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

Smits, Peter  
psmits@berkeley.edu

Finnegan, Seth  
sethf@berkeley.edu

A tenet of conservation paleobiology 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.) For each group we examine how extinction probability varies over time as a function of species age, time of observation, current geographic range, change in geographic range, climate state, and change in climate state. Although the best-performing model includes time-varying effects and historical covariates (change in range and change in climate state) the improvement in predictive power over models with constant effects and no historical covariates is very minor. This model has an approximately 78% median probability of correctly ranking the relative extinction risk any two randomly selected species. It performs nearly as well at out-of-sample prediction (70-80%), implying that determinants of extinction risk have varied only modestly through time. An important caveat is that human impacts may substantially disrupt range-risk dynamics so that the future will be less predictable than it has been in the past.

**Keywords:** conservation, paleobiology, extinction, forecasting

# 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. Such predictions are 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 suggestions that some risk assessments might be improved by incorporating palaeontological data (Finnegan et al., 2015, Kiessling and Kocsis 2016). The fossil record preserves information about the full histories, including ultimate extinction, of thousands of lineages, and this information can help to augment the shorter-term higher-resolution data used to make risk assessments of extant taxa.

Past paleobiological studies of extinction have frequently focused on identifying and measuring the effect of various potential predictors on extinction risk (Harnik 2011; Smits 2015; Peters 2008; Payne and Finnegan 2007; Harnik, Simpson, and Payne 2012; Ezard et al. 2011; Foote 2006) or on how to identify or measure these effects (J. Alroy 2010; Alroy 2014; Alroy et al. 2001; Alroy, Koch, and Zachos 2000; Alroy 2000; Foote 2001).

This focus means that while we have a good understanding of which factors are strong and general determinants of extinction risk, we have less knowledge of how accurate or strong our predictions about the differences in extinction risk are likely to be.

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

Both the intensity (average rate) and the selectivity (difference in risk between taxa) of extinction varies through time (Payne and Finnegan 2007; Payne et al. 2016; Ezard et al. 2011). Given this variation, how accurate are our assessments based on past events likely to be when applied to the future? Putting aside the important question of how human activities will alter the determinants of future extinction risk, we can address this uncertainty by specifically including and modeling the temporal variation in extinction risk across a range of extinction intensities and selectivities.

Numerous studies have established that geographic range is one of the most important determinants of extinction risk in the fossil record, and that geographic range trajectories are highly variable (ARGH, somehow cut the various refs you had in to make this point -can you restore them from an earlier draft?). The degree to which the past can help to predict the future fates of species depends in part on the degree to which species’ geographic range trajectories are deterministic versus Markovian. In the former case, knowledge of the specific past trajectory of a species -whether its range has expanded or contracted over a given timespan- might help to improve assessments of its current risk. In the latter case, only the current range of the species conveys useful information about its current risk, although we can still use prior extinction patterns to augment predictions by evaluating the relative extinction risk of species that had similar ranges in the past. Discriminating among these alternative models is thus very important for determining how best to incorporate fossil data in present risk assessments.

For this exercise we chose to analyze one of the best-sampled and studied fossil records – the Cenozoic record of skeletonized marine planktonic microorganisms (Foraminifera, Radiolaria, Diatoms, and Coccolithophores). These data are readily available through the Neptune database, an online repository of species occurrences obtained through the Deep Sea Drilling Program and the Ocean Drilling Project (Lazarus 1994; Spencer-Cervato 1999). This database provides abundant samples in space and time, a high degree of temporal resolution for the entirety of the Cenozoic, and has an internally consistent taxonomic identification strategy – as close to ideal data for this analysis as possible. The Neptune database records multiple phyla-scale taxonomic groups for over 60 million years, with incredible temporal resolution supported by the various age-models of the deep-sea cores from which the occurrences are recorded. Analyzing patterns of extinction and global occurrence at fine temporal scales means we can better elucidate how well we can predict species extinction at human-relevant scales.

# Materials and Methods

## Data Specifications

We analyzed microfossil occurrence information which was downloaded from the Neptune Database <http://www.nsb-mfn-berlin.de/nannotax> (Lazarus 1994; Spencer-Cervato 1999). All occurrence information was downloaded for calcareous nannoplankton, diatoms, foraminifera, and radiolarians.This dataset of occurrences was then filtered to just those species which have their first occurrence at most 63 Mya. This choice means that our analysis avoids those taxa which survived the K/Pg boundary, those taxa which arose just after the K/Pg boundary, and means that our occurrence histories line up with the temperature time-series which was used as a predictor of extinction as discussed below.

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

Temperature data used as covariates in this analysis are based on Magnesium/Calcium isotope ratios sourced from Cramer et al. (2011). 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. (2011), provides temperature estimates for every 0.1 My from 0 to 63 Mya. We binned these estimates into 1 My intervals as we did with the fossil occurrences. The temperature estimate for each 1 My interval was calculated as the mean of all estimates within that interval.

See Section [sec:data\_desc] for a further explanation on how observations were temporally binned and how our covariates were standardized and transformed prior to analysis.

## Model Specifications

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

Importantly, we include species geologic age at time of observation as a non-nested varying intercept term. This factor may or may not contribute to differences in species extinction risk over time (Smits 2015; Finnegan, Payne, and Wang 2008; Ezard et al. 2012; Van Valen 1973; Liow et al. 2011; Crampton et al. 2016), but its inclusion in our model is critical to its nature as a survival model (Tutz and Schmid 2016). Again, see Section [sec:model\_desc] for how this term was specified as part of our model.

Models and their definitions

|  |  |  |  |
| --- | --- | --- | --- |
| Code | Description | Covariates | R Formula Syntax |
| C | Constant effects, no historical cov. |  |  |
| V | Varying effects, no historical cov. |  |  |
| CP | Constant effects, historical cov. |  |  |
| VP | Varying effects, historical cov. |  |  |

See Equation [eq:model] for full statistical model definition.

Species observation where 1 if time of last observation, otherwise 0.

Species geographic range in log km. Mean centered, scaled to sd = 1.

Global temperature in degrees C. Mean centered, scaled to sd = 1.

Species are at observation in millions of years.

Taxonomic group of species (i.e. Foraminifera, Diatoms, Radiolarians, Calcareous nannoplankton).

Change in geographic range since last observation.

Temperature at previous observation.

[tab:model\_def]

## In-sample and out-of-sample forecasting

We are interested in model performance into two contexts: in-sample and out-of-sample predictive performance i.e. forecasting.

In-sample forecasting means we are estimating how well our model predicts extinction probability for observations that that model was fit to. This is a posterior predictive check in that we are comparing the posterior predictive distribution to our observed data.

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). These in-sample measures, however, are not necessarily good estimates of the model’s ability to predict data from the future. 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 time points. Keep in mind, however, that each time point corresponds to many (100-1000) individual observations.

Here is a description of how k-fold cross-validation for time series precedes (Arlot and Celisse 2009; Bergmeir, Hyndman, and Koo 2018). The model is fit first 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 first and second golds, and the posterior estimates of that fit are used to forecast the extinction probability of the third fold. This process continues with the model fit to the first three folds used to forecast extinction probabilities for the fourth fold, and then the model fit to the first four folds being used to forecast the fifth fold.

Relative and absolute model adequacy of the four variant models was compared using the area under the receiver operating characteristic curve or AUC (Fawcett 2006; Mason and Graham 2002). 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 (Fawcett 2006; Mason and Graham 2002). AUC values of approximately 0.8 or greater can be considered “good” (Tang, He, and Tu 2012), which means that values between 0.7 and 0.8 could then be considered “fair.”

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 (Wickham et al. 2018), and purr (Henry and Wickham 2018), and tidybayes (Kay 2018). Additionally, all of our models were written using the brms (Bürkner 2017; Bürkner 2018) R package, which implements Stan-based Bayesian models which are fit via Hamiltonian Monte Carlo (Stan Development Team 2017).

# Results

0.45 In-sample ([fig:auc\_hist]) and out-of-sample ([fig:fold\_auc]) AUC estimates for each of our four models. These estimates are calculated from the models posterior predictive distribution or from predictions made to new data, respectively. Models with a higher AUC values indicate better performance over models with lower AUC values. AUC is bounded between 0.5 and 1. See Table [tab:model\_def] for a description of each of the four models.

0.45 In-sample ([fig:auc\_hist]) and out-of-sample ([fig:fold\_auc]) AUC estimates for each of our four models. These estimates are calculated from the models posterior predictive distribution or from predictions made to new data, respectively. Models with a higher AUC values indicate better performance over models with lower AUC values. AUC is bounded between 0.5 and 1. See Table [tab:model\_def] for a description of each of the four models.

## In-sample forecasting 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 the models have approximately equal in-sample performance (Fig. [fig:auc\_hist]). The parameter rich model VP has the greatest median in-sample AUC when compared to the other three models, but there is substantial overlap in their posterior distributions. Additionally, while our parameter rich model VP is possibly the most adequately performing model, the difference or improvement to performance is minimal at best – all four models have approximately equal in-sample AUC posterior distributions. All the in-sample AUC estimates from our models are concentrated on an AUC value of 0.77. It is therefore hard to conclude that there is one “best” model which we can rely upon as they are all nearly functional equivalent.

There are one or more time periods where posterior predictive in-sample AUC has a median value less than or equal to 0.5 for the various taxon/model combinations (Fig. [fig:auc\_taxon\_time]). However, this pattern is absent for the posterior predictive distribution of Foraminfera and Radiolaria for the VP model. Additionally, these periods of low model performance are rarer for the posterior predictive distribution of the VP model for calcareous nannoplankton and Dinoflagellates when compared to the other three models.

Comparison of in-sample forecasting performance measured by AUC for each of the four models, arranged over time and by taxonomic group. These estimates reflect each model’s fit to the various taxonomic groups over time. The black line corresponds to the median AUC value, while the envelopes correspond to multiple credible intervals as indicated in the legend. In all cases, higher AUC values indicate greater predictive performance versus lower AUC values. See Table [tab:model\_def] for a description of each of the four models.

Comparison of in-sample forecasting performance measured by AUC for each of the four models, arranged over time and by taxonomic group. These estimates reflect each model’s fit to the various taxonomic groups over time. The black line corresponds to the median AUC value, while the envelopes correspond to multiple credible intervals as indicated in the legend. In all cases, higher AUC values indicate greater predictive performance versus lower AUC values. See Table [tab:model\_def] for a description of each of the four models.

## Out-of-sample forecasting performance

Expected out-of-sample forecasting performance was estimated using five-fold cross-validation for time series (Arlot and Celisse 2009; Bergmeir, Hyndman, and Koo 2018). The resulting distribution when all folds are combined is very multimodal by their very nature of being fit to and estimated from different data sets and amounts of data (Hastie, Tibshirani, and Friedman 2009). We find the visual degree of this multimodality increases with model complexity (Fig. [fig:fold\_auc]), most likely because the complex models allow for predictor effects to vary with time which allows for a greater range in possible parameter values which in turn yield a greater range of posterior predictions.

Comparison between the in-sample forecasting performance distributions of expected out-of-sample AUC reveals a similar range in plausible values for all models (Fig. [fig:auc\_hist], [fig:fold\_auc]). Interestingly, the differences in the posterior predictive distributions of AUC for the models have decreased. For example, model VP not clearly better than either models V or CP (Fig. [fig:fold\_auc]), which were shown earlier to be obviously worse-performing models based on in-sample performance (Fig. [fig:auc\_hist]). These differences mean 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 interpretations based only on the median values is incorrect – the models are effectively indistinguishable in their out-of-sample forecast performance.

We can also compare the posterior predictive distribution of expected out-of-sample AUC over time and taxonomic group for each of the four models (Fig. [fig:fold\_auc\_taxon\_time]). The time-series of posterior predictive values for each model are broadly congruent between each taxonomic group.

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

Comparison of out-of-sample AUC values over time as aggregated by taxonomic group for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogeneity in performance within and between folds. This presentation decomposes each of the 12-million year folds by each of the taxonomic groups into the predictions made for each of the million-year intervals. The black line corresponds to the median AUC estimate, with the envelopes corresponding to multiple credible intervals as indicated in the legend. See Table [tab:model\_def] for a description of each of the four models.

Comparison of out-of-sample AUC values over time as aggregated by taxonomic group for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogeneity in performance within and between folds. This presentation decomposes each of the 12-million year folds by each of the taxonomic groups into the predictions made for each of the million-year intervals. The black line corresponds to the median AUC estimate, with the envelopes corresponding to multiple credible intervals as indicated in the legend. See Table [tab:model\_def] for a description of each of the four models.

We quantitatively compared 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. [fig:oos\_ins\_diff]). This calculation means that differences close to 0 indicate complete congruence between 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, quantifying the broad congruence that the visual comparison of the two types of AUC time series suggest (Fig. [fig:auc\_taxon\_time], [fig:fold\_auc\_taxon\_time]). 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. [fig:oos\_ins\_diff]).

Comparison between our out-of-sample forecasts and in-sample forecasts for all models. This value is calculated as the values presented in Figure [fig:fold\_auc\_taxon\_time] minus those values presented in Figure [fig:auc\_taxon\_time]. 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 [tab:model\_def] for a description of each of the four models.

Comparison between our out-of-sample forecasts and in-sample forecasts for all models. This value is calculated as the values presented in Figure [fig:fold\_auc\_taxon\_time] minus those values presented in Figure [fig:auc\_taxon\_time]. 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 [tab:model\_def] for a description of each of the four models.

# Discussion

The results of this paper set out our baseline ability to predict relative differences in extinction risk. We find that all of our model have an approximate 77% to 79% probability of correctly rank the extinction risk of two randomly selected in-sample observations. Similarly, these models are expected to correctly forecast the rank order in 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). Additionally, a potential decrease in precision when forecasting the future of extinction risk is almost to be expected as the future is not necessarily the same as the past. However, the similarly between the in-sample and out-of-sample results indicates that our model is fairly robust to how extinction intensity has changed over the Cenozoic.

One of the most striking results of this analysis is that the in-sample and out-of-sample forecasts made with our models demonstrate that while models where the historical covariates are predictors of species extinction have better in-sample performance (Fig. [fig:auc\_hist], [fig:auc\_taxon\_time]), all of them forecast the extinction risk of out-of-sample data with a similar degree of success (Fig. [fig:fold\_auc], [fig:fold\_auc\_taxon\_time]).

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

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

The principal reason we were not able to include more biological information in our models is that we lack almost any life history or ecological information on most marine micro- and nanno-plankton. Forams are the exception to this problem – there is life history, ecological, and physiological information for a selection of foram species (Ezard et al. 2011). However, not this information does not exist all foram species. If we want to include this type of information in a predictive model of extinction risk, we would be able to analyze only a single taxonomic subset of the fossil occurrences present the Neptune Database and then only a limited selection of those species. This means ignoring the majority of occurrence information present in the Neptune database.

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

# Supplement to Materials and Methods

## Data Specifications

### Binning fossil occurrences

Because the estimated ages of each occurrence is a product of core-specific age-models and can be overly precise, the hope is that by binning the data this smooths over the between-core heterogeneity and thus homogenizes our disparate data sources. The occurrence histories of each species were then given binary codes used to model the presence or extinction of those species. For every occurrence of a species, except the last, that species is considered to have survived and was marked with a 0. The last occurrence of that species is considered the bin in which the taxon has gone extinct – and is assigned a 1. This protocol means that we are reading the fossil record “as written,” a practice that is potentially dangerous as it is an overconfident statement of preservation and may be shortening the actual duration of that species (J Alroy 2010; Alroy 2000; Alroy 2014; Foote 1997; Foote and Sepkoski 1999; Foote 2001; Foote and Raup 1996; Lloyd et al. 2012; Marshall 1995; Wang and Marshall 2016). However, this practice is common with marine microfossil data due to their exceptional preservation rate (Ezard, Thomas, and Purvis 2013; Ezard and Purvis 2016; Ezard et al. 2011; Liow et al. 2010). 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.

### Covariate transforms and standardizations

Prior to analysis, geographic range was then log-plus-one transformed and standardized by mean-centering the data and then dividing by the standard deviation of the distribution of geographic ranges. This standardization means that a regression coefficient associated with this covariate describes the change in extinction probability per change in standard deviation of geographic range, that coefficients associated with similarly standardized covariates will be directly comparable in magnitude, and that the intercept term corresponds to the expected value of the outcome at when geographic range is its average value (Gelman and Hill 2006). Change in geographic range between observations was measured from the standardized geographic range values and was not standardized separately.

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

## 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 -th interval given that it survived up until and the relevant covariate information for all 1 My intervals (Tutz and Schmid 2016). For the discrete time intervals , extinction is defined as . The discrete time hazard function is defined as

The hazard function (Eq. [eq:hazard]) is easily reparameterized as a logistic regression by defining that where is a logit inverse-link function and is the probability of a taxon going extinction during interval (Tutz and Schmid 2016). 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 (Gelman and Hill 2006).

Our covariates matrix is a matrix where is the total number of observations and is the total number of covariates. The first column of is entirely 1’s as it corresponds to the intercept term in the regression model. The next two columns of are two aspects of geographic range as continuous covariates: geographic range during interval , and the difference between the geographic range at and . Change in geographic range was calculated from the transformed and standardized geographic range values; this means that change in geographic range is in units of changes in standard deviations. The final two columns are two aspects of global temperature: mean temperature during interval , and the lag of mean temperature (i.e. mean temperature during interval .) 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 – a by matrix, were is the number of time temporal intervals and is the number of phyla. The elements of this matrix, the regression coefficients, are themselves modeled as being multivariate normally distributed with vector of means describing the average intercept and regression coefficient estimates of each Coefficients for each phylum . These phylum averages are themselves modeled as multivariate normally distributed with mean vector describing the overall average regression coefficients, including the intercept. has length and is ordered intercept, range coefficient, change in range coefficient, temperature coefficient, temperature lag coefficient.

The effect of species age on the log-odds of species extinction is modeled as a non-nested random intercept (Tutz and Schmid 2016). This term describes how the log-odds of extinction varies along a species duration, and how this effect can differ between the phyla. is a by matrix, where is the age at observation of a species and is its phylum. is modeled as following a multivariate normal distribution with phylum means being the vector and covariance matrix . The covariation between the elements of vector are modeled as a multivariate normal distribution with a mean vector of all 0s and covariance matrix .

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. ) 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 (Stan Development Team 2017).

In total, our model can be expressed as:

with indexing the observation and bracket subscripts referencing the class of the th observation where is the time of the -th observation, is the phylum of the -th observation, and is the age of the -th observation.

## Model Parameter Estimation

We implemented our model (Eq. [eq:model] using the

rstanarm

package for the R programming language (Stan Development Team 2017). This package provides an interface to the Stan probabilistic programming language for writing hierarchical/mixed-effects models in native R. Posterior estimates were obtained through Hamiltonian Monte Carlo, using 2000 steps divided equally between warm-up and sampling. In order to prevent divergent transitions adapt delta was increased to 0.9999; all other HMC/NUTS sampling parameters were kept at the defaults for rstanarm 2.18.2 (Gabry and Goodrich 2018).

An implementation of our full model in

rstanarm

, given a data.frame of all necessary data in a data.frame called “data”, is coded as:

form <- event ~ range + range\_diff + temp + temp\_lag +   
 (1 + range + range\_diff + temp + temp\_lag | mybin/phylum) +   
 (1 | age/phylum),   
stan\_glmer(formula = form,  
 data = data,   
 family = 'binomial',  
 prior = normal(c(-1, 0, 0, 0), rep(1, 4), autoscale = FALSE),   
 prior\_intercept = normal(-2, 5, autoscale = FALSE),   
 prior\_aux = cauchy(0, 1, autoscale = FALSE),   
 chains = 4,  
 thin = 4,  
 adapt\_delta = 0.999999)

Similarly, our full model can also be implemented using the

brms

Stan interface (Bürkner 2017; Bürkner 2018) as:

priors <- c(set\_prior('normal(-2, 5)', class= 'Intercept'),  
 set\_prior('normal(0, 1)', class = 'b'),  
 set\_prior('normal(-1, 1)', class = 'b', coef = 'range'),  
 set\_prior('cauchy(0, 1)', class = 'sd'),  
 set\_prior('lkj(1)', class = 'cor'))  
form <- bf(event ~ range + range\_diff + temp + temp\_lag +  
 (1 + range + range\_diff + temp + temp\_lag | mybin/phylum) +  
 (1 | age/phylum))  
brmfit <- brm(formula = form,  
 data = data,   
 family = bernoulli(),   
 prior = priors,  
 chains = 4,   
 thin = 4,  
 control = list(adapt\_delta = 0.999999)

Posterior convergence was determined using the general and HMC-specific diagnostic criteria: scale reduction factor (; target ), effective sample size (eff; target value eff/steps ), 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 ). For further explanation of these diagnostic criteria, see the Stan Manual (Stan Development Team 2017).

Alroy, J. 2010. “Fair sampling of taxonomic richness and unbiased estimation of origination and extinction rates.” *Quantitative Methods in Paleobiology. Paleontological Society Papers.* 16: 55–80. [https://www.nceas.ucsb.edu/{~}alroy/pdfs/2010-PSPapers-16-55.pdf{\%}5Cnfile:///Users/tmsmiley/Documents/Papers2/Articles/2010/Alroy/Quantitative methods in paleobiology. Paleontological Society Papers. 2010 Alroy.pdf{\%}5Cnpapers2://publication/uuid/E23F7702-48A8](https://www.nceas.ucsb.edu/%7b~%7dalroy/pdfs/2010-PSPapers-16-55.pdf%7b\%25%7d5Cnfile:///Users/tmsmiley/Documents/Papers2/Articles/2010/Alroy/Quantitative%20methods%20in%20paleobiology.%20Paleontological%20Society%20Papers.%202010%20Alroy.pdf%7b\%25%7d5Cnpapers2://publication/uuid/E23F7702-48A8).

Alroy, J. 2010. “The Shifting Balance of Diversity Among Major Marine Animal Groups.” *Science* 329 (5996): 1191–4. doi:[10.1126/science.1189910](https://doi.org/10.1126/science.1189910).

Alroy, John. 2000. “New methods for quantifying macroevolutionary patterns and processes.” *Paleobiology* 26 (4): 707–33. doi:[10.1666/0094-8373(2000)026<0707:NMFQMP>2.0.CO;2](https://doi.org/10.1666/0094-8373(2000)026%3c0707:NMFQMP%3e2.0.CO;2).

———. 2014. “Accurate and precise estimates of origination and extinction rates.” *Paleobiology* 40 (03): 374–97. doi:[10.1666/13036](https://doi.org/10.1666/13036).

Alroy, John, Paul L Koch, and James C Zachos. 2000. “Global Climate Change and North American Mammalian Evolution.” *Paleotological Society* 26 (4): 259–88. doi:[10.1666/0094-8373(2000)26](https://doi.org/10.1666/0094-8373(2000)26).

Alroy, John, C R Marshall, Richard K. Bambach, K Bezusko, Mike. Foote, F. T. Fursich, T A Hansen, et al. 2001. “Effects of sampling standardization on estimates of Phanerozoic marine diversification.” *Proceedings of the National Academy of Sciences* 98 (11): 6261–6. doi:[10.1073/pnas.111144698](https://doi.org/10.1073/pnas.111144698).

Arlot, Sylvain, and Alain Celisse. 2009. “A survey of cross-validation procedures for model selection” 4: 40–79. doi:[10.1214/09-SS054](https://doi.org/10.1214/09-SS054).

Bergmeir, Christoph, Rob J. Hyndman, and Bonsoo Koo. 2018. “A note on the validity of cross-validation for evaluating autoregressive time series prediction.” *Computational Statistics and Data Analysis* 120. Elsevier B.V.: 70–83. doi:[10.1016/j.csda.2017.11.003](https://doi.org/10.1016/j.csda.2017.11.003).

Bürkner, Paul-Christian. 2017. “brms: An R Package for Bayesian Multilevel Models Using Stan.” *Journal of Statistical Software* 80 (1): 1–28. doi:[10.18637/jss.v080.i01](https://doi.org/10.18637/jss.v080.i01).

———. 2018. “Advanced Bayesian Multilevel Modeling with the R Package brms.” *The R Journal* 10 (1): 395–411.

Cramer, B. S., K. G. Miller, P. J. Barrett, and J. D. Wright. 2011. “Late Cretaceous-Neogene trends in deep ocean temperature and continental ice volume: Reconciling records of benthic foraminiferal geochemistry (18O and Mg/Ca) with sea level history.” *Journal of Geophysical Research: Oceans* 116 (12): 1–23. doi:[10.1029/2011JC007255](https://doi.org/10.1029/2011JC007255).

Crampton, James S, Roger A Cooper, Peter M Sadler, and Michael Foote. 2016. “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. doi:[10.1073/pnas.1519092113](https://doi.org/10.1073/pnas.1519092113).

Ezard, T. H. G., P. N. Pearson, T. Aze, and Andy Purvis. 2012. “The meaning of birth and death (in macroevolutionary birth-death models).” *Biology Letters* 8 (1): 139–42. doi:[10.1098/rsbl.2011.0699](https://doi.org/10.1098/rsbl.2011.0699).

Ezard, Thomas H G, and Andy Purvis. 2016. “Environmental changes define ecological limits to species richness and reveal the mode of macroevolutionary competition.” *Ecology Letters*, 1–8. doi:[10.1111/ele.12626](https://doi.org/10.1111/ele.12626).

Ezard, Thomas H G, Tracy Aze, Paul N Pearson, and Andy Purvis. 2011. “Interplay Between Changing Climate and Species’ Ecology Drives Macroevolutionary Dynamics.” *Science* 332 (6027): 349–51. doi:[10.1126/science.1203060](https://doi.org/10.1126/science.1203060).

Ezard, Thomas H G, Gavin H Thomas, and Andy Purvis. 2013. “Inclusion of a near-complete fossil record reveals speciation-related molecular evolution.” *Methods in Ecology and Evolution* 4 (8): 745–53. doi:[10.1111/2041-210X.12089](https://doi.org/10.1111/2041-210X.12089).

Fawcett, Tom. 2006. “An introduction to ROC analysis.” *Pattern Recognition Letters* 27 (8): 861–74. doi:[10.1016/j.patrec.2005.10.010](https://doi.org/10.1016/j.patrec.2005.10.010).

Finnegan, S, J L Payne, and S C Wang. 2008. “The Red Queen revisited: reevaluating the age selectivity of Phanerozoic marine genus extinctions.” *Paleobiology* 34 (3): 318–41. doi:[10.1666/07008.1](https://doi.org/10.1666/07008.1).

Finnegan, Seth, Sean C Anderson, Paul G Harnik, Carl Simpson, Derek P Tittensor, Jarrett E Byrnes, Zoe V Finkley, et al. 2015. “Paleontological baselines for evaluating extinction risk in the modern oceans.” *Science* 348 (6234): 567–70.

Foote, M. 1997. “Estimating Taxonomic Durations and Preservation Probability.” *Paleobiology* 23 (3): 278–300. doi:[10.1017/S0094837300019692](https://doi.org/10.1017/S0094837300019692).

Foote, M, and J J Sepkoski. 1999. “Absolute measures of the completeness of the fossil record.” *Nature* 398 (6726): 415–17. doi:[10.1038/18872](https://doi.org/10.1038/18872).

Foote, Michael. 2001. “Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis.” *Paleobiology* 27 (4): 602–30. doi:[10.1666/0094-8373(2001)027<0602:ITPOPO>2.0.CO;2](https://doi.org/10.1666/0094-8373(2001)027%3c0602:ITPOPO%3e2.0.CO;2).

———. 2006. “Substrate affinity and diversity dynamics of Paleozoic marine animals.” *Paleobiology* 32 (3): 345–66. doi:[10.1666/05062.1](https://doi.org/10.1666/05062.1).

———. 2007. “Symmetric waxing and waning of marine invertebrate genera.” *Palaeobiology* 33 (4): 517–29. doi:[10.1666/06084.1](https://doi.org/10.1666/06084.1).

Foote, Michael, and D M Raup. 1996. “Fossil preservation and the stratigraphic ranges of taxa.” *Paleobiology* 22 (2): 121–40.

Gabry, Jonah, and Ben Goodrich. 2018. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. <https://CRAN.R-project.org/package=rstanarm>.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Harnik, P. G., C. Simpson, and J. L. Payne. 2012. “Long-term differences in extinction risk among the seven forms of rarity.” *Proceedings of the Royal Society B: Biological Sciences* 279 (1749): 4969–76. doi:[10.1098/rspb.2012.1902](https://doi.org/10.1098/rspb.2012.1902).

Harnik, Paul G. 2011. “Direct and indirect effects of biological factors on extinction risk in fossil bivalves.” *Proceedings of the National Academy of Science* 108 (33): 13594–9. doi:[10.1073/pnas.1100572108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1100572108](https://doi.org/10.1073/pnas.1100572108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1100572108).

Harnik, Paul G., Heike K. Lotze, Sean C. Anderson, Zoe V. Finkel, Seth Finnegan, David R. Lindberg, Lee Hsiang Liow, et al. 2012. “Extinctions in ancient and modern seas.” *Trends in Ecology and Evolution* 27 (11). Elsevier Ltd: 608–17. doi:[10.1016/j.tree.2012.07.010](https://doi.org/10.1016/j.tree.2012.07.010).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. New York, NY: Springer. doi:[10.1007/b94608](https://doi.org/10.1007/b94608).

Henry, Lionel, and Hadley Wickham. 2018. *Purrr: Functional Programming Tools*. <https://CRAN.R-project.org/package=purrr>.

Jablonski, D., and K. Roy. 2003. “Geographical range and speciation in fossil and living molluscs.” *Proceedings of the Royal Society B: Biological Sciences* 270 (1513): 401–6. doi:[10.1098/rspb.2002.2243](https://doi.org/10.1098/rspb.2002.2243).

Jablonski, David. 2008. “Species Selection: Theory and Data.” *Annual Review of Ecology, Evolution, and Systematics* 39 (1): 501–24. doi:[10.1146/annurev.ecolsys.39.110707.173510](https://doi.org/10.1146/annurev.ecolsys.39.110707.173510).

Jablonski, David, and Gene Hunt. 2006. “Larval Ecology, Geographic Range, and Species Survivorship in Cretaceous Mollusks: Organismic versus Species‐Level Explanations.” *The American Naturalist* 168 (4): 556–64. doi:[10.1086/507994](https://doi.org/10.1086/507994).

Kay, Matthew. 2018. *tidybayes: Tidy Data and Geoms for Bayesian Models*. doi:[10.5281/zenodo.1308151](https://doi.org/10.5281/zenodo.1308151).

Kiessling, Wolfgang, and Ádám T Kocsis. 2016. “Adding fossil occupancy trajectories to the assessment of modern extinction risk.” *Biology Letters* 12 (10): 20150813. doi:[10.1098/rsbl.2015.0813](https://doi.org/10.1098/rsbl.2015.0813).

Lazarus, Dave. 1994. “Neptune: A marine micropaleontology database.” *Mathematical Geology* 26 (7): 817–32. doi:[10.1007/BF02083119](https://doi.org/10.1007/BF02083119).

Liow, L. H., and N. C. Stenseth. 2007. “The rise and fall of species: implications for macroevolutionary and macroecological studies.” *Proceedings of the Royal Society B: Biological Sciences* 274 (1626): 2745–52. doi:[10.1098/rspb.2007.1006](https://doi.org/10.1098/rspb.2007.1006).

Liow, Lee Hsiang, Carl Simpson, Frédéric Bouchard, John Damuth, Benedikt Hallgrimsson, Gene Hunt, Dan W. McShea, et al. 2011. “Pioneering paradigms and magnificent manifestos–Leigh Van Valen’s priceless contributions to evolutionary biology.” *Evolution; International Journal of Organic Evolution* 65 (4): 917–22. doi:[10.1111/j.1558-5646.2011.01242.x](https://doi.org/10.1111/j.1558-5646.2011.01242.x).

Liow, Lee Hsiang, Hans Julius Skaug, Torbjørn Ergon, and Tore Schweder. 2010. “Global occurrence trajectories of microfossils: environmental volatility and the rise and fall of individual species.” *Paleobiology* 36 (2): 224–52. doi:[10.1666/08080.1](https://doi.org/10.1666/08080.1).

Lloyd, G T, P N Pearson, J R Young, and A B Smith. 2012. “Sampling bias and the fossil record of planktonic foraminifera on land and in the deep sea.” *Paleobiology* 38 (4): 569–84. doi:[10.5061/dryad.8ts3p](https://doi.org/10.5061/dryad.8ts3p).

Marshall, C. R. 1995. “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): 731–34. doi:[10.1130/0091-7613(1995)023<0731:DBSAGE>2.3.CO](https://doi.org/10.1130/0091-7613(1995)023%3c0731:DBSAGE%3e2.3.CO).

Mason, S J, and NE Graham. 2002. “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: 2145–66.

Payne, Jonathan L, and Seth Finnegan. 2007. “The effect of geographic range on extinction risk during background and mass extinction.” *Proceedings of the National Academy of Sciences* 104 (25): 10506–11. doi:[10.1073/pnas.0701257104](https://doi.org/10.1073/pnas.0701257104).

Payne, Jonathan L., Andrew M. Bush, Ellen T. Chang, Noel A. Heim, Matthew L. Knope, and Sara B. Pruss. 2016. “Extinction intensity, selectivity and their combined macroevolutionary influence in the fossil record.” *Biology Letters* 12 (10): 20160202. doi:[10.1098/rsbl.2016.0202](https://doi.org/10.1098/rsbl.2016.0202).

Peters, Shanan E. 2008. “Environmental determinants of extinction selectivity in the fossil record.” *Nature* 454 (7204): 626–29. doi:[10.1038/nature07032](https://doi.org/10.1038/nature07032).

Smits, Peter D. 2015. “Expected time-invariant effects of biological traits on mammal species duration.” *Proceedings of the National Academy of Sciences* 112 (42): 13015–20. doi:[10.1073/pnas.1510482112](https://doi.org/10.1073/pnas.1510482112).

Spencer-Cervato, Cinzia. 1999. “The Cenozoic deep sea microfossil record: explorations of the DSDP/ODP sample set using the Neptune database.” *Palaeontologia Electronica* 2 (2): 4–286. [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](http://scholar.google.com/scholar?hl=en%7b\&%7dbtnG=Search%7b\&%7dq=intitle:THE+CENOZOIC+DEEP+SEA+MICROFOSSIL+RECORD+:+EXPLORATIONS+OF+THE+DSDP+/+ODP+SAMPLE+SET+USING+THE+NEPTUNE+DATABASE%7b\#}0).

Stan Development Team. 2017. “Stan Modeling Language Users Guide and Reference Manual.” <http://mc-stan.org>.

Tang, Wan, Hua He, and Xin M Tu. 2012. *Applied categorical and count data analysis*. Boca Raton, FL: CRC Press.

Tutz, Gerhard, and Matthis Schmid. 2016. *Modeling discrete time-to-event data*. Springer International Publishing. doi:[10.1007/978-3-319-28158-2](https://doi.org/10.1007/978-3-319-28158-2).

Van Valen, Leigh. 1973. “A new evolutionary law.” *Evolutionary Theory* 1: 1–30.

Wang, Steve C, and Charles R Marshall. 2016. “Estimating times of extinction in the fossil record.” *Biology Letters* 12 (4): 20150989.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2018. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.