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. Our models have a 70-80% probability of correctly forecast the rank order of extinction risk for a random out-of-sample species pair, implying that determinants of extinction have varied only modestly through time. We find that models which include either historical covariates or account for variation in covariate effects over time yield equivalent forecasts, but a model including both is overfit and yields biased forecasts. 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 suggestion 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.

Extinction intensity (average rate) and selectivity (difference in risk among taxa) have varied greatly through time, and the relative risk of extinction exhibited by different taxonomic and ecological groups can provide insights into the drivers of both background and mass extinction [3–8]. Many studies have examined the effects of various potential predictors on extinction risk through time [3, 5, 9–14] or refined methods for identifying and measuring these effects [15–20]. These studies have produced a growing body of knowledge regarding which factors have been general determinates of extinction risk in the geological past.

A related question that has received much less attention is how successful we might expect to be when using this knowledge to attempt to predict future extinction events.

Because future extinctions are unobservable we cannot directly evaluate the ultimate performance of such predictions. However, we can take a given point in the geological past, develop a predictive model based on extinction patterns prior to that point, and assessing the predictive performance of this model on subsequent (e.g. "future", from the point of view of the model) extinction/survival events. Putting aside the very important question of how human activities will alter the determinants of future extinction risk, such an approach provides a framework for evaluating the expected accuracy of future risk assessments based on past extinction events.

Here we take this approach, using as a model system the Cenozoic record of skeletonized marine planktonic microorganisms (Foraminifera, Radiolaria, Diatoms, and Coccolithophores). This record has several key strengths for our purposes: planktonic microorganisms are widespread and abundant in pelagic habitats, have high preservation potential, and because of their utility for biostratigraphic, paleoclimatic, and oceanographic study they have been the focus of an extensive international coring and study effort [21, 22]. A compliation of these data is readily available through the Neptune database, an online repository of species occurrences obtained through the Deep Sea Drilling Program and the Ocean Drilling Project [21, 22]. This database provides abundant samples in space and time, a high degree of temporal resolution for the entirety of the Cenozoic, and has an taxonomic synonymization framework for dealing with 50+ years of taxonomic opinion [21] – as close to ideal data for this analysis as possible. 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.

The overall question of how well models based on past extinction patterns perform at forecasting future extinctions depends in part on model complexity. Simple models requiring only a few parameters are in general preferable because more complex models run a greater risk of being overfit to the observations on which they are trained. In addition, many traits that might influence extinction risk among extant species are difficult to assign confidently to extinct species. For these reasons we elect to focus on baseline models which include only a few parameters that have been shown to be important and/or consistent determinants of extinction risk in the marine fossil record. 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 [3, 23–29]. In addition to geographic range, we also considered global climate state and change in climate state since previous observation in order to evaluate the influence of climate or climate change trajectory on extinction risk. Finally, we included species age, both because previous studies

of planktonic taxa have found it to be a determinant of extinction risk and because its inclusion in our models is critical to their nature as survival models (see Model Specifications below). We reiterate that our primary objective is to evaluate the predictive performance of simple models that include only a few general parameters; more complex models including other likely determinants of extinction risk such as skeletal mineralogy, trophic ecology, and thermal tolerance range might well perform better.

There are a number of ways in which past extinction patterns might be used to model present risk. The simplest case assumes that relationships between predictors (hereafter covariates) and extinction risk have been constant through time. A more complex but more realistic case allows relationships between covariates and extinction risk to vary through time, consistent with evidence for temporal variation in extinction selectivity regime. Finally, an important consideration is the degree to which species geographic range trajectories exhibit deterministic versus Markovian behavior [25, 26, 30, 31]. In the former case, knowledge of the specific past trajectory of a species whether its range has expanded or contracted from some point in the past to the present might help to improve assessments of its current risk. In the latter case only the current geographic range of the species would convey useful information about current and future risk (although that assessment would still be based on the relative extinction risk of species that had similar ranges at different points in the geological past).

Below, we evaluate four models along a spectrum from simplest (fixed covariate effects, Markovian range dynamics) to most complex (varying covariate effects, deterministic range dynamics). We ask (1) how well they perform at classifying species as extinctions or survivors in the data they were fit to, and (2) how well they perform at classifying species as extinctions or survivors in future data that were not used in fitting the models.

2 Materials and Methods

2.1 Data Specifications

We analyzed microfossil occurrence information from the Neptune Database `http://www.nsb-mfn-berlin.de/nannotax` [21, 22]. This occurrence-based dataset includes calcareous nannoplankton, diatoms, planktonic foraminifera, and radiolarians. Occurrences were filtered to include only those species with first occurrences no earlier than 63 Mya (millions of years ago). This filtering criterion excludes taxa that survived the K/Pg extinction or arose during this 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 (million year) bins based on the estimated age of the fossil occurrence as listed in the Neptune Database. After binning, each species' geographic range was calculated for each of the 1 My bins in which it occurred. Geographic range was calculated as the minimum spanning tree distance between all observations of that taxon during that temporal bin; this distance was measured in kilometers. Minimum spanning tree distance was calculated using the `GeoRange` package for R [32].

We also included how a taxon's geographic range has changed since its last three observation
111 times. We measured this change in geographic range by calculating the difference in geographic range between an observation and that taxon's three previous occurrences. Change between the most recent and the three previous occurrences was calculated individually for each of
114 those lags. If there are not enough previous observations to calculate, then that value is recorded as a 0. These differences were calculated after minimum spanning tree distance was transformed and standardized (see Supplement Section S1.1.2)

117 Average global temperature of each 1 My bin was calculated from estimates based on Magnesium/Calcium isotope ratios Cramer et al. [33]. We use Mg/Ca rather than oxygen isotopes to avoid confounding effect of varying ice-volume – this property is of particular importance
120 for this analysis as polar ice-caps develop midway through the Cenozoic. Estimating temperature over long periods of time from Mg/Ca ratios also suffers from complications because Mg/Ca based temperature estimates requires benthic foram and seawater Mg/Ca isotope
123 ratio information. Because seawater Mg/Ca ratio has changed over time, the method to estimate temperature used Cramer et al. [33] attempts to account for this unknown in order to obtain accurate, albeit uncertain, temperature estimates. Our data source, Cramer et al.
126 [33], estimated temperature for every 0.1 My interval between 0 to 63 Mya. The temperature estimate for each 1 My interval was calculated as the mean of all estimates within that interval.

129 We also included the global temperature from the previous time that taxon was observed. If there are not enough previous observations to calculate, then that value is recorded as a 0. This lag was calculated after global temperature was transformed and standardized (see
132 Supplement Section S1.1.2).

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

See Supplemental Section S1.1 for a further explanation on how observations were temporally
141 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
144 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 [34]. We considered four different models involving different permutations of covariate effects (fixed or time-varying)
147 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 varying-intercept term. This factor may or may not contribute to differences in species extinction risk over time [10, 35–39], but its inclusion in our model is critical to its nature as a survival model [34]. The effect of species age is allowed to vary by taxonomic group.

Similarly, we included time of obseration as an additional varying-intercept term to account for changes average global extinction risk over time that are not related to the covariates included in this model. This varying-intercept is further allowed to vary by taxonomic group. This varying-intercept term allows us to tease apart the differences in extinction risk associated with time of observation versus age since first observation. An important note is that for our V and VP models, the covariation between this varying-intercept and the varying-slopes of our covariates is explicitly modeled (see Supplement Section S1.2).

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 Supplement Section S1.2. Additionally, the full description of how these models were implemented and coded, including choice of priors, is available in Supplement Section S1.2.

2.3 In-sample and out-of-sample forecasting

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

In-sample forecasting is a posterior predictive check in that we are estimating our model's ability to correctly classify the data to which it was fit. Posterior predictive checks are a type of sensitivity analysis because we are checking the quality of model's fit to the data. If our models have poor in-sample forecasting performance, than our models are not adequate descriptors of the data and will most likely make poor out-of-sample predictions. In-sample forecasting measures, however, are not the same as understanding our models' ability to forecast future extinctions or if our models are overfit to our data and produce biased out-of-sample estimates [40].

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 [40–42]. Prior

to cross-valdiation, the data is divided into k nearly even segments or folds – for a time series, this means the data is divided into k continuous sequences. Next, 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. Next, the model is fit to the first four folds combined and then is used to forecast the fifth fold. This process continues until $k - 1$ folds are included in the fitting the model and the final fold is predicted from this model. When combined, the results from these forecasts are then combined to yield our estimate of expected out-of-sample performance. In 5-fold cross validaition, the data is divided into five folds the cross-validation procedure yields predictions for four of the folds.

Cross-validation is a procedure for estimating a models expected out-of-sample error. Information criteria such as AIC or WAIC are approximations of out-of-sample predictive error as estimated by cross-validation [40, 43]. Cross-validation implicitly takes into account model complexity because when a model is overfit to its data, out-of-sample predictions will be biased and inaccurate [40]. A high degree of similarity between out-of-sample and in-sample estimates indicates that the model is not overfit to the data (though it is not necessarily an adequate descriptor of the data). Cross-validation is preferable to simple metrics such as AIC because instead of a single value it produces, an entire posterior distribution of estimates.

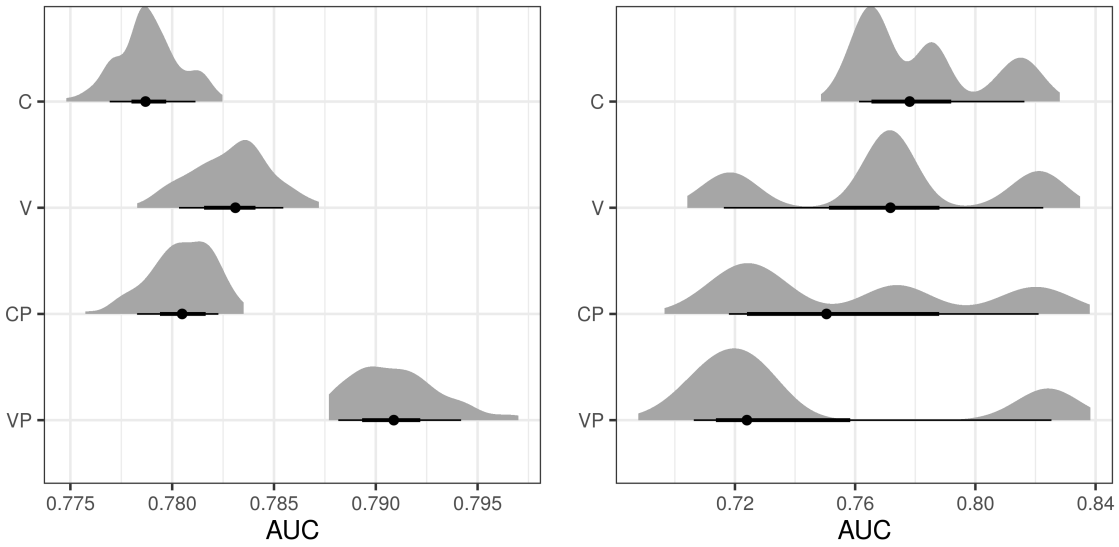
The relative adequacy of the four model variants was compared using the area under the receiver operating characteristic curve or AUC [44, 45]. This measure is commonly used to measure the performance of classification models as it has the desirable characteristic of comparing the model's true positive 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 [44, 45]. AUC values of approximately 0.8 or greater can be considered "good" [46], so we consider values between between 0.7 and 0.8 as "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. The entire analysis was coded in R and uses tidyverse and tidyverse adjacent tools such as `dplyr` [47], `purr` [48], and `tidybayes` [49]. All of our models were written using the `brms` [50, 51] R package, which implements Stan-based Bayesian models which are fit via Hamiltonian Monte Carlo [52].

3 Results

The primary focus of this study is on understanding how well our models forecast future extinction events by comparing our in-sample and out-of-sample forecast estimates. A presen-

tation of the posterior estimates for the regression coefficient estimates from our VP model (Table 1) is available our Supplemental Materials (Section S2).



(a) In-sample

(b) Out-of-sample

Figure 1: Comparisons of measures of model performance for both in-sample (1a) and out-of-sample (1b) cross-validation. The area under the receiver operating charactered curve (AUC) was calculated for each model. these estimates are calculated from the models posterior predictive distribution (1a) or from predictions made to new data (1b), respectively. Marked below the posterior distributions are the median AUC and 50% and 80% posterior intervals for all observations in our dataset. Models with higher AUC values indicate better performance over models with lower AUC values. AUC is bounded between 0.5 and 1. See Table 1 for an explanation of the four models (C, V, CP, VP).

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, acting as quality control and sensitivity analysis. Comparison between the posterior distributions of in-sample AUC for each of the four models demonstrates that the parameter rich model VP has the greatest median in-sample AUC when compared to the other three models, while there is substantial overlap in the posterior distributions of the forecasts from the other three models (Fig. 1a).

However, the actual difference in forecast AUC result between the VP model and the other three models is extremely small (0.01 AUC unit), and all of the in-sample AUC estimates from our models are concentrated between an AUC value of 0.775 and 0.795 (Fig. 1a). This result indicates that the practical difference in performanc between these models might be so small that there is no practical benefit that the VP model over the other three. Ultimately, determining which of these models produces the best forecasts of future extinctions requires comparing these in-sample results to our out-of-sample results (see below).

The in-sample forecasts from our four models over time are broadly similar between taxonomic groups (Fig. 2). Our in-sample forecasts for Diatoms are the weakest of the taxonomic groups as all four models have an equal number of intervals with no predictive power (AUC approximately 0.5). In contrast, our best in-sample forecast results are for Radiolarians where for any of our models there is at most 1 interval with almost no predictive power. Ultimately, our in-sample forecasts over time by taxonimc group are broadly consistent between our four models.

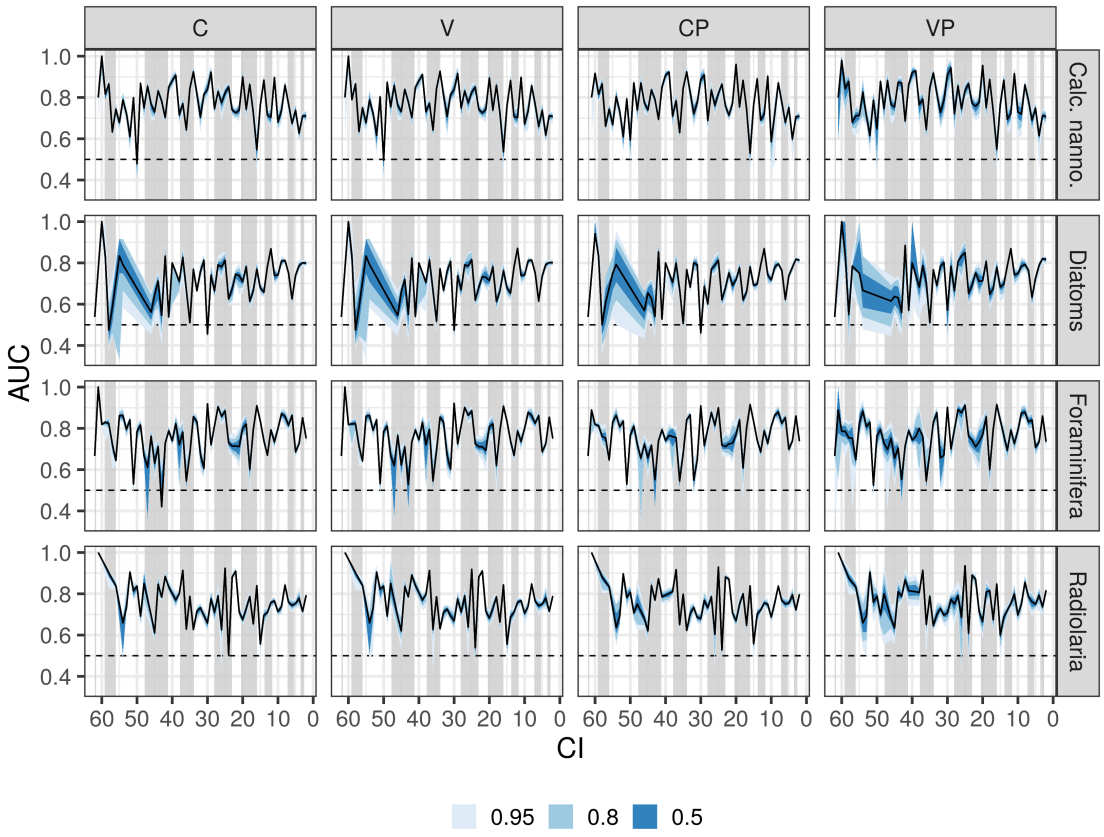


Figure 2: Understanding model adequacy over time and taxonomic group by comparing in-sample forecasting performance measured by AUC for each of the four models. These estimates reflect each model's fit to the various taxonomic groups over time. The black line corresponds to the median AUC value, while 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. The grey intervals mark the geologic ages of the Cenozoic. See Table 1 for a description of each of the four models (C, V, CP, VP).

3.2 Out-of-sample forecasting performance

Expected out-of-sample forecasting performance was estimated using five-fold cross-validation for time series [41, 42]. Our out-of-sample forecast AUC estimates demonstrate a broader

range of results, with AUC estimates ranging between approximately 0.7 and 0.85 (Fig. 1a, 1b). While our VP model was the model with greatest in-sample forecasting performance (Fig. 1a), this model has the greatest decrease in out-of-sample forecasting performance (Fig. 1b). In comparison, the other three models demontrstrate a much smaller difference between in-sample forecast and out-of-sample forecast AUC values which indicates that our VP model is overfit to our data, and that one of the simpler model would be preferable for predicting future extinctions. This result means that our model which includes the historical covariates (e.g. change in geographic range) and allowing these effects to vary over time produces biased extinciton forecasts. Interestingly, including either historical covariates like change in geographic range or allowing the effects of geographic range and other covariate effects to vary over time have approximately equal effect on forecasting future extinction events (Fig. 1).

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 diatoms, 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 forecast which species of randomly selected extinct-extant pair is more likely to go extinct between 70% to 80% of the time (Fig. 1b). These results confirm that past extinction patterns can provide valuable information about which extant species are most threatened with extinction in the near geological future.

Three of the four models we evaluated 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). The model with the best in-sample forecasting results includes the historical covariates and allows

all covariate effects to vary over time. However, the out-of-sample forecasts from this model are biased forecasts, indicating that it is overfit to our data 1b). The model that includes historical covariates such as geographic range trajectory yields out-of-sample forecasts with nearly identical results to the model that allows covariate effects to vary over time but does not include historical covariates.

We note that both including historical covariates such as change in geographic range and allowing covariate effects to vary over time are different ways of encoding information from the past. Including historical covariates in a model but not allowing covariate effects to vary over time encodes the past explicitly but assumes that covariates effects are constant over time. Allowing covariate effects to vary over time, on the other hand, does not explicitly encode "the past" into our model but instead models how covariate effects vary over time which allows the past to be implicitly included in our model. By modeling this variation, forecasts made for future extinction events can allow for a wide range of potential effects of our covariates on predicting extinction. Comparing our out-of-sample forecast results indicates that these approaches yield approximately equal forecasting performance (Fig. 1b). Our results supplement those of Kiessling and Kocsis [2] which examined using differences in geographic range over time to predict extinction risk.

An 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 predicatable than it has been in the past [1, 13]. Our model with historical covariates assumes that extinction selectivity with respect to these covariates is constant through time, but given growing evidence that human impacts substantially alter extinction risk dynamics [1, 3, 13], this assumption may not be valid and may limit or bias our ability to predict extinction in truly novel environmental regimes. Thus, it might be preferable to model the variation in extinction risk and selectivity over time instead of relying solely on measures from a taxon's past such as change in geographic range over time. For this reason, while our CP and V models yield similar out-of-sample forecasts, we believe the V model offers more practical benefits for predicting extinction risk in future, anthropogenically impacted environments.

The relative quality and consistency between in-sample and out-of-sample forecasting performance for three of the four models 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 the models used here is because we lack additional life history or ecological information for many of the marine micro- and nannoplankton included in this study. Foraminifera are an exception to this problem as aspects of life history, ecology, and physiology are known for many foram species [5]. However, comparable information does not exist for 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 simple taxonomic, geographic, and historical predictors. Although a model that includes historical covariates such as change in geographic range and change in climate between observations while also allowing covariate effects to vary over time performs best at in-sample prediction, this model is overfit to our data and produces less accurate out-of-sample forecasts than our three less complex models. The remaining three models yield nearly equivalent out-of-sample forecasts, suggesting that including historical information via either explicit modeling of historical covariate effects or modeling how covariate effects have changed over time does not diminish and may ultimately improve our ability to forecast future extinctions. The results of this simple exercise suggest that conservation decisions could indeed be bolstered by including fossil data.

5 Acknowledgements

We thank Jonathan Payne, Thomas Ezard, Erin Saup, and Peter Roopnarine for useful discussions. Wolfgang Kiessling, Matthew Clapham, and an anonymous reviewer provided thoughful reviews that strengthened the manuscript consideraly. This work was funded by a David and Lucile Packard Fellowship to SF.

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] 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.

[4] 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.

[5] 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.

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

[7] S. C. Wang and A. M. Bush. Adjusting global extinction rates to account for taxonomic susceptibility. *Paleobiology* 34 (4) (Dec. 2008), 434–455.

[8] A. H. Knoll, R. K. Bambach, J. L. Payne, S. Pruss, and W. W. Fischer. Paleophysiology and end-Permian mass extinction. *Earth and Planetary Science Letters* 256 (2007), 295–313.

[9] 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.

[10] 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.

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

[12] 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.

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

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

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

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

[17] 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.

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

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

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

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

[22] 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.

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

[24] 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.

[25] 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.

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

[27] 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.

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

[29] 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.

[30] 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.

[31] 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).

[32] J. Boyle. *GeoRange: Calculating Geographic Range from Occurrence Data*. R package version 0.1.0. 2017.

[33] 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.

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

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

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

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

[42] 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.

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

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

[45] 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.

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

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

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

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

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

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

[52] S. D. Team. Stan Modeling Language Users Guide and Reference Manual. 2017.

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 | time[e]/phylum[f]) + (1 | age[g]/phylum)
V	Varying effects, no historical cov.	Geographic range, temperature	event ∼ range + temp + (1 + range + temp | time/phylum) + (1 | age/phylum)
CP	Constant effects, historical cov.	Geographic range, change in geographic range, temperature, previous temperature	event ∼ + range_diff1[g] + range_diff2[h] + range_diff3[h] + temp + temp_lag[i] + (1 | time/phylum) + (1 | age/phylum)
VP	Varying effects, historical cov.	Geographic range, change in geographic range, temperature, previous temperature	event ∼ range + range_diff1 + range_diff2 + range_diff3 + temp + temp_lag + (1 + range + range_diff1 + range_diff2 + range_diff3 + temp + temp_lag | time/phylum) + (1 | age/phylum)

[a] See Supplemental 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] Time of observation.
[f] Taxonomic group of species (i.e. Foraminifera, Diatoms, Radiolarians, Calcareous nannoplankton).
[g] Age at observation.
[h] Change in geographic range since last observation (number indicates how lags).
[i] Temperature at previous observation.

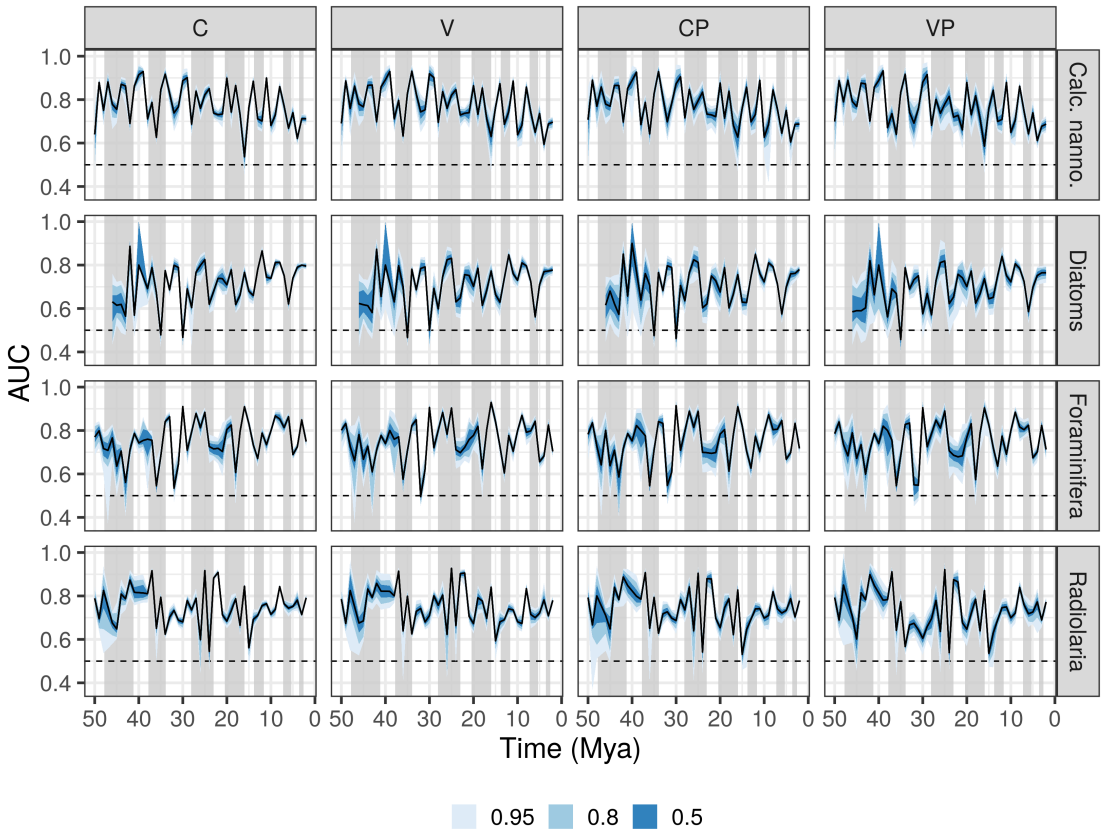


Figure 3: Comparison of our models ability to forecast future extinction events as measured by out-of-sample AUC values over time as aggregated by taxonomic group for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogeneity in performance 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. The grey intervals mark the geologic ages of the Cenozoic. See Table 1 for a description of each of the four models (C, V, CP, VP).

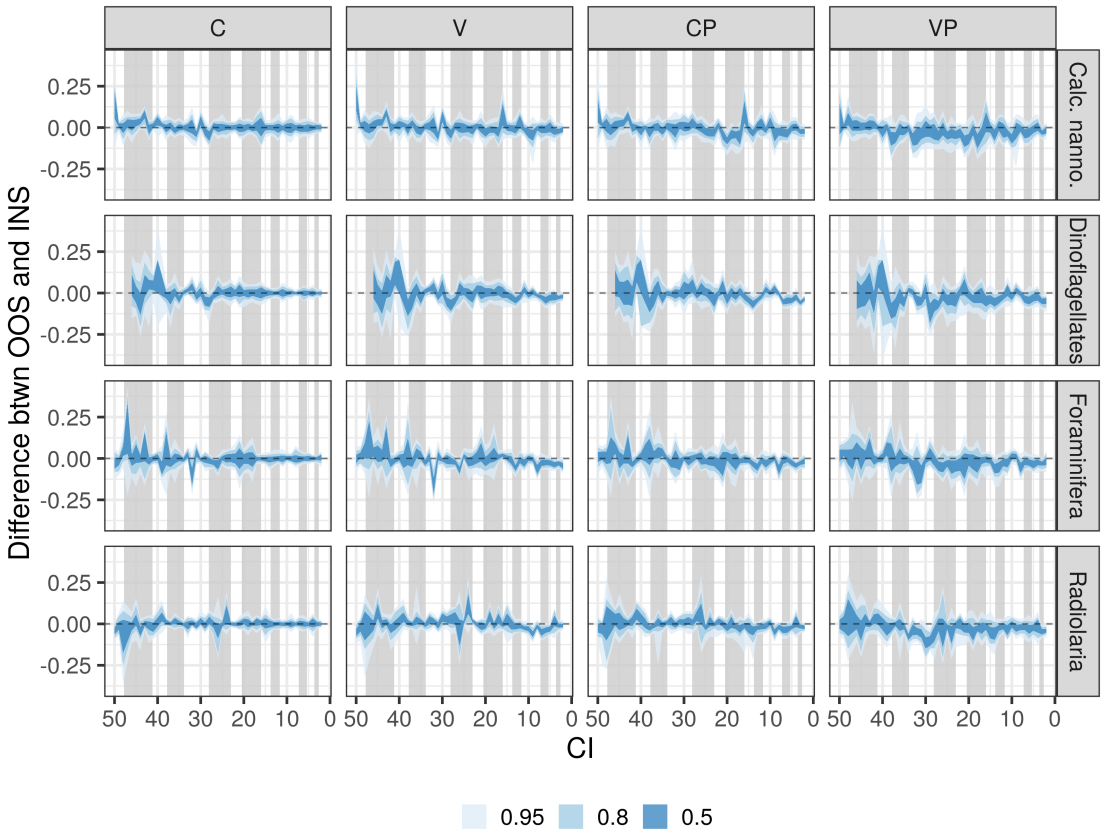


Figure 4: Comparing our models' ability to forecast out-of-sample extinction compared to our models' in-sample forecasts. Congruence between these measures indicates that our models are not necessarily overfit to the data. This value is calculated as the values presented in Figure 3 minus those values 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 (C, V, CP, VP).