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

Smits, Peter  
psmits@berkeley.edu

Finnegan, Seth  
sethf@berkeley.edu

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). 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 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. We 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 because our in-sample model performance measures are approximately equal to out-of-sample performance. 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. 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. 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.

# 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 (P. G. Harnik, Simpson, and Payne 2012; Seth Finnegan et al. 2015). 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) (Jonathan L Payne and Finnegan 2007; Jonathan L. Payne et al. 2016; T. H. G. Ezard et al. 2011). 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 (Harnik 2011; Smits 2015; Peters 2008; Jonathan L Payne and Finnegan 2007; P. G. Harnik, Simpson, and Payne 2012; T. H. G. Ezard et al. 2011; Michael Foote 2006) or on how to identify or measure these effects (J Alroy 2010; John Alroy 2014; John Alroy et al. 2001; John Alroy, Koch, and Zachos 2000; John Alroy 2000; Michael Foote 2001). 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 (Gelman and Hill 2006). 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 (Michael Foote 2007; Liow et al. 2010; L. H. Liow and Stenseth 2007; Kiessling and Kocsis 2016). We also know that a species’ geographic range size is a good predictor of differences extinction risk (Jonathan L Payne and Finnegan 2007; D. Jablonski and Roy 2003; David Jablonski 2008; David Jablonski and Hunt 2006). 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 (Michael Foote 2007; Liow et al. 2010; L. H. Liow and Stenseth 2007; Kiessling and Kocsis 2016), the potential predictive impact of this change has been under-evaluated (but see Kiessling and Kocsis (2016)). 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 (Smits 2015; S Finnegan, Payne, and Wang 2008; T. H. G. Ezard et al. 2012; Van Valen 1973; Liow et al. 2011; Crampton et al. 2016). 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 (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.

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 (Zachos, Dickens, and Zeebe 2008) – 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 (Zachos, Dickens, and Zeebe 2008).

# 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 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 (J Alroy 2010; John Alroy 2000; John Alroy 2014; M Foote 1997; M Foote and Sepkoski 1999; Michael Foote 2001; Michael Foote and Raup 1996; Lloyd et al. 2012; C. R. Marshall 1995; S. C. Wang and Marshall 2016). However, this practice is common with marine microfossil data due to their exceptional preservation rate (T. H. Ezard, Thomas, and Purvis 2013; T. H. Ezard and Purvis 2016; T. H. G. 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.

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 Earth) between any two occurrences of that species; this measure is also called a geodesic. This distance was measured in kilometers. Geographic range was then log-plus-one transformed, and standardized by zero-centering the data and then dividing by the standard deviation of the distribution of geographic ranges. This standardization means that the analyzed values of 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, that coefficients associated with similarly standardized covariates will be directly comparable in magnitude, and that the interecpt 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 transformed and standardized geographic range values and not standardized seperately.

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. Temperature was then transformed and standardized the in the same manner as geographic range (above). Difference in temperature between between observations was measured from the transformed and standardized temperature values and not standardized seperately.

## 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].

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, Calcaeous nannoplankton).

Change in geographic range since last observation.

Temperature at previous observation.

[tab:model\_def]

## 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 (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).

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 (Wickham et al. 2018), purr (Henry and Wickham 2018), and tidybayes (Kay 2018), thus some familiarity with that package ecosystem is necessary to fully comprehend how we’ve processed our data and results.

# 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.

## 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. [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 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.

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

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

## Cross-validation

Expected out-of-sample predictive performance was estimated using five-fold cross-validation, modified for time series data (Hastie, Tibshirani, and Friedman 2009). 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 (Hastie, Tibshirani, and Friedman 2009). Additionally, multimodality increases with model complexity (Fig. [fig:fold\_auc]) – 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. [fig:fold\_auc]) reveals a similar range in plausible values for all models as the in-sample AUC posterior predictive distributions (Fig. [fig:auc\_hist]). Interestingly, the differences between the posterior predictive distributions for the models have decreased. For example, model VP not clearly better than either models V or CP (Fig. [fig:fold\_auc]), which were shown earlier to be obviously worse-performing models based on in-sample performance (Fig. [fig:auc\_hist]). 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. [fig:auc\_hist]), indicating that our models would yield consistent results when generalized to future extinctions.

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

Comparison of out-of-sample AUC values over time as aggregated by taxonomic group for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogentity in performance within and between folds. This presentation decomposes each of the 12-million year folds by each of the taxonomic groups into the predictions made for each of the million-year intervals. The black line corresponds to the median AUC estimate, with the envelopes corresponding to multiple credible intervals as indicated in the legend. See Table [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 heterogentity in performance within and between folds. This presentation decomposes each of the 12-million year folds by each of the taxonomic groups into the predictions made for each of the million-year intervals. The black line corresponds to the median AUC estimate, with the envelopes corresponding to multiple credible intervals as indicated in the legend. See Table [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 rank the extinction risk of two randomly selected out-of-sample observations approximately 70% to 80% of the time. A slight decrease in performance when dealing with out-of-sample observations makes sense: each of the models fit during cross-validation is based on less data than the model fit on the full data (between 1/5th to 4/5ths of the original). The similarly between the in-sample and out-of-sample results indicates that our model is fairly robust to how extinction intensity has changed over the Cenozoic.

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

Other noteworthy aspects of our in-sample results are that AUC estimates for our models models differ significantly between models and that all of these estimates are in a narrow range of possible AUC values (Fig. [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.

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 intercept is greater than 2 or less than -2 because the inverse-logit of those values are close to 0 or 1, respectively (Gelman and Hill 2006).

The success of model 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 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 (T. H. G. 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.

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.

# Supplementary Material

## 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 . 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 , 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 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 (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 (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 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 (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 (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>.

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<0707:NMFQMP>2.0.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).

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., T. Aze, P. N. Pearson, and A. 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, T. H. G., P. N. Pearson, T. Aze, and A. 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., 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<0602:ITPOPO>2.0.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).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. “The Elements of Statistical Learning.” *Bayesian Forecasting and Dynamic Models* 1: 1–694. 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<0731:DBSAGE>2.3.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>.

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

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>.

Zachos, James C, Gerald R Dickens, and Richard E Zeebe. 2008. “An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics.” *Nature* 451 (7176): 279–83. doi:[10.1038/nature06588](https://doi.org/10.1038/nature06588).