. This approach to assigning priors to a covariance matrix is recommended by the Stan Manual [17].

In total, our model can be expressed as:

$$
\begin{aligned}
t_i &\sim \text{Bernoulli}(\Theta) \\
\Theta_i &= \text{logit}^{-1}(X_i B_{w[i],p[i]} + A_{l[i],p[i]}) \\
B_{w,p} &\sim MVN(\alpha_p, \Sigma_B) \\
\alpha_p &\sim MVN(\mu, \Sigma_\alpha) \\
A_{l,p} &\sim MVN(\delta_p, \Sigma_A) \\
\delta_p &\sim \text{N}(0, \sigma_\delta) \\
\mu_d &\sim \begin{cases} N(-2, 5) & \text{if } d = \text{intercept} \\ N(-1, 1) & \text{if } d = \text{geo. range} \\ N(0, 1) & \text{else} \end{cases} \\
\delta &\sim N(0, 1) \\
\Sigma_B &= diag(\tau_B)\Omega_B diag(\tau_B) \\
\Sigma_\alpha &= diag(\tau_\alpha)\Omega_\alpha diag(\tau_\alpha) \\
\Sigma_A &= diag(\tau_A)\Omega_A diag(\tau_A) \\
\tau_B &\sim C^+(1) \\
\tau_\alpha &\sim C^+(1) \\
\tau_A &\sim C^+(1) \\
\Omega_B &\sim LKJ(1) \\
\Omega_\alpha &\sim LKJ(1) \\
\Omega_A &\sim LKJ(1)
\end{aligned}
\tag{S2}
$$

with i indexing the observation and bracket subscripts referencing the class of the ith observation where $w[i]$ is the time of the i-th observation, $p[i]$ is the phylum of the i-th observation, and $d[i]$ is the age of the i-th observation.

S1.3   Model Parameter Estimation

We implemented our model (Eq. S2 using the `rstanarm` package for the R programming language [17]. This package provides an interface to the Stan probabilistic programming language for writing hierarchical/mixed-effects models in native R. Posterior estimates were obtained through Hamiltonian Monte Carlo, using 2000 steps divided equally between warm-up and sampling. In order to prevent divergent transitions the adapt delta value was increased to 0.9999; all other HMC/NUTS sampling parameters were kept at the defaults for rstanarm 2.18.2 [18].

An implementation of our full model in `rstanarm`, given a data.frame of all necessary data in a data.frame called "data", is coded as:

```
form <- event ~ range + range_diff + temp + temp_lag +
                (1 + range + range_diff + temp + temp_lag | mybin/phylum) +
                (1 | age/phylum),
stan_glmer(formula = form,
           data = data,
           family = 'binomial',
           prior = normal(c(-1, 0, 0, 0), rep(1, 4), autoscale = FALSE),
           prior_intercept = normal(-2, 5, autoscale = FALSE),
           prior_aux = cauchy(0, 1, autoscale = FALSE),
           chains = 4,
           thin = 4,
           adapt_delta = 0.999999)
```

Similarly, our full model can also be implemented using the brms Stan interface [19, 20] as:

```
priors <- c(set_prior('normal(-2, 5)', class= 'Intercept'),
            set_prior('normal(0, 1)', class = 'b'),
            set_prior('normal(-1, 1)', class = 'b', coef = 'range'),
            set_prior('cauchy(0, 1)', class = 'sd'),
            set_prior('lkj(1)', class = 'cor'))
form <- bf(event ~ range + range_diff + temp + temp_lag +
           (1 + range + range_diff + temp + temp_lag | mybin/phylum) +
           (1 | age/phylum))
brmfit <- brm(formula = form,
              data = data,
              family = bernoulli(),
              prior = priors,
              chains = 4,
              thin = 4,
              control = list(adapt_delta = 0.999999)
```

Posterior convergence was determined using the general and HMC-specific diagnostic criteria: scale reduction factor ($\hat{R}$; target < 1.1), effective sample size (eff; target value eff/steps < 0.0001), number of samples that saturated the maximum trajectory length for avoiding infinite loops (treedepth; target value 0), sample divergence, and the energy Bayesian Fraction of Mission Information (E-BFMI; target value > 0.2). For further explanation of these diagnostic criteria, see the Stan Manual [17].

Supplementary References

[1] J. Alroy. "Fair sampling of taxonomic richness and unbiased estimation of origination and extinction rates". In: *Quantitative methods in paleobiology. Paleontological Society Papers.* 16 (2010), pp. 55–80.

[2] J. Alroy. "New methods for quantifying macroevolutionary patterns and processes". In: *Paleobiology* 26.4 (2000), pp. 707–733.

[3]   J. Alroy. "Accurate and precise estimates of origination and extinction rates". In: *Paleobiology* 40.03 (2014), pp. 374–397.

[4]   M. Foote. "Estimating Taxonomic Durations and Preservation Probability". In: *Paleobiology* 23.3 (1997), pp. 278–300.

[5]   M. Foote and J. J. Sepkoski. "Absolute measures of the completeness of the fossil record." In: *Nature* 398.6726 (1999), pp. 415–417.

[6]   M. Foote. "Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis". In: *Paleobiology* 27.4 (2001), pp. 602–630.

[7]   M. Foote and D. M. Raup. "Fossil preservation and the stratigraphic ranges of taxa". In: *Paleobiology* 22.2 (1996), pp. 121–140.

[8]   G. T. Lloyd et al. "Sampling bias and the fossil record of planktonic foraminifera on land and in the deep sea". In: *Paleobiology* 38.4 (2012), pp. 569–584.

[9]   C. R. Marshall. "Distinguishing between sudden and gradual extinctions in the fossil record: predicting the position of the Cretaceous-Tertiary iridium anomaly using the ammonite fossil record on Seymour Island, Antarctica". In: *Geology* 23.8 (1995), pp. 731–734.

[10] S. C. Wang and C. R. Marshall. "Estimating times of extinction in the fossil record". In: *Biology Letters* 12.4 (2016), p. 20150989.

[11] T. H. G. Ezard, G. H. Thomas, and A. Purvis. "Inclusion of a near-complete fossil record reveals speciation-related molecular evolution". In: *Methods in Ecology and Evolution* 4.8 (2013), pp. 745–753.

[12] T. H. G. Ezard and A. Purvis. "Environmental changes define ecological limits to species richness and reveal the mode of macroevolutionary competition". In: *Ecology Letters* (2016), pp. 1–8.

[13] T. H. G. Ezard et al. "Interplay Between Changing Climate and Species' Ecology Drives Macroevolutionary Dynamics". In: *Science* 332.6027 (2011), pp. 349–351.

[14] L. H. Liow et al. "Global occurrence trajectories of microfossils: environmental volatility and the rise and fall of individual species". In: *Paleobiology* 36.2 (2010), pp. 224–252.

[15] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press, 2006.

[16] G. Tutz and M. Schmid. *Modeling discrete time-to-event data*. Springer International Publishing, 2016.

[17] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual*. 2017.

[18] J. Gabry and B. Goodrich. *rstanarm: Bayesian Applied Regression Modeling via Stan*. R package version 2.18.2. 2018.

[19] P.-C. Bürkner. "brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software* 80.1 (2017), pp. 1–28.

[20] P.-C. Bürkner. "Advanced Bayesian Multilevel Modeling with the R Package brms". In: *The R Journal* 10.1 (2018), pp. 395–411.