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

Smits, Peter
psmits@berkeley.edu

Finnegan, Seth
sethf@berkeley.edu

Abstract

A tenet of conservation palaeobiology is that knowledge of past extinction patterns can help us to better predict future extinctions. Although the future is unobservable, we can test the strength of this proposition by asking how well models conditioned on past observations would have predicted subsequent extinction events at different points in the geological past. To answer this question we analyze the well-sampled fossil record of Cenozoic planktonic microfossil taxa (foramanifera, radiolarians, diatoms, and calcareous nanoplankton.) We examine how extinction probability varies over time as a function of species age, time of observation, current geographic range, change in geographic range, climate state, and change in climate state. Although the best-performing model includes time-varying effects and historical covariates (change in range and change in climate) the improvement in predictive power over models with constant effects and no historical covariates is very minor. Our models have a 70-80% probability of correctly forecast the rank order of extinction risk for a random out-of-sample pair, implying that determinants of extinction have varied only modestly through time. 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.

Keywords: conservation, palaeobiology, extinction, forecasting

1 Introduction

The intensifying biodiversity crisis confronts conservation biologists with the difficult task of trying to predict which species are most threatened with extinction in the near future. Predicting which species will go extinct is difficult because reliable population and geographic range time series are typically known for only the past few decades in even the best-studied groups, and because few modern extinctions have been adequately documented. This has led to the suggestions that some risk assessments might be improved by incorporating palaeontological data [1, 2]. The fossil record preserves information about the full histories, include ultimate extinction, of thousands of lineages, and this information can help to augment the shorter-term higher-resolution data used to make risk assessments of extant taxa.

Past palaeobiological studies of extinction have frequently focused on identifying and measuring the effect of various potential predictors on extinction risk [3–9] or on how to identify or measure these effects [10–15]. This focus means that while we have a good understanding of which factors are strong and general determinants of extinction risk, we have less knowledge of how accurate or strong our predictions about the differences in extinction risk are.

Here we ask how precise risk predictions based on fossil data might be. Because future extinctions are unobservable we cannot directly evaluate the ultimate performance of such predictions. However, we can address the question by taking a given point in the geological past, developing a predictive model based on extinction patterns prior to that point, and assessing the predictive performance of this model on unobserved (e.g. "future", from the point of view of the model) extinction/survival events.

Extinction intensity (average rate) and selectivity (difference in risk between taxa) vary through time 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 [6, 8, 16, 17]. This variation also raises a question: given that extinction intensity and selectivity change over time, how accurate are our assessments based on past events likely to be when applied to the future? Putting aside the important question of how human activities will alter the determinants of future extinction risk, we can address this uncertainty by specifically including and modeling the temporal variation in extinction risk across a range of extinction intensities and selectivities.

Numerous studies have established that geographic range is one of the most important determinants of extinction risk in the fossil record, and that a species geographic range can be highly variable over geologic time [6, 18–24]. The degree to which the past can help to predict the future fates of species depends in part on the degree to which species geographic range trajectories are deterministic versus Markovian [20, 21, 25, 26]. In the former case, knowledge of the specific past trajectory of a species – whether its range has expanded or contracted over a given time span – might help to improve assessments of its current risk. In the latter case, only the current range of the species conveys useful information about its current risk, although we can still use prior extinction patterns to augment predictions by evaluating the relative extinction risk of species that had similar ranges in the past. Discriminating among these alternative models is thus very important for determining how best to incorporate fossil data in present risk assessments.

For this exercise we chose to analyze one of the best-sampled and studied fossil records – the Cenozoic record of skeletonized marine planktonic microorganisms (Foraminifera, 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 [27, 28]. 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 from the Neptune Database `http://www.nsb-mfn-berlin.de/nannotax` [27, 28]. This occurrence-based dataset includes calcareous nannoplankton, diatoms, foraminifera, and radiolarians.Occurrences were filtered to include only those species with first occurrence was no more than 63 Mya. This filtering criterion excludes taxa that survived the K/Pg extinction or arose during its ecologically unusual recovery interval, and ensures that our occurrence histories fully overlap with the temperature time-series used as a potential extinction risk predictor (see below).

All fossil occurrences were assigned to 1 My bins based on the estimated age of the fossil occurrence as listed in the Neptune Database. 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.

Average global temperature of each 1 My bin was calculated from estimates based on Magnesium/Calcium isotope ratios Cramer et al. [29]. We use Mg/Ca rather than oxygen isotopes to avoid confounding effect of varying ice-volume – this property is of particular importance for this analysis as polar ice-caps develop midway through the Cenozoic. Our data source, Cramer et al. [29], estimated temperature for every 0.1 My from 0 to 63 Mya. 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 used a discrete-time survival modelling framework to estimate how well we can predict extinction risk at one million year time scales. At its core, our model is a multilevel logistic regression with taxon age in millions of years as a varying intercept [30]. We considered four different models involving different permutations of covariate effects (fixed or time-varying) and historical covariates: covariate effects constant over time and no historical covariates included (Model C), covariate effects allowed to vary over time but no historical covariates included (Model V), covariate effects constant over time and historical covariates included (Model CP), and covariate effects allowed to vary over time and historical covariates are included (Model VP). The C and P models attempt to predict based only on present state,

whereas the CP and VP models allow for the possibility of non-Markovian behaviour by including change in state from the previous time increment.

We always included species age at time of observation (i.e. observed prior duration) as a non-nested varying intercept term. This factor may or may not contribute to differences in species extinction risk over time [4, 31–35], but its inclusion in our model is critical to its nature as a survival model [30].

See Table 1 for further explanation of how the four models we considered differ from each other. A complete description of the statistical model used in this analysis is available in Section S1.2. Additionally, the full description of how these models were implemented and coded, including choice of priors, is available Section S1.3.

Table 1: Models and their definitions

Code	Description	Covariates	R Formula Syntax[a]
C	Constant effects, no historical cov.	Geographic range, temperature	event[b] $\sim$ range[c] + temp[d] + (1 | age[e]/phylum[f])
V	Varying effects, no historical cov.	Geographic range, temperature	event $\sim$ range + temp + (1 + range + temp | phylum) + (1 | age/phylum)
CP	Constant effects, historical cov.	Geographic range, change in geographic range, temperature, previous temperature	event $\sim$ 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 $\sim$ 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, Calcareous nannoplankton).
[g] Change in geographic range since last observation.
[h] Temperature at previous observation.

2.3 In-sample and out-of-sample forecasting

We are interested in our models' performance in two contexts: in-sample performance, and out-of-sample predictive performance (i.e. forecasting).

In-sample forecasting means we are estimating how well our model predicts extinction probability for observations that that model was fit to. This is a posterior predictive check in that we are comparing the posterior predictive distribution to our observed data. In-sample forecasting measures, however, are not necessarily good estimates of the model's ability to predict data from the future [36].

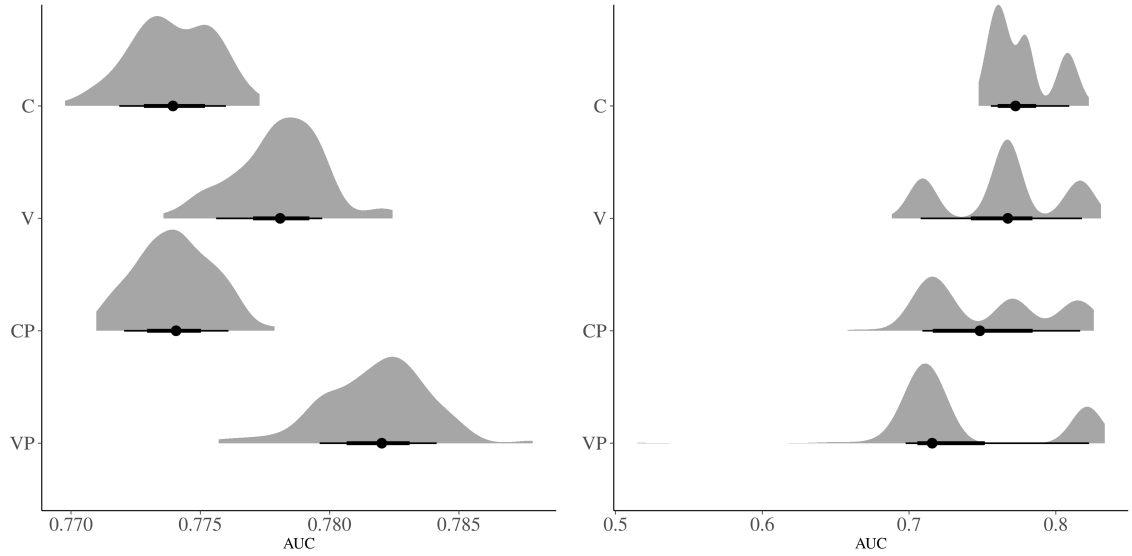
We are particularly interested in understanding how well our model forecasts extinction probability of data from the future that the model was not fit to (out-of-sample data). To quantify our ability to forecast species' extinction risk, we estimated average out-of-sample forecasting performance using 5-fold time-series cross-validation. For time-series data, the folds (data partitions) are approximately equal segments of time. Each fold represents a sequence of time points. With 63 time points, each of the five folds represents approximately 13 million-year time increments. It is important to bear in mind, however, that each time increment includes many (100s-1000s) individual observations.

k-fold cross-validation for time series follows a specific sequence of procedures [37, 38]. First, the model is fit to the first fold (time segment), and the posterior estimates of that fit are then used to forecast the extinction probability of the second fold (i.e. the future). Then the model is fit to the combined first and second folds, and the posterior estimates of that fit are used to forecast the extinction probability of the third fold. Continuing, the model is then fit to the first three folds combined and is then used to forecast extinction probabilities for the fourth fold. Finally, the model is fit to the first four folds combined and then is used to forecast the fifth fold. The results from these forecasts are then combined to yield our estimate of expected out-of-sample performance.

The relative adequacy of the four model variants was compared using the area under the receiver operating characteristic curve or AUC [39, 40]. This measure is commonly used to measure the performance of classification models 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 true positives. AUC ranges between 0.5 and 1, with 0.5 indicating no difference in classification from random and 1 indicating perfect classification. AUC can be interpreted as the probability that our model correctly ranks the relative extinction risks of a randomly selected extinct-extant species pair [39, 40]. AUC values of approximately 0.8 or greater can be considered "good" [41], so we interpret values between 0.7 and 0.8 could then be considered "fair," and values between 0.6 and 0.7 as "poor."

See our code repository at https://github.com/psmits/trident for full code details. This entire analysis was coded in R and uses tidyverse and tidyverse adjacent tools such as `dplyr` [42], `purr` [43], and `tidybayes` [44]. Additionally, all of our models were written using the `brms` [45, 46] R package, which implements Stan-based Bayesian models which are fit via Hamiltonian Monte Carlo [47].

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 In-sample forecasting adequacy

The in-sample model comparisons are useful for comparing the relative ability of our models to represent the data they were fit to. Comparison between the posterior distributions of in-sample AUC for each of the four models demonstrates that all of our models have approximately equal in-sample forecasting 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 the in-sample AUC estimates from our models are concentrated on an AUC value 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.

Depending on the taxon-model combination, there are between zero and 4 time intervals where our posterior distribution of in-sample AUC has a median value less than or equal to 0.5 (Fig. 2). However, this pattern is absent for the posterior estimates of in-sample AUC for Foraminfera and Radiolaria as fit by the VP model. In contrast, there are fewer periods of

low model performance for calcareous nannoplankton and Dinoflagellates as estimated from
174 our VP model than in those estimates from the other three model variations.

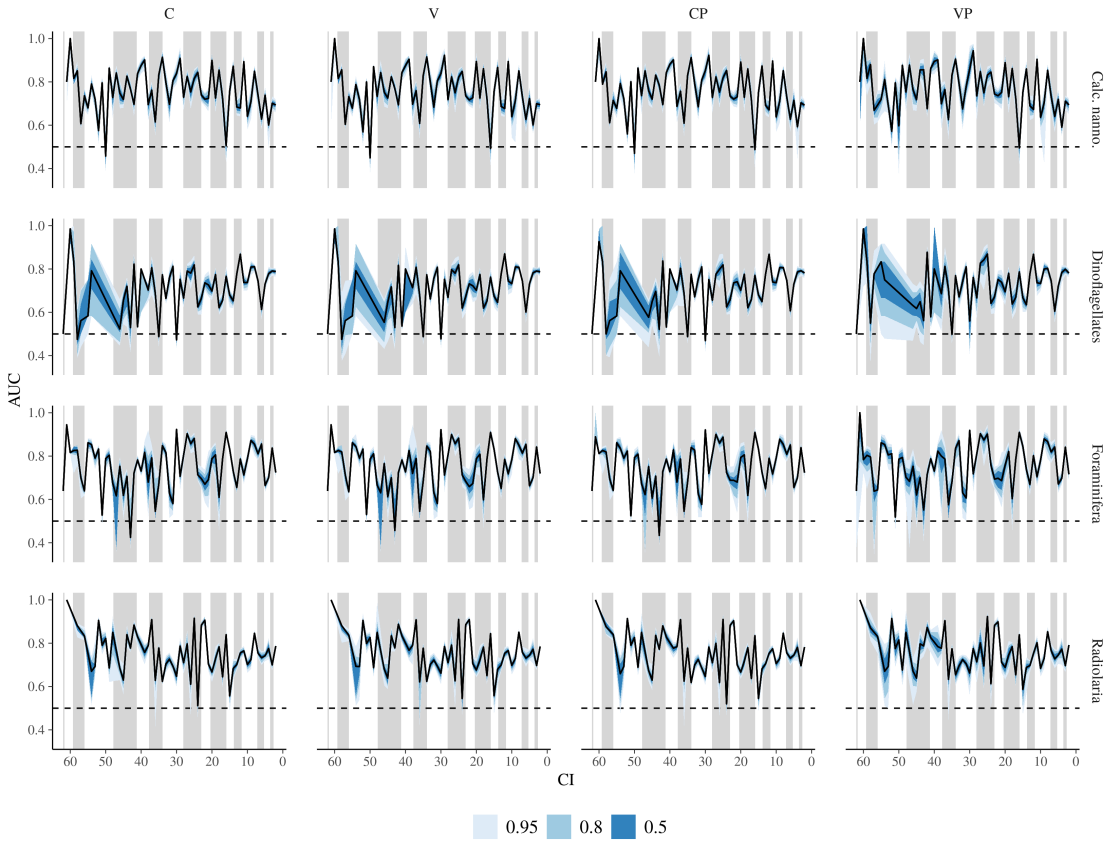


Figure 2: Comparison of in-sample forecasting performance measured by AUC for each of the four models, arranged over time and by taxonomic group. These estimates reflect each model's fit to the various taxonomic groups over time. The black line corresponds to the median AUC value, while the envelopes 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 Out-of-sample forecasting performance

Expected out-of-sample forecasting performance was estimated using five-fold cross-validation
177 for time series [37, 38]. The resulting distribution when all folds are combined is highly multimodal as expected given that they are fit to and estimated from different data sets with different numbers of observations [36]. This multimodality increases with model complexity
180 (Fig. 1b), most likely because the complex models allow for predictor effects to vary with time, allowing a greater range in possible parameter values which in turn yield a greater range of posterior predictions.

183 Comparison the in-sample AUC estimates to the expected out-of-sample AUC estimates

reveals a similar range in performance for all models (Fig. 1a, 1b). Interestingly, the differences in posterior predictive distributions of AUC between the four model variations is reduced in out-of-sample prediction. For example, the VP model no longer has the greatest median AUC of the four models (Fig. 1b). For this reason the rank order of median out-of-sample AUC is different from the rank order of median in-sample AUC. However, making interpretations based only on the median AUC estimates is incorrect – our estimates are the full posterior, not individual points. Because of this, the four models are effectively indistinguishable in their out-of-sample forecast performance (Fig. 1b).

The posterior predictive distribution of expected out-of-sample AUC over time and across taxonomic groups are nearly identical for each of the four models (Fig. 3).

In the analysis of the in-sample forecast performance of the four models, we noted that there were time intervals where our predictions were no better than random (Fig. 2). This occurrence is generally much rarer for the posterior distribution of AUC from the out-of-sample forecasts. The major exception to this pattern are our estimates for the Dinoflagellates, which have at least one time interval for all four models in which the median AUC of the out-of-sample forecasts were no better random. 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 the V model.

We compared the difference in our AUC estimates from the out-of-sample forecasts to the AUC estimates from our in-sample forecasts by subtracting the in-sample AUC estimates from the out-of-sample AUC estimates (Fig. 4). A difference in AUC close to 0 indicates complete congruence between the in-sample and out-of-sample forecasts. A positive difference indicates that our out-of-sample forecasts are actually higher performing than our in-sample forecasts, while negative difference indicates poorer out-of-sample performance than in-sample forecast. Divergences between our out-of-sample and in-sample forecasts are rare and tend to not form multimillion year patters, consistent with the broad visual congruence between the in-sample and out-of-sample forecast performance (Fig. 2, 3). The only major multimillion year pattern indicating significantly poorer out-of-sample forecast performance than in-sample forecast performance is for Radiolaria based on the VP model concentrated around 30 Mya (Fig. 4).

4 Discussion

We find that all of our models are expected to correctly forecasting the rank order of extinction probability for a randomly selected extinct-extant pair of future observations between 70% to 80% of the time (Fig. 1b). One of the most striking aspects of these results is the similarity in forecasting performance for in-sample and out-of-sample observations. 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 fewer data than the model fit on the full data (between 1/5th to 4/5ths of the original). Additionally, a potential decrease in precision when forecasting the future of extinction risk is to be expected as the future will always differ from the past in some respect. However, the similarly between the in-sample and out-of-sample results indicates that our model is fairly robust to variation in extinction intensity over the

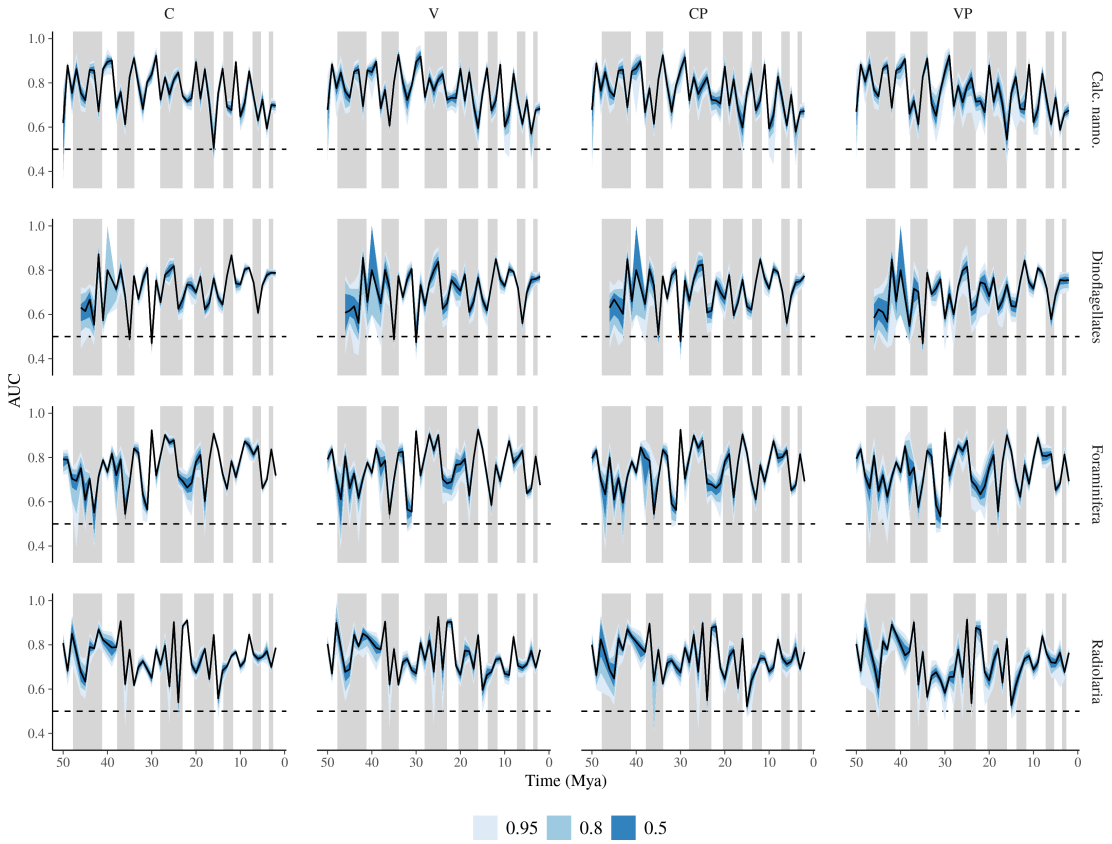


Figure 3: Comparison of out-of-sample AUC values over time as aggregated by taxonomic group for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogeneity in performance within and between folds. This presentation decomposes each of the 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.

Cenozoic. A extremely important caveat, of course, is that human impacts may substantially alter present and future extinction risk dynamics relative to the average Cenozoic state, so that the future may become less predictable than it has been in the past [1, 48].

We also find that our four models are practically identical in their ability to make in-sample and out-of-sample forecasts. Although the in-sample AUC estimates differ between models, all of these estimates are in a narrow range of possible AUC values (Fig. 1a). While the "best" model does include the historical covariates and allows all covariate effects to vary over time, its practical difference in performance versus the other models is negligible (Fig. 1b). Thus even though the VP model has a statistically greater AUC for in-sample forecasts than the other three models, this result is not practically or scientifically significant. The results of the out-of-sample forecasts illustrate this point in full, as all four models have functionally identical out-of-sample performance. Thus, including geographic range trajectory results in only very minor improvements in forecasting accuracy. This contrasts with findings from coral

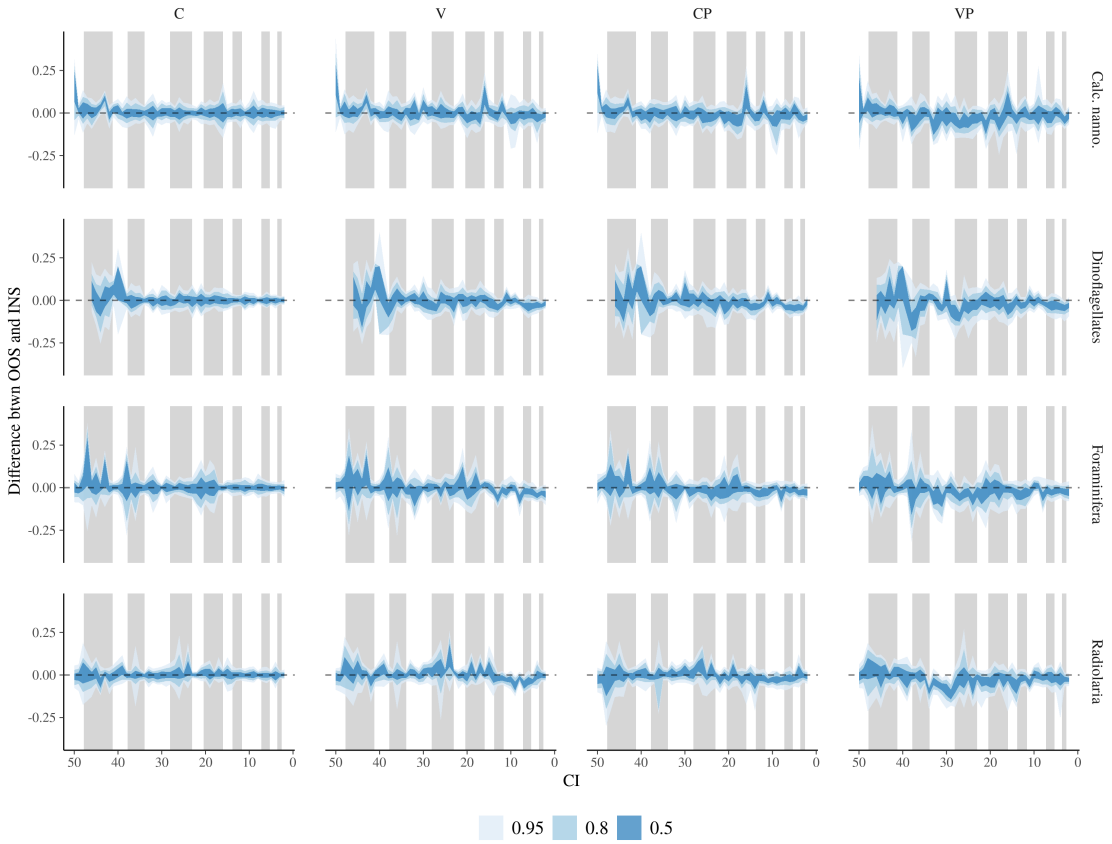


Figure 4: Comparison between our out-of-sample forecasts and in-sample forecasts for all models. This value is calculated as the values presented in Figure 3 minus those values presented in Figure 2. A differences close to 0 indicate complete congruence between in-sample and out-of-sample forecasts, while a positive difference indicates that our out-of-sample forecasts are actually higher performing than our in-sample forecasts, and a negative difference indicates poorer out-of-sample performance than in-sample forecast. See Table 1 for a description of each of the four models.

genera [2] and suggests that further investigation of other taxa, timescales, and environments may help to elucidate the conditions under which past geographic range trajectories are most informative about future extinction risk.

The relative quality and consistency of our models' out-of-sample forecasting performance is encouraging given that these estimates are based on very limited biological and environmental information about the studied taxa. Even our most complex models only account for a few simple aspects of geographic range, prior history, and phylogenetic affinity. The principal reason we were not able to include more biological information in our models is that we lack suitably detailed life history or ecological information for many marine micro- and nannoplankton. Foraminifera are an exception to this problem as aspects of life history, ecology, and physiology are known for many foram species [8]. However, comparable information does not exist all foram species, nor does this type of data exist for the other three taxonomic groups studied here. Future analyses including this type of information and focused more narrowly on the

foraminifera may be informative.

In summary, our results suggest that models trained on prior extinction/survival patterns do modestly well at predicting relative extinction probability of randomly selected species pairs based on a small number of taxonomic, geographic, and historical predictors. Our results are directly comparable to conservation determinations because both are expressed in terms of a continuum of risk, from most to least. The results of this simple exercise suggest that conservation decisions would be bolstered by including fossil data.

References

[1] S. Finnegan et al. Paleontological baselines for evaluating extinction risk in the modern oceans. *Science* 348 (6234) (2015), 567–570.

[2] W. Kiessling and Á. T. Kocsis. Adding fossil occupancy trajectories to the assessment of modern extinction risk. *Biology letters* 12 (2016), 20150813.

[3] P. G. Harnik. Direct and indirect effects of biological factors on extinction risk in fossil bivalves. *Proceedings of the National Academy of Science* 108 (33) (2011), 13594–13599.

[4] P. D. Smits. Expected time-invariant effects of biological traits on mammal species duration. *Proceedings of the National Academy of Sciences* 112 (42) (2015), 13015–13020.

[5] S. E. Peters. Environmental determinants of extinction selectivity in the fossil record. *Nature* 454 (7204) (2008), 626–629.

[6] J. L. Payne and S. Finnegan. The effect of geographic range on extinction risk during background and mass extinction. *Proceedings of the National Academy of Sciences* 104 (25) (2007), 10506–10511.

[7] P. G. Harnik, C. Simpson, and J. L. Payne. Long-term differences in extinction risk among the seven forms of rarity. *Proceedings of the Royal Society B: Biological Sciences* 279 (1749) (2012), 4969–4976.

[8] T. H. G. Ezard, T. Aze, P. N. Pearson, and A. Purvis. Interplay Between Changing Climate and Species' Ecology Drives Macroevolutionary Dynamics. *Science* 332 (6027) (2011), 349–351.

[9] M. Foote. Substrate affinity and diversity dynamics of Paleozoic marine animals. *Paleobiology* 32 (3) (2006), 345–366.

[10] J. Alroy. The Shifting Balance of Diversity Among Major Marine Animal Groups. *Science* 329 (5996) (2010), 1191–1194.

[11] J. Alroy. Accurate and precise estimates of origination and extinction rates. *Paleobiology* 40 (03) (2014), 374–397.

[12] J. Alroy et al. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences* 98 (11) (2001), 6261–6266.

[13] J. Alroy, P. L. Koch, and J. C. Zachos. Global Climate Change and North American Mammalian Evolution. *Paleotological Society* 26 (4) (2000), 259–288.

[14] J. Alroy. New methods for quantifying macroevolutionary patterns and processes. *Paleobiology* 26 (4) (2000), 707–733.

[15] M. Foote. Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis. *Paleobiology* 27 (4) (2001), 602–630.

[16] J. L. Payne, A. M. Bush, E. T. Chang, N. A. Heim, M. L. Knope, and S. B. Pruss. Extinction intensity, selectivity and their combined macroevolutionary influence in the fossil record. *Biology Letters* 12 (10) (2016), 20160202.

[17] P. D. Smits. How macroecology affects macroevolution : the interplay between extinction intensity and trait-dependent extinction in brachiopods. *bioRxiv* (2019), 523811.

[18] M. Foote. Symmetric waxing and waning of marine invertebrate genera. *Palaeobiology* 33 (4) (2007), 517–529.

[19] L. H. Liow, H. J. Skaug, T. Ergon, and T. Schweder. Global occurrence trajectories of microfossils: environmental volatility and the rise and fall of individual species. *Paleobiology* 36 (2) (2010), 224–252.

[20] L. H. Liow and N. C. Stenseth. The rise and fall of species: implications for macroevolutionary and macroecological studies. *Proceedings of the Royal Society B: Biological Sciences* 274 (1626) (2007), 2745–2752.

[21] W. Kiessling and Á. T. Kocsis. Adding fossil occupancy trajectories to the assessment of modern extinction risk. *Biology Letters* 12 (10) (2016), 20150813.

[22] D. Jablonski and K. Roy. Geographical range and speciation in fossil and living molluscs. *Proceedings of the Royal Society B: Biological Sciences* 270 (1513) (2003), 401–406.

[23] D. Jablonski. Species Selection: Theory and Data. *Annual Review of Ecology, Evolution, and Systematics* 39 (1) (2008), 501–524.

[24] D. Jablonski and G. Hunt. Larval Ecology, Geographic Range, and Species Survivorship in Cretaceous Mollusks: Organismic versus Species-Level Explanations. *The American Naturalist* 168 (4) (2006), 556–564.

[25] M. Foote, J. S. Crampton, A. G. Beu, B. A. Marshall, R. A. Cooper, P. A. Maxwell, and I. Matcham. Rise and fall of species occupancy in Cenozoic fossil mollusks. *Science* 318 (November) (2007), 1131–1134.

[26] A. L. Pigot, I. P. Owens, and C. D. L. Orme. Speciation and extinction drive the appearance of directional range size evolution in phylogenies and the fossil record. *PLoS Biology* 10 (2) (2012).

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

[28] C. Spencer-Cervato. The Cenozoic deep sea microfossil record: explorations of the DSDP/ODP sample set using the Neptune database. *Palaeontologia Electronica* 2 (2) (1999), 4–286.

[29] B. S. Cramer, K. G. Miller, P. J. Barrett, and J. D. Wright. 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. *Journal of Geophysical Research: Oceans* 116 (12) (2011), 1–23.

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

[31] S. Finnegan, J. L. Payne, and S. C. Wang. The Red Queen revisited: reevaluating the age selectivity of Phanerozoic marine genus extinctions. *Paleobiology* 34 (3) (2008), 318–341.

[32] T. H. G. Ezard, P. N. Pearson, T. Aze, and A. Purvis. The meaning of birth and death (in macroevolutionary birth-death models). *Biology Letters* 8 (1) (2012), 139–142.

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

[34] L. H. Liow et al. Pioneering paradigms and magnificent manifestos–Leigh Van Valen's priceless contributions to evolutionary biology. *Evolution; international journal of organic evolution* 65 (4) (2011), 917–922.

[35] J. S. Crampton, R. A. Cooper, P. M. Sadler, and M. Foote. Greenhouse–icehouse transition in the Late Ordovician marks a step change in extinction regime in the marine plankton. *Proceedings of the National Academy of Sciences* 113 (6) (2016), 1498–1503.

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

[37] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. 4 (2009), 40–79. arXiv: 0907.4728.

[38] C. Bergmeir, R. J. Hyndman, and B. Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis* 120 (2018), 70–83.

[39] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8) (2006), 861–874.

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

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

[42] H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*. R package version 0.7.8. 2018.

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

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

[45] P.-C. Bürkner. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80 (1) (2017), 1–28.

[46] P.-C. Bürkner. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* 10 (1) (2018), 395–411.

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

[48] P. G. Harnik et al. Extinctions in ancient and modern seas. *Trends in Ecology and Evolution* 27 (11) (2012), 608–617.

S1 Supplement to Materials and Methods

S1.1 Data Specifications

S1.1.1 Binning fossil occurrences

The estimated age of each occurrence is based on the core-specific age-model that observation is from and can be overly precise. To alleviate this overprecision, we coarsened our temporal information in an effort to limit the effects of between-core heterogeneity in age. The occurrence histories of each species was then summarized as a series of binary codes indicating the presence or last occurrence of that species. For every occurrence of a species, except the last, that species existence and survival is recorded as a 0. The last occurrence of that species is considered the bin in which the taxon has gone extinct – and is recorded as 1. This protocol means that we are reading the fossil record "as written," a practice that is potentially dangerous as it is an overconfident statement of preservation and may be shortening the actual durations of the studied 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 [15–18].

S1.1.2 Covariate transformation and standardization

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 each covariate describes the change in extinction probability per change in standard deviation of that covariate, 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 [19]. 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. The change in temperature between an observation and its previous observation 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 covariate information. 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 [20]. 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 [20]. $h(\Theta)$ is then modeled as with any regression. In this case, we opted for a hierarchical/mixed-effects model with multiple non-nested varying intercepts and slopes [19].

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. Change 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 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 [20]. 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 the 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 [21].

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 [21]. 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 adapt delta was increased to 0.999999; all other HMC/NUTS sampling parameters were kept at the defaults for rstanarm 2.18.2 [22].

To implement our VP model in rstanarm, where "data" is a data.frame object of all necessary data (response, covariates), 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 VP model can be implemented using the brms Stan interface [23, 24] 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 [21].

Supplementary References

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

[2] J. Alroy. New methods for quantifying macroevolutionary patterns and processes. *Paleobiology* 26 (4) (2000), 707–733.

[3] J. Alroy. Accurate and precise estimates of origination and extinction rates. *Paleobiology* 40 (03) (2014), 374–397.

[4] M. Foote. Estimating Taxonomic Durations and Preservation Probability. *Paleobiology* 23 (3) (1997), 278–300.

[5] M. Foote and J. J. Sepkoski. Absolute measures of the completeness of the fossil record. *Nature* 398 (6726) (1999), 415–417.

[6] M. Foote. Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis. *Paleobiology* 27 (4) (2001), 602–630.

[7] M. Foote and D. M. Raup. Fossil preservation and the stratigraphic ranges of taxa. *Paleobiology* 22 (2) (1996), 121–140.

[8] G. T. Lloyd, P. N. Pearson, J. R. Young, and A. B. Smith. Sampling bias and the fossil record of planktonic foraminifera on land and in the deep sea. *Paleobiology* 38 (4) (2012), 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. *Geology* 23 (8) (1995), 731–734.

[10] S. C. Wang and C. R. Marshall. Estimating times of extinction in the fossil record. *Biology Letters* 12 (4) (2016), 20150989.

[11] T. H. G. Ezard, G. H. Thomas, and A. Purvis. Inclusion of a near-complete fossil record reveals speciation-related molecular evolution. *Methods in Ecology and Evolution* 4 (8) (2013), 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. *Ecology Letters* (2016), 1–8.

[13] T. H. G. Ezard, T. Aze, P. N. Pearson, and A. Purvis. Interplay Between Changing Climate and Species' Ecology Drives Macroevolutionary Dynamics. *Science* 332 (6027) (2011), 349–351.

[14] L. H. Liow, H. J. Skaug, T. Ergon, and T. Schweder. Global occurrence trajectories of microfossils: environmental volatility and the rise and fall of individual species. *Paleobiology* 36 (2) (2010), 224–252.

[15] F. Mekik and R. Anderson. Is the core top modern? Observations from the eastern equatorial Pacific. *Quaternary Science Reviews* 186 (2018), 156–168.

[16] W. Broecker, K. Matsumoto, E. Clark, I. Hajdas, and G. Bonani. Radiocarbon age differences between coexisting foraminiferal species. *Paleoceanography* 14 (4) (1999), 431–436.

[17] F. Mekik. Radiocarbon dating of planktonic foraminifer shells: A cautionary tale. *Paleoceanography* 29 (1) (2014), 13–29.

[18] T.-H. Peng and W. S. Broecker. The impacts of bioturbation on the age difference between benthic and planktonic foraminifera in deep sea sediments. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 35 (1984), 346–352.

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

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

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

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

[23] P.-C. Bürkner. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80 (1) (2017), 1–28.

[24] P.-C. Bürkner. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* 10 (1) (2018), 395–411.