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

Smits, Peter
psmits@berkeley.edu

Finnegan, Seth
sethf@berkeley.edu

February 11, 2019

Abstract

One of the great promises of paleobiology is that by studying the past we can better predict the future. This promise is particularly pertinent given as risk assessments for some modern species could potentially be improved by examining past extinction patterns and by using paleontological records to establish geographic range and abundance trajectories on geological timescales. Any effort to assess future risk based on past extinctions and range trajectories must address two key questions: (1) At a given timescale, are geographic range and extinction risk trajectories deterministic (past trends are likely to continue into the future) or Markovian (the future depends only on the present state)? (2) Given knowledge of past extinction/survival patterns and the present geographic ranges of extant taxa, how accurate are extinction risk predictions?

To address these questions we analyze the fossil record of Cenozoic planktonic microfossil taxa (foramanifera, radiolarians, diatoms, and calcareous nanoplankton). Using a model of species survival, we analyzed how survival probability changes over time as a function of species age, time of observation, current geographic range, most recent change in geographic range, global temperature average, and the lag of global temperature. Our best supported model includes the historical covariates, change in geographic range and lag of global temperature, which indicates that the past improves our estimates of the present and future.

Our results show that our best performing model has an approximately 78% median probability of correctly ranking the relative extinction risk any two randomly selected species. Including the historical covariates and allowing their effects to vary over time yields marginally better predictions than not including them. However, the improvement in predictive power by including these historical covariates is modest at best, absent at worst, and ultimately reflects the extremely stochastic nature of species extinction. Using k-fold cross-validation, we found that in-sample model performance measures are approximately equal to out-of-sample performance, meaning that we can have confidence that our conclusions about our ability to predict species extinction risk in the future with similar accuracy to when we predict species extinction risk in the past. These results imply that at million-year timescales geographic range trajectories are nearly Markovian, perhaps because the processes driving geographic range changes vary on substantially shorter timescales. The effect of change in geographic range on survival most likely stands for many interacting and unobserved processes which in-turn produce that species' geographic range and its affect on survival.

We find that including information on a species' change in geographic range size on average improves our predictions of species survival at million-year timescales. However, the effect of change in geographic range is much smaller than the effect of current geographic range, and highly variable through time as the effect changes sign and there are times where there is little evidence for any effect of past geographic range. These results reflect the difficulty of estimating species extinction, and that while including historical covariates does improve model performance, that gain is very small. The results of this study reinforce the importance of the promise of paleontology and using the past to predict the future.

1 Introduction

Being able to predict which species are more likely to go extinct than others is critical for making good conservation decisions to limit the impact of the current biodiversity crisis. We cannot know, however, we do not yet know which species are going to go extinct because this has not happened yet – it is unobservable. We approach this this problem by analyzing the past in order to

predict the future. The fossil record preserves past extinction events, allowing us to develop a predictive model of species extinction based on this record and the properties of the observed species, both extinct and extant [16, 26]. By assessing the predictive performance of this model on unobserved data, we can quantify how precise our best estimates will be for future extinctions – we ask the probability that, given two random species, we correctly rank their relative risks of extinction.

By studying how species vary in their extinction risk over time and we can assess which species are at greater risk under unobserved conditions. We know that a species' risk of extinction varies over time in both intensity (average rate) and selectivity (difference in risk between taxa) [12, 41, 42]. Species, after all, can go extinct at any "moment" and the relative risk of extinction exhibited by different taxonomic groups and how that risk varies over time is an important dynamic which shapes the rate and structure of extinction. What has not been evaluated is that as extinction intensity and selectivity change over time, how accurate are our assessments based on past events likely to be when applied to the future? By specifically including and modeling the temporal variation in extinction risk, we are able to improve our overall predictions because we incorporate and explicitly model differences between observations from across a range extinction intensities and selectivities.

By analyzing extinction and survival data from the fossil record, the hope is this can aide in predicting the extinction risk of extant species – after all, the present must at some level be a function of the past. Past paleobiological studies of extinction have frequently focused on identifying and measuring the effect of various predictors on extinction risk [12, 19, 25, 26, 41, 43, 44] or on how to identify or measure these effects [1–5, 18]. This focus means that while we have a good understanding of which factors are strong and general determinates of extinction risk, we have less knowledge of how accurate or strong our predictions about the differences in extinction risk are. For example, while a predictor may be "significant" when comparing the odds of extinction risk between groups, the practical difference that predictor makes on prediction can be minimal [24]. By including the kinds of biological and abiotic predictors that have been shown to affect differences in extinction risk, we can quantify their actual, as opposed to relative, effects on predictive performance.

A related question is if the changes to biotic or abiotic predictors, and not

just their values, are similarly important factors for predicting extinction. For example, we know that a species' global geographic range changes over its duration [20, 33, 35, 36]. We also know that a species' geographic range size is a good predictor of differences extinction risk [29–31, 41]. This begs the question: how does a species' extinction risk change over its duration? While the phenomenon of species' geographic range change over time has been studied [20, 33, 35, 36], the potential predictive impact of this change has been under-evaluated (but see Kiessling and Kocsis [33]). For example, does a species' extinction risk increase if that species decreased in global geographic range size over 1 million years? Here, we explicitly model and quantify the effects of changing geographic range as well as differences in global climate on how well we can predict species extinction. Similarly, we include species geologic age at time of observation as a potential predictor of extinction – a factor that may or may not contribute to differences in species extinction risk over time [9, 13, 15, 37, 44, 48]. Importantly, the inclusion of these "historical" predictors allows us to more fully evaluate the question of how much information about a species' past is necessary or useful when predicting a species' risk of extinction.

For this kind of exercise, we chose to analyze what is the longest continuous and best resolved fossil record – that of skeletonized marine plantonic microorganisms from the Cenozoic such as Foraminifera, Radiolarians, Diatoms, and calcareous nannofossils (e.g. coccolithophores). This data is available through the Neptune database, an online repository of species occurrences obtained through the Deep Sea Drilling Program and the Ocean Drilling Project [34, 45]. 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.

Rarely are we able to analyze long periods of geological time at fine resolutions – below the 5-10 My scale. Due to substantial effort and the unique biology of the system, the microfossil provides us the unique opportunity to analyze ecological and evolutionary patterns at approximately million-year time scales. Typical "exceptional" fossil records tend to be of individual taxonomic groups and for rarely longer than 10 million years. 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 the occurrences are recorded from – there is no equivalent fossil record.

By analyzing patterns of extinction and global occurrence at fine temporal scales, we can better elucidate how well we can predict species extinction at human-relevant scales.

Being able to analyze over 60 million years of fossil occurrences allows to actually quantify how accurate our predictions are in general, but also how much variation there is in predictive accuracy over time and in many different environmental contexts. Specifically, we might expect that our model's predictive performance is best during prolonged periods of similar stress, such as the Eocene-Miocene transition [51] – more samples from similar environments inherently improves future predictions in unobserved, but similar conditions. Alternatively, we would expect our model based predictions of extinction surrounding the Paleocene-Eocene Thermal Maximum may be less accurate because there are inherently fewer samples from the rapid climatic event [51].

2 Materials and Methods

2.1 Data Specifications

We analyzed microfossil occurrence information which was downloaded from the Neptune Database `http://www.nsb-mfn-berlin.de/nannotax` [34, 45]. All occurrence information was downloaded for calcareous nannofossils, diatoms, foraminifera, and radiolarians – these occurrences span the entire globe between 120 and 0 million years ago (Mya). This dataset of occurrences was then filtered to just those species which have their first occurrence at most 63 Mya. This choice means that our analysis avoids those taxa which survived the K/Pg boundary, those taxa which arose just after the K/Pg boundary, and means that our occurrence histories line up with the temperature time-series which was used as a predictor of extinction (discussed below).

All fossil occurrences were assigned to 1 My bins based on the estimated age of the fossil occurrence. Because the estimated ages of each occurrence is a product of core-specific age-models and can be overly precise, the hope is that by binning the data this smooths over the between-core heterogeneity and thus homogenizes our disparate data sources. The occurrence histories of each species were then given binary codes used to model the presence or extinction of those species. For every occurrence of a species, except the last,

that species is considered to have survived and was marked with a 0. The last occurrence of that species is considered the bin in which the taxon has gone extinct – and is assigned a 1. This protocol means that we are reading the fossil record "as written," a practice that is potentially dangerous as it is a overconfident statement of preservation and may be shortening the actual duration of that species [1–3, 17, 18, 21, 22, 38, 39, 49]. However, this practice is common with marine microfossil data due to their exceptional preservation rate [10–12, 36]. In fact, with marine microfossils collected from cores a bigger problem may be over extending the duration of a species due to mixing and smearing within the cores CITATIONS.

A 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 globe) between any two occurrences; this measure is also called a geodesic. This distance was measured in kilometers. The geodesic distance was then log-plus-one transformed and then standardized by zero-centering the data and then dividing by its standard deviation so that geographic range had mean 0 and standard deviation 1. This standardization means that a regression coefficient associated with this covariate describes the change in extinction probability per change in standard deviation of geographic range, and that coefficients associated with similarly standardized covariates will be directly comparable in magnitude [24].

Temperature data used as covariates in this analysis are based on Magnesium/Calcium isotope ratios sourced from Cramer et al. [8]. These elemental ratios are considered more accurate estimates of past global temperature when compared to the frequently used Oxygen isotope-based estimates; this is because Mg/Ca based estimates are not effected by ice-volume and fresh-water input (e.g. meteoric water) which can alter Oxygen isotope ratios without reflecting changes to the climate itself. This property is of particular importance for this analysis as polar ice-caps develop midway through the Cenozoic. Our data source, Cramer et al. [8], provides temperature estimates for every 0.1 My from 0 to 63 Mya. We binned these estimates into 1 My intervals as we did with the fossil occurrences. The temperature estimate for each 1 My interval was calculated as the mean of all estimates within that interval (Fig. 1). Temperature was then transformed and standardized the in the same manner as geographic range (above).

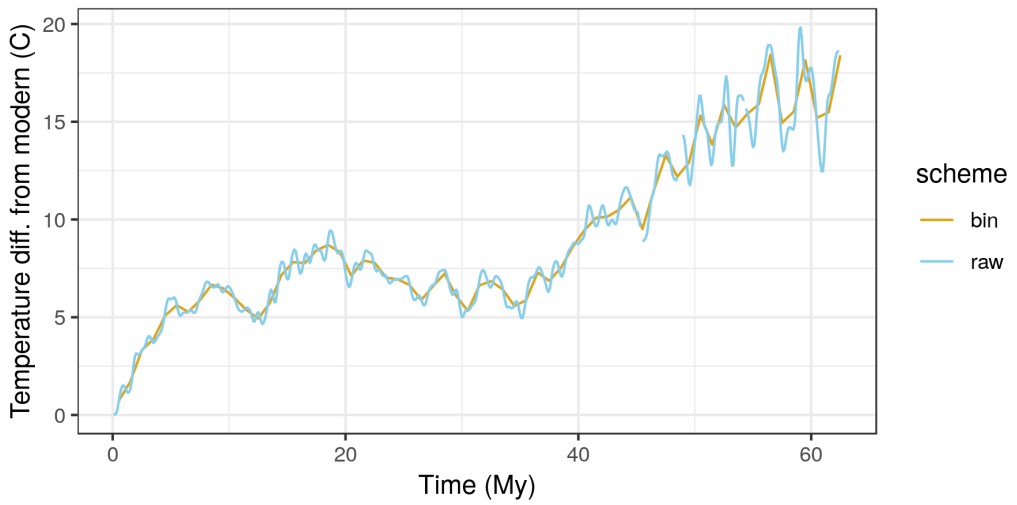

Figure 1: Comparison of initial temperature estimates from Cramer et al. [8] (goldenrod) versus the binned values used in this analysis (blue). The initial values are for every 0.1 My while our bins are defined for every 1 My.

2.2 Model Specifications

In survival analysis, the hazard function describes the instantaneous rate of extinction of a species given its age and relevant covariates. The hazard function is defined as the conditional probability of a species going extinct by the end of the t-th interval given that it survived up until t and the relevant covariate information X for all k 1 My intervals [47]. For the discrete time intervals $T = 1, \cdots, k$, extinction is defined as $T = t$. The discrete time hazard function is defined as

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

The hazard function (Eq. 1) 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 [47]. $h(\Theta)$ is then modeled as with any regression as it is defined for all real-values. In this case, we opted for a hierarchical/mixed-effects model with multiple non-nested varying intercepts and slopes [24].

Our covariates matrix X is a $N \times D$ matrix where N is the total number of

observations and D is the total number of covariates. The first column of X is entirely 1's as it corresponds to the intercept term in the regression model. The next two columns of X are two aspects of geographic range as continuous covariates: geographic range r during interval t, and the difference d between the geographic range at $t-1$ and t. The difference in geographic range was calculated from the transformed and standardized geographic range values; this means that change in geographic range is in units of changes in standard deviations. The final two columns are two aspects of global temperature: mean temperature during interval t, and the lag of mean temperature (i.e. mean temperature during interval $t-1$.) As with change to geographic range, the lag of temperature is based on the transformed and standardized temperature estimates.

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

This logistic regression model, minus the final selection of priors, is expressed as

$$\begin{aligned} t_i &\sim \text{Bernoulli}(\Theta) \\ \Theta_i &= \text{logit}^{-1}(X_i B_{w[i],p[i]} + A_{d[i],p[i]}) \\ B_{w,p} &\sim MVN(\alpha_p, \Sigma_B) \\ \alpha_p &\sim MVN(\mu, \Sigma_\alpha) \\ A_{w,p} &\sim MVN(\delta_p, \Sigma_A) \\ \delta_p &\sim \text{N}(0, \sigma_\delta) \end{aligned} \tag{2}$$

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.

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 [46].

In total, these prior choices can be expressed as

$$
\begin{aligned}
\mu_d &\sim \begin{cases} N(-2,5) & \text{if } d = \text{intercept} \\ N(-1,1) & \text{if } d = \text{geo. range} \\ N(0,1) & \text{else} \end{cases} \\
\delta &\sim N(0,1) \\
\Sigma_B &= diag(\tau_B)\Omega_B diag(\tau_B) \\
\Sigma_\alpha &= diag(\tau_\alpha)\Omega_\alpha diag(\tau_\alpha) \\
\Sigma_A &= diag(\tau_A)\Omega_A diag(\tau_A) \\
\tau_B &\sim C^+(5) \\
\tau_\alpha &\sim C^+(5) \\
\tau_A &\sim C^+(5) \\
\Omega_B &\sim LKJ(1) \\
\Omega_\alpha &\sim LKJ(1) \\
\Omega_A &\sim LKJ(1)
\end{aligned} \tag{3}
$$

We considered four variants of the model described above: 1) historical covariates with time-varying intercepts and slopes, 2) no historical coviarates with time-varying intercepts and slopes, 3) historical covariates without time-varying intercepts and slopes, and 4) no historical covariates and no time-varying intercepts and slopes (except age effect). The second and fourth models modify the number of columns in X by removing two of the covariates. The third and fourth model simplify the first model to just varying intercepts without varying slopes.

2.3 Model Parameter Estimation

We implemented our model using the `rstanarm` package for the R programming language [46]. This package provides an interface to the Stan probabilistic programming language for writing hierarchical/mixed-effects models in native R. Posterior estimates were obtained through Hamiltonian Monte Carlo, using 2000 steps divided equally between warm-up and sampling. In order to prevent divergent transitions the adapt delta value was increased to 0.9999; all other HMC/NUTS sampling parameters were kept at the defaults for rstanarm 2.18.2 [23].

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

```
form <- event ~ range + range_diff + temp + temp_lag +
                (1 + range + range_diff + temp + temp_lag | mybin/phylum) +
                (1 | age/phylum),
stan_glmer(formula = form,
           data = data,
           family = 'binomial',
           prior = normal(c(-1, 0, 0, 0), rep(1, 4), autoscale = FALSE),
           prior_intercept = normal(0, 10, autoscale = FALSE),
           prior_aux = cauchy(0, 3, autoscale = FALSE),
           chains = 4,
           thin = 4,
           adapt_delta = 0.999999)
```

Similarly, our full model can be implemented in brms [6, 7] as:

```
priors <- c(set_prior('normal(0, 10)', class= 'Intercept'),
            set_prior('normal(0, 1)', class = 'b'),
            set_prior('normal(-1, 1)', class = 'b', coef = 'range'),
            set_prior('cauchy(0, 3)', 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 [46].

2.4 Model adequacy

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

Relative and absolute model adequacy of the four variant models was compared using the area under the receiver operating characteristic curve or AUC [14, 40]. This measure is commonly used in classification problems as it has the desirable characteristic of comparing the model's true positive rate with its false positive rate, as opposed to accuracy which only considers the count of true positives. AUC ranges between 0.5 and 1, with 0.5 indicating no improvement in performance from random and 1 indicating perfect performance. AUC can be interpreted as the probability that our model correctly ranks the relative extinction risks of any two randomly selected species [14, 40].

The differences in in-sample predictive performance between the models was visualized in multiple ways: whole data set by model, taxonomic group by model, model performance over time, and model performance by taxonomic groups over time. These comparisons demonstrate the relative and absolute adequacy of the models in describing the dataset they were fit to.

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

corresponds to many (100-1000) individual observations.

See our code repository LINK for full code details. Our code uses "tidyverse" tools such as `dplyr` [50], `purr` [28], and `tidybayes` [32], thus some familiarity with that package ecosystem is necessary to fully comprehend how we've processed our data and results.

3 Results

3.1 Model adequacy

The in-sample model comparisons are for determining their relative adequacy, or a model's ability to represent the data it was fit to. Comparison between the posterior predictive estimates of in-sample AUC for each of the four models demonstrates that, overall, all of the models have approximately equal in-sample performance (Fig. 2). The parameter rich "past and vary" model 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 "past and vary" model is possibly the most adequately performing model, the difference or improvement to performance is minimal at best – all four models have approximately equal in-sample AUC posterior distributions. All of the in-sample AUC estimates from our models are concentrated around an AUC of 0.77 which is interpreted as "fine but not good" performance. It is then hard to conclude that there is one "best" model which we can rely upon.

When the posterior predictive distributions of the in-sample AUC estimates are presented over time, the similarity in adequacy between the models becomes more apparent (Fig. 3). There are few major or obvious differences in model adequacy between the four models.

When the posterior predictive distributions of the in-sample AUC estimates are presented by taxonomic group, some heterogeneity in model adequacy is revealed (Fig. 4). While in all cases the model with the highest average in-sample AUC is the parameter-rich "past and vary" model, the amount of difference between the models varies by taxonomic group in ways not observable from the pooled estimates (Fig. 2). For example, the difference between

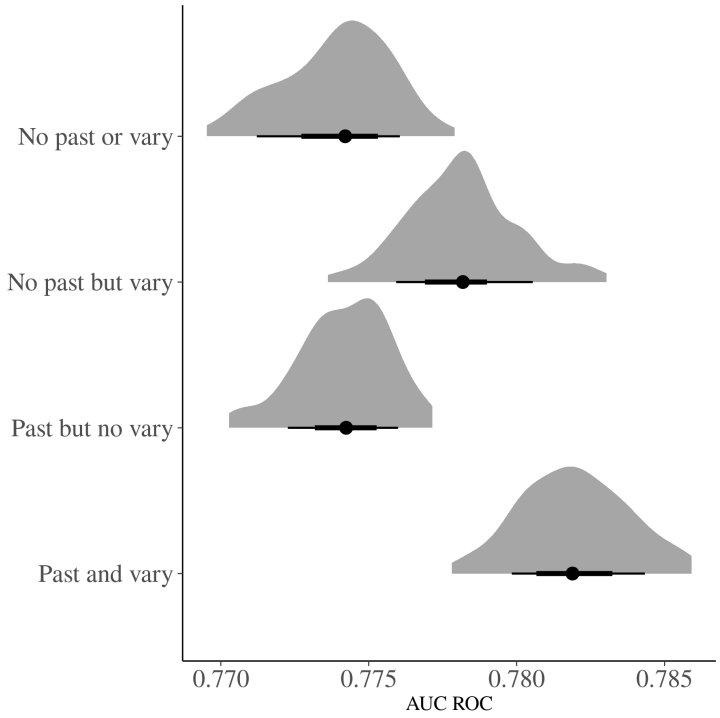

Figure 2: Posterior predictive AUC estimates for each of the four models being compared. These estimates are calculated from each of the models posterior predictive distribution compared to the empirical values. Models with a higher AUC values indicate better performance over models with lower AUC values. AUC is bounded between 0.5 and 1.

the "past and vary" model and the others is more pronouced for Calcareous nannoplankton and Dinoflagellates, and smaller for the Foraminifera and Radiolaria.

For many taxon/model combinations there are one or more time periods where posterior predictive in-sample AUC has a median value less than or equal to 0.5 – AUC value of 0.5 indicates that the model's predictions are no better than random (Fig. 5). However, this pattern is absent for the posterior predictive distribution of Foraminfera and Radiolaria for the "past and vary" model. Additionally, these periods of low model performance are rarer for the posterior predictive distribution of the "past and vary" model for calcareous nannoplankton and Dinoflagellates when compared to the other three models.

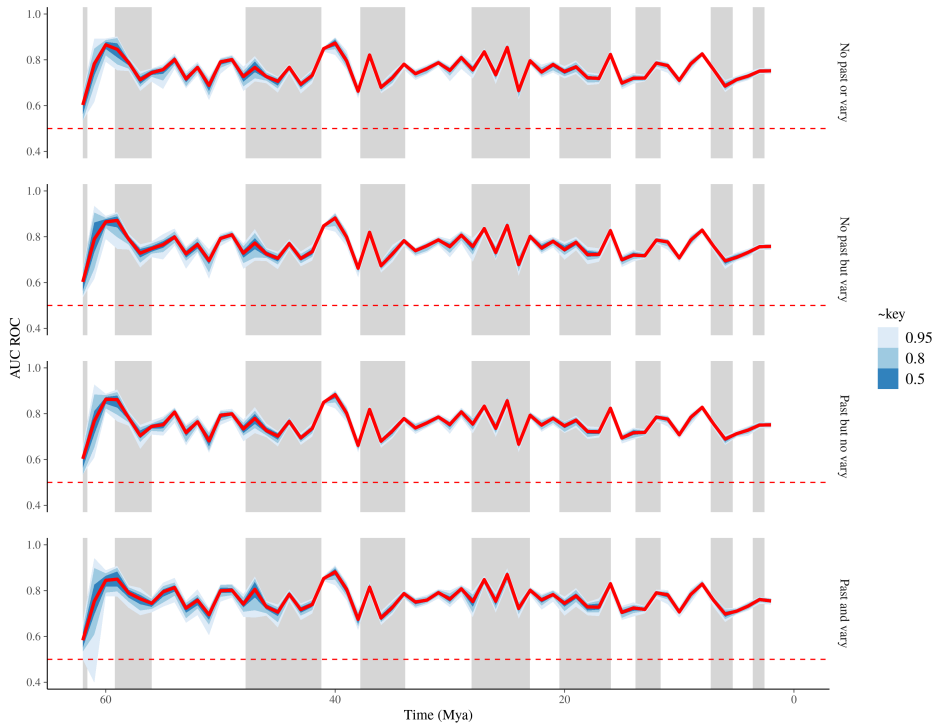

Figure 3: Comparison between the posterior predictive AUC estimates for each of the time intervals for each of the four models. These estimates are reflections of each model's fit to the various time intervals. The red line corresponds to the median AUC value, while the envelopes correspond to multiple credible intervals as indicated in the legend. In all cases, higher AUC values indicate greater predictive performance versus lower AUC values.

3.2 Cross-validation

Expected out-of-sample predictive performance was estimated using five-fold cross-validation, modified for time series data [27]. This procedure yields four posterior (predictive) distributions, each corresponding to AUC values calculated from model-based predictions compared to the extinction state of the hold-out data. These four posterior predictive distributions are pooled to yield a posterior predictive distribution of expected out-of-sample performance

– the resulting distributions tend to be very multimodal due to their very nature being fit to and estimated from different data sets and amounts of data [27]. Additionally, multimodality increases with model complexity (Fig. 6) – this makes sense as the more complex models allow for predictor effects to vary with time, allowing for a greater range in possible parameter values which in turn yield a greater range of posterior predictions.

Comparison between the posterior predictive distributions of expected out-of-sample AUC (Fig. 6) reveals a similar range in plausible values for all models as the in-sample AUC posterior predictive distributions (Fig. 2). Interestingly, the differences between the posterior predictive distributions for the models have decreased. For example, the "past and vary" model not clearly better than either of the "no past or vary" and "past and no vary" models (Fig. 6), which were shown earlier to be obviously worse-performing models based on in-sample performance (Fig. 2). These differences means that the rank order of median out-of-sample AUC is different from the rank order of median in-sample AUC. However, the shapes of the posterior distributions means interepting from the median values is incorrect – the models are effectively indistinguishable in their expected out-of-sample AUC values.

Additionally, the quality of expected out-of-sample performance is not great, with average out-of-sample AUC for each of our models estimated to be between 0.7 and 0.8 which is far from perfect. This result means that we would expect to correctly rank two species in order of most to least likely to go extinct 70-80% of the time. However, this expected out-of-sample performance is approximately the same as the in-sample performance results (Fig. 2), indicating that our models would yield consistent results when generalized to future extinctions.

When the posterior predictive distribution of expected out-of-sample AUC is presented as a time series, the similarity between the models is even more apparent (Fig. 7). While the width of the credible intervals at various time points varies between the models, the overall picture of expected out-of-sample AUC is almost identical when you compare the models – periods of relatively better or worse performance map identically between the time series.

We can also compare expected out-of-sample AUC by taxonomic group for each of the models (Fig. 8). These comparisons reveal a lot about the differences in predictive potential of the taxonomic groups. For example, the posterior predictive distributions of out-of-sample AUC for Foraminifera from

all four models are approximately identical. In contrast, expected out-of-sample AUC for Radiolaria exhibits the same or similar pattern in relative model performance to the pooled comparisons (Fig. 6). Additionally, we can state that our out-of-sample predictions for calcareous nannoplankton and dinoflagellates are not necessarily as precise as our estimates for Foraminifera or even Radiolaria. These results indicate that out-of-sample predictions may be easier for some taxonomic groups than others (e.g. Foraminifera versus Dinoflagellates.

Finally, we can present the posterior predictive distribution of expected out-of-sample AUC over time and taxonomic group for each of the four models (Fig. 9). For each taxonomic group, the time-series of posterior predictive values for each model are broadly congruent.

In the analysis of the posterior predictive distributions of the in-sample AUC values for the four models, we noted that there were time intervals where the models' predictions were no better than random (Fig. 5). This occurrence is generally much rarer for the posterior predictive distribution of out-of-sample AUC values – the major exception to this is Dinoflagellates, which for all four models has at least one time interval where the median the AUC of out-of-sample data were no better random. In contrast, the only other group for which median posterior predictive estimate of out-of-sample AUC reaches 0.5 is calcareous nannoplankton, and then only with the "no past or vary" model.

4 Discussion

The results of this paper are a measure of baseline ability to predict relative differences in extinction risk. By focusing on two aspects of species macroecology – geographic range and global temperature – we developed a rather naive model for estimating the probability of a species extinction based on their occurrence records.

We find that all of our model have an approximate 77% to 79% probability of correctly rank the extinction risk of two randomly selected in-sample observations. Similarly, these models are expected to correctly rank the extinction risk of two randomly selected out-of-sample observations approximately 70%

to 80% of the time. A slight decrease in performance when dealing with out-of-sample observations makes sense: each of the models fit during cross-validation is based on less data than the model fit on the full data (between 1/5th to 4/5ths of the original). The similarly between the in-sample and out-of-sample results indicates that our model is fairly robust to how extinction intensity has changed over the Cenozoic.

The striking aspects of our in-sample results are the statistical differences in AUC between the models, and the incredible narrow band of estimated AUC values. For our in-sample results, while the statistically best model does include the historical covariates and allows all covariate effects to vary over time, its practical difference in performance compared to the other models is virtually negligible. This means that even though one model has a statistically greater AUC, the practical difference between the models is not scientifically significant.

This result is an important reminder about understanding the practical interpretation of our analyses, which can be lost when we do not consider the predictive aspects of our analyses. By focusing on determining which covariates are "significant" and which model is best through simple comparisons means that the practical importance of the results are ignored. For example, in logistic regression a covariate can be considered significant on the log-odds scale but have no practical difference on the probability scale because the range of values is too small as to matter such as when the inverse-logit of the intercept is close to 0 or 1 CITATION.

The success of model is partially driven by the size of our dataset and the structure of our model. Our estimates are based on rather limited information about the taxa themselves, and our model only takes into account some aspects of species geographic range and their rough taxonomic grouping. Instead of relying on large amounts of ecological information to shape individual differences, our model leverages the entire Neptune dataset through a multilevel model to constrain and improve our parameter estimates by sharing information about those parameters across taxa and time.

The reason we were not able to include more biological information in our models about the species studied is that we lack most any life history or ecological information on most marine micro- and nano-plankton. Forams are the exception to this problem – there is life history, ecological, and physiological information for a selection of foram species.

If we want to include this type of information in a predictive model of extinction risk, we would be able to analyze only a subset of the fossil occurrences present the Neptune Database. This means ignoring the vast trove of occurrence information present in the rest of the database.

This presents an interesting conundrum about how to improve upon our results. A simple hypothesis for how to improve upon our results is that if we were to include more biological information in our model, our estimates of species relative extinction risk would be improved. However, if we decrease the amount of data in our model, our results by definition decrease in their quality. For example, compare the results from our models fit using full in-sample dataset, and the results from the cross-validation where the fit to each fold is by definition based on less than full information.)

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

References

[1] J. Alroy. New methods for quantifying macroevolutionary patterns and processes. *Paleobiology*, 26(4):707–733, 2000. ISSN 0094-8373. doi: 10.1666/0094-8373(2000)026⟨0707:NMFQMP⟩2.0.CO;2.

[2] J. Alroy. Fair sampling of taxonomic richness and unbiased estimation of origination and extinction rates. *Quantitative methods in paleobiology. Paleontological Society Papers.*, 16:55–80, 2010. URL `https://www.nceas.ucsb.edu/{~}alroy/pdfs/2010-PSPapers-16-55.pdf{%}5Cnfile:///Users/tmsmiley/Documents/Papers2/Articles/2010/Alroy/Quantitativemethodsinpaleobiology.PaleontologicalSocietyPapers.2010Alroy.pdf{%}5Cnpapers2://publication/uuid/E23F7702-48A8`.

[3] J. Alroy. Accurate and precise estimates of origination and extinction rates. *Paleobiology*, 40(03):374–397, 2014. ISSN 0094-8373. doi:

10.1666/13036. URL https://www.cambridge.org/core/product/identifier/S0094837300001871/type/journal{_}article.

[4] J. Alroy, P. L. Koch, and J. C. Zachos. Global Climate Change and North American Mammalian Evolution. *Paleotological Society*, 26(4): 259–288, 2000. doi: 10.1666/0094-8373(2000)26.

[5] J. Alroy, C. R. Marshall, R. K. Bambach, K. Bezusko, M. Foote, F. T. Fursich, T. A. Hansen, S. M. Holland, L. C. Ivany, D. Jablonski, D. K. Jacobs, D. C. Jones, M. A. Kosnik, S. Lidgard, S. Low, A. I. Miller, P. M. Novack-Gottshall, T. D. Olszewski, M. E. Patzkowsky, D. M. Raup, K. Roy, J. J. Sepkoski, M. G. Sommers, P. J. Wagner, and A. Webber. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences*, 98(11): 6261–6266, 2001. ISSN 0027-8424. doi: 10.1073/pnas.111144698. URL http://www.pnas.org/cgi/doi/10.1073/pnas.111144698.

[6] P.-C. Brkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

[7] P.-C. Brkner. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411, 2018.

[8] 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):1–23, 2011. ISSN 21699291. doi: 10.1029/2011JC007255.

[9] 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):1498–1503, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1519092113.

[10] T. H. Ezard and A. Purvis. Environmental changes define ecological limits to species richness and reveal the mode of macroevolutionary competition. *Ecology Letters*, pages 1–8, 2016. ISSN 14610248. doi: 10.1111/ele.12626.

[11] T. H. 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):745–753, 2013. ISSN 2041210X. doi: 10.1111/2041-210X.12089.

[12] 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):349–351, 2011. ISSN 0036-8075. doi: 10.1126/science.1203060. URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1203060`.

[13] 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):139–142, 2012. ISSN 1744-9561. doi: 10.1098/rsbl.2011.0699. URL `http://rsbl.royalsocietypublishing.org/cgi/doi/10.1098/rsbl.2011.0699`.

[14] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 01678655. doi: 10.1016/j.patrec.2005.10.010.

[15] 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):318–341, 2008. ISSN 0094-8373. doi: 10.1666/07008.1.

[16] S. Finnegan, S. C. Anderson, P. G. Harnik, C. Simpson, D. P. Tittensor, J. E. Byrnes, Z. V. Finkley, D. R. Lindberg, L. H. Liow, A. O'Dea, and J. M. Pandolfi. Paleontological baselines for evaluating extinction risk in the modern oceans. *Science*, 348(6234):567–570, 2015.

[17] M. Foote. Estimating Taxonomic Durations and Preservation Probability. *Paleobiology*, 23(3):278–300, 1997. ISSN 0094-8373. doi: 10.1017/S0094837300019692.

[18] M. Foote. Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis. *Paleobiology*, 27(4):602–630, 2001. ISSN 0094-8373. doi: 10.1666/0094-8373(2001)027⟨0602:ITPOPO⟩2.0.CO;2.

[19] M. Foote. Substrate affinity and diversity dynamics of Paleozoic marine animals. *Paleobiology*, 32(3):345–366, 2006. ISSN 0094-8373. doi: 10.1666/05062.1. URL `http://www.bioone.org/doi/abs/10.1666/05062.1`.

[20] M. Foote. Symmetric waxing and waning of marine invertebrate genera. *Palaeobiology*, 33(4):517–529, 2007. ISSN 0094-8373. doi: 10.1666/06084.1.

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

[22] M. Foote and J. J. Sepkoski. Absolute measures of the completeness of the fossil record. *Nature*, 398(6726):415–417, apr 1999. ISSN 0028-0836. doi: 10.1038/18872.

[23] J. Gabry and B. Goodrich. *rstanarm: Bayesian Applied Regression Modeling via Stan*, 2018. URL `https://CRAN.R-project.org/package=rstanarm`. R package version 2.18.2.

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

[25] 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):13594–13599, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1100572108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1100572108.

[26] 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):4969–4976, 2012. ISSN 0962-8452. doi: 10.1098/rspb.2012.1902. URL `http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2012.1902`.

[27] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. *Bayesian Forecasting and Dynamic Models*, 1:1–694, 2009. ISSN 0172-7397. doi: 10.1007/b94608. URL `http://www.springerlink.com/index/10.1007/b94608`.

[28] L. Henry and H. Wickham. *purrr: Functional Programming Tools*, 2018. URL `https://CRAN.R-project.org/package=purrr`. R package version 0.2.5.

[29] D. Jablonski. Species Selection: Theory and Data. *Annual Review of Ecology, Evolution, and Systematics*, 39(1):501–524, 2008. ISSN 1543-592X. doi: 10.1146/annurev.ecolsys.39.110707.

173510. URL http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.39.110707.173510.

[30] D. Jablonski and G. Hunt. Larval Ecology, Geographic Range, and Species Survivorship in Cretaceous Mollusks: Organismic versus SpeciesLevel Explanations. *The American Naturalist*, 168(4):556–564, 2006. ISSN 0003-0147. doi: 10.1086/507994. URL http://www.journals.uchicago.edu/doi/10.1086/507994.

[31] D. Jablonski and K. Roy. Geographical range and speciation in fossil and living molluscs. *Proceedings of the Royal Society B: Biological Sciences*, 270(1513):401–406, 2003. ISSN 0962-8452. doi: 10.1098/rspb.2002.2243. URL http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2002.2243.

[32] M. Kay. *tidybayes: Tidy Data and Geoms for Bayesian Models*, 2018. URL http://mjskay.github.io/tidybayes/. R package version 1.0.3.

[33] W. Kiessling and Á. T. Kocsis. Adding fossil occupancy trajectories to the assessment of modern extinction risk. *Biology Letters*, 12 (10):20150813, 2016. ISSN 1744-9561. doi: 10.1098/rsbl.2015.0813. URL http://www.ncbi.nlm.nih.gov/pubmed/28120797{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5095184{%}0Ahttp://rsbl.royalsocietypublishing.org/lookup/doi/10.1098/rsbl.2015.0813.

[34] D. Lazarus. Neptune: A marine micropaleontology database. *Mathematical Geology*, 26(7):817–832, 1994. ISSN 08828121. doi: 10.1007/BF02083119.

[35] 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):2745–2752, 2007. ISSN 0962-8452. doi: 10.1098/rspb.2007.1006. URL http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2007.1006.

[36] 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):224–252, 2010. ISSN 0094-8373. doi: 10.1666/08080.1.

[37] L. H. Liow, C. Simpson, F. Bouchard, J. Damuth, B. Hallgrimsson,

G. Hunt, D. W. McShea, J. R. Powell, N. C. Stenseth, M. K. Stoller, and G. Wagner. Pioneering paradigms and magnificent manifestos–Leigh Van Valen's priceless contributions to evolutionary biology. *Evolution; international journal of organic evolution*, 65(4):917–922, 2011. ISSN 15585646. doi: 10.1111/j.1558-5646.2011.01242.x.

[38] 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):569–584, 2012. doi: 10.5061/dryad.8ts3p.

[39] 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):731–734, 1995. ISSN 00917613. doi: 10.1130/0091-7613(1995)023⟨0731:DBSAGE⟩2.3.CO.

[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*, 128:2145–2166, 2002.

[41] 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):10506–10511, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0701257104.

[42] 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):20160202, 2016. ISSN 1744-9561. doi: 10.1098/rsbl.2016.0202. URL `http://rsbl.royalsocietypublishing.org/lookup/doi/10.1098/rsbl.2016.0202`.

[43] S. E. Peters. Environmental determinants of extinction selectivity in the fossil record. *Nature*, 454(7204):626–629, 2008. ISSN 00280836. doi: 10.1038/nature07032.

[44] P. D. Smits. Expected time-invariant effects of biological traits on mammal species duration. *Proceedings of the National Academy of Sciences*, 112(42):13015–13020, 2015. ISSN 0027-8424. doi: 10.1073/pnas.

1510482112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1510482112.

[45] C. Spencer-Cervato. The Cenozoic deep sea microfossil record: explorations of the DSDP/ODP sample set using the Neptune database. *Palaeontologia Electronica*, 2(2):4–286, 1999. ISSN 10948074. URL http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:THE+CENOZOIC+DEEP+SEA+MICROFOSSIL+RECORD+:+EXPLORATIONS+OF+THE+DSDP+/+ODP+SAMPLE+SET+USING+THE+NEPTUNE+DATABASE{#}0.

[46] S. D. Team. Stan Modeling Language Users Guide and Reference Manual, 2017. URL http://mc-stan.org.

[47] G. Tutz and M. Schmid. *Modeling discrete time-to-event data*. Springer International Publishing, 2016. ISBN 978-3-319-28156-8. doi: 10.1007/978-3-319-28158-2.

[48] L. Van Valen. A new evolutionary law. *Evolutionary Theory*, 1:1–30, 1973.

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

[50] H. Wickham, R. Franois, L. Henry, and K. Mller. *dplyr: A Grammar of Data Manipulation*, 2018. URL https://CRAN.R-project.org/package=dplyr. R package version 0.7.8.

[51] J. C. Zachos, G. R. Dickens, and R. E. Zeebe. An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. *Nature*, 451(7176):279–283, jan 2008. ISSN 1476-4687. doi: 10.1038/nature06588.

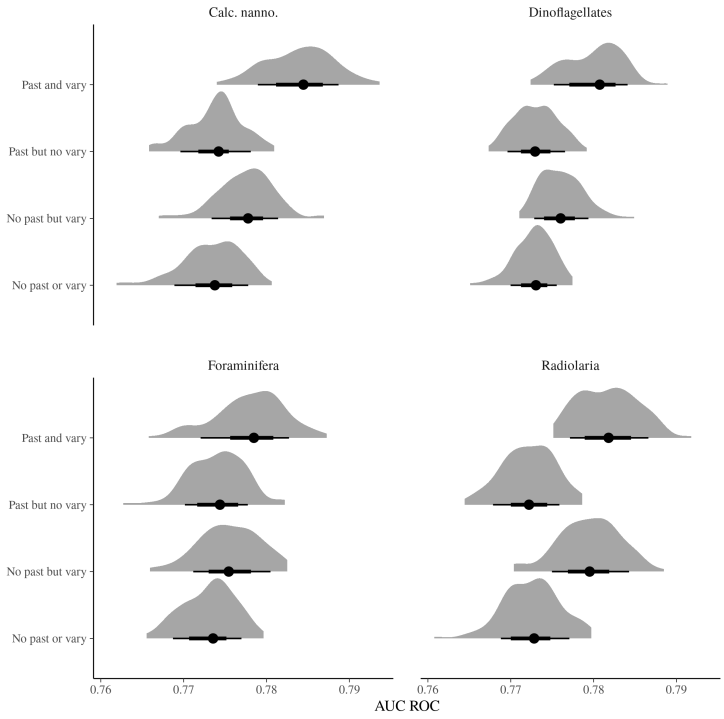

Figure 4: Comparison of posterior predictive AUC estimates for each of the four models, arranged by taxonomic group. These estimates reflect each model's fit to the various taxonomic groups present in this analysis. The densities reflect the posterior distribution of the estimates, and below each density is marked the median AUC value along with the 50% and 80% credible intervals. In all cases, higher AUC values indicate greater predictive performance versus lower AUC values.

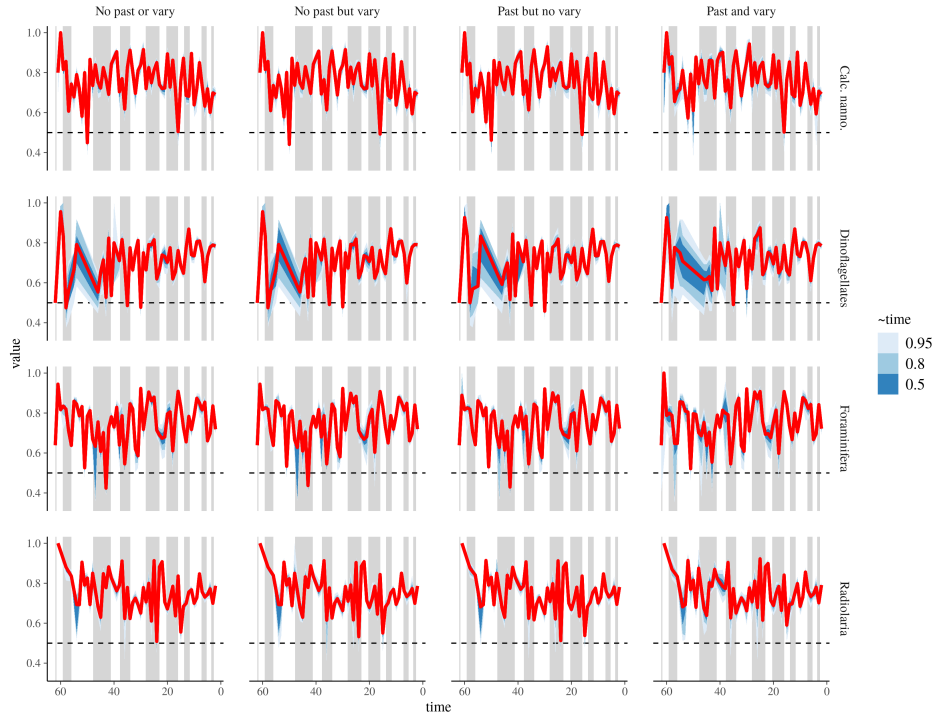

Figure 5: Comparison of posterior predictive AUC estimates for each of the four models, arranged over time and by taxonomic group. These estimates reflect each model's fit to the various taxonomic groups over time. The red line corresponds to the median AUC value, while the envelopes correspond to multiple credible intervals as indicated in the legend. In all cases, higher AUC values indicate greater predictive performance versus lower AUC values.

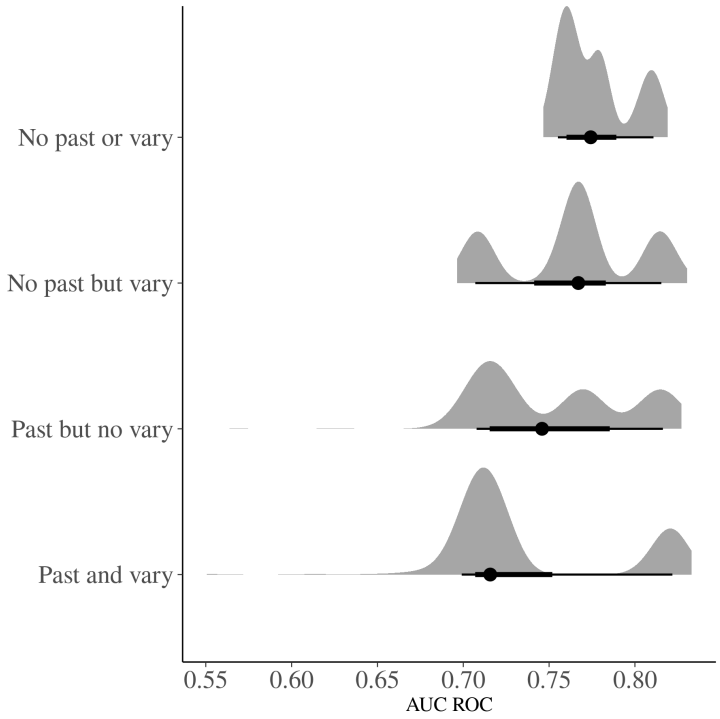

Figure 6: Results from our five-fold cross-validation of the time-series. Each labeled distribution of AUC values correspond to expected out-of-sample performance as estimated from that fold. Each fold represents a section of data being predicted from a model fit to all data before the start of that fold. Given that there are only five folds, performance is measured from predictions for four of the folds.

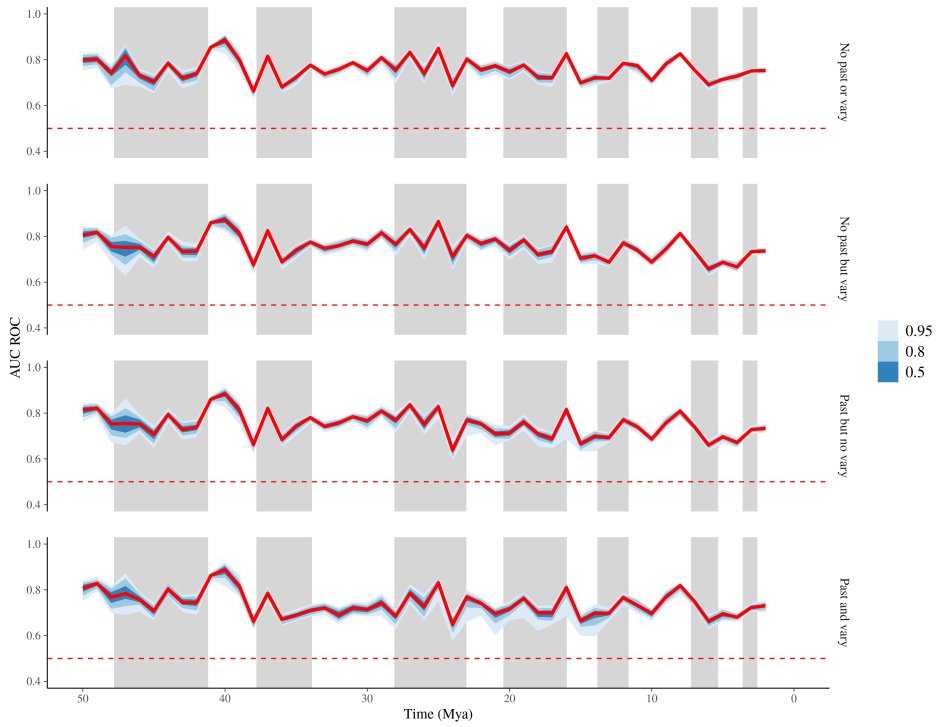

Figure 7: Comparison of out-of-sample AUC values calculated for each of the My intervals for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogentity in performance within and between folds. This presentation decomposes each of the 12-million year folds (Fig. 6) into the predictions made for each of the million-year intervals. The red line corresponds to the median AUC estimate, with the envelopes corresponding to multiple credible intervals as indicated in the legend.

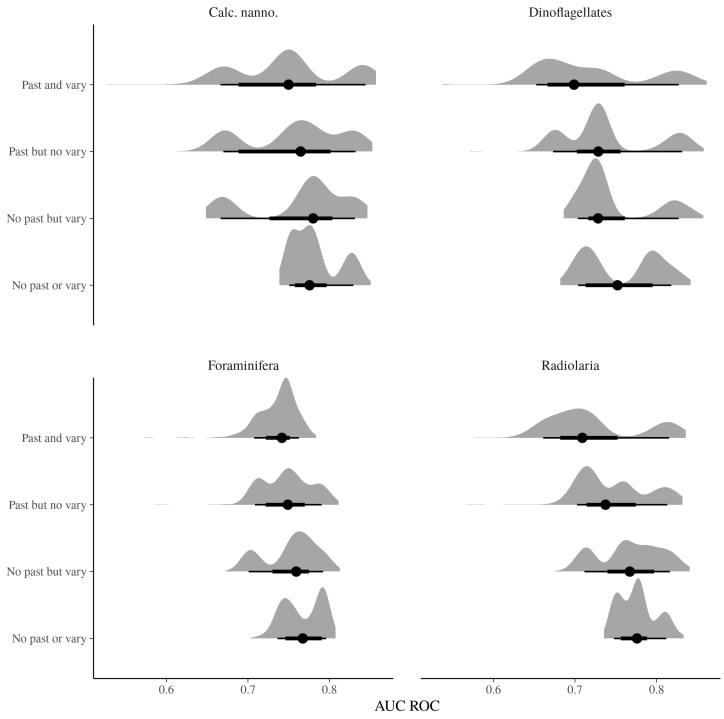

Figure 8: Comparison of out-of-sample AUC values aggregated by taxonomic group for each of the four models. Depicted for each taxon-model combination is an aggregate density of all posterior predictive estimates for each of four folds – cross-validation estimates are commonly multi-modal as each fold presents its own challenges for prediction. Beneath these densities is marked the median estimate along with 50% and 80% credible intervals.

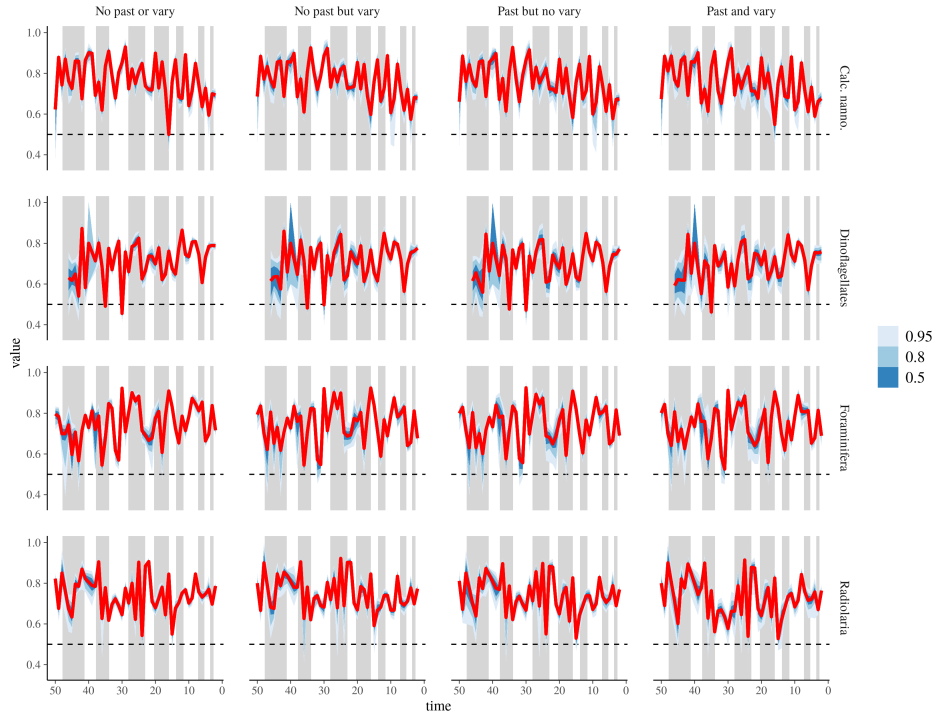

Figure 9: Comparison of out-of-sample AUC values over time as aggregated by taxonomic group for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogentity in performance within and between folds. This presentation decomposes each of the 12-million year folds by each of the taxonomic groups (Fig. 8) into the predictions made for each of the million-year intervals. The red line corresponds to the median AUC estimate, with the envelopes corresponding to multiple credible intervals as indicated in the legend.