Materials and Methods

Taxon occurrences and species-level information

All fossil occurrence information information was downloaded from the Paleobiology Database. Occurrences (PBDB) were restricted to all Mammalia sampled in North America between the Maastrichtian and Gelasian stages. Taxonomic, stratigraphic, and ecological metadata for each occurrence was included. The raw data is available for download at `http://goo.gl/2slgeU`.

This raw data was then sorted, cleaned, and manipulated programmatically prior to analysis. Species taxonomic assignments given by the PBDB were updated for accuracy and consistency. For example, species classified in the order Artidodactyla were reclassified as Cetartiodactyla. These re-assignments follow **?**] and were based on CITATIONS. Additionally, Taxa who's life habit was classified as either volant (i.e. Chiroptera) or aquatic (e.g. Cetacea) were excluded from this analysis because of both differences in fossilization potential and applicability to the study of terrestrial species pools.

The life habit and dietary categories provided through the PBDB where coarsened to increase per ecotype sample size; this coarsening follows the same procedure as **?**]. Additionally, life habit category was further modified to break-up the vague "ground-dwelling" category; re-classifying these species by ankle posture gives more precise information about that species' environmental context. Ground-dwelling taxa were reassigned following **?**] by species taxonomic context. Species ecotype is defined as the interaction between life habit and diet categories. Ecotype categories with less than 10 species havig ever been that combination were excluded, yielding a total of 18 of 21 possible ecotypes.

Species mass information was gathered from multiple different sources where a plurality of the body size estimates are from the PBDB. Body part measurements for many species are also available through the PBDB. Just as with **?**], these measurements and corresponding regression equations were used to get mass estimates for more species. Additional mass estimates and body part measurements were sourced from CITATIONS, the Neogene Old World Database as in **?**]. Mass was log-transformed and then mean-centered and rescaled by dividing by two-times its standard deviation; this insures

that the magnitude of effects for both continuous and discrete covariates are comparable and follows CITATION.

All fossil occurrences from 64 to 2 million years ago (Mya) were binned into 31 2 million year (My) bins. This temporal length was chosen because it is approximately the resolution of the North American mammal fossil record.

Environmental and temporal covariates

The group-level covariates in this study are descriptors of species' environmental context, such as global temperature and floral interval.

Global temperature was calculated from Mg/Ca isotope data across the Cenozoic CRAMER CITATION. WHY DID I DO THIS? TOM EZARD HAS USED THIS AND IT IS PROBABLY BETTER THAN THE ZACHOS CURVE. Two aspects of the Mg/Ca-based temperature curve were included in this analysis: mean and range. Both were calculated as the mean of all respective estimates for each 2 My temporal bins. Both mean and range were then rescaled as above: subtract mean, divide by twice the standard deviation.

Floral interval is a holistic descriptor of which taxonomic groups are present and their relative modernity of those taxa for a given temporal interval CITATION. Graham CITATION defines four intervals from the Cretaceous to the Pliocene; only three of these intervals are included in this analysis.

Modelling species occurrence

Two different models were used in this study: a pure-presence model and a birth-death model. Both models at their core are hidden Markov model with an absorbing state. The difference between these two models is if the probability of a species origination and survival are considered equal or different.

Data augmentation

All presence/absence observations are incomplete. The hidden Markov model at the core of this analysis allows for observed absences to be used meaningfully

to estimate the number of unobserved species. Of specific concern in this analysis is the unknown "true" size of the dataset; how many species could have actually been observed? While many species have been observed, the natural incompleteness of all observations, especially in the case of paleontological data, there are obviously many species which were never sampled.

Let N by the total number of observed species, M be the upper limit of possible species that could have existed given a model of species presence, and N^* is the all-zero histories where $N^* = M - N$. This approach assumes that $\hat{N} \sim \text{Binomial}(M, \psi)$ where $\hat{N}$ is the estimated "true" number of species and ψ is the probability that any augmented species should actually be "present." Because M is user defined, this approach effectively gives ψ a uniform prior over N to M CITATION. For a more detailed explanation of data augmentation and its application here, please see CITATION, CITATION, and CITATION.

Data imputation is the process of estimating missing data for partially observed covariates CITATION, this is simple in a Bayesian context because data are also parameters CITATION. Augmented species also have no known mass so a mass estimate must be imputed for each possible species CITATION ROYLE. This procedure assumes that mass values for augmented species are from the same distribution as observed species. The distribution of observed mass values is estimated as part of the model, and new mass values are then generated from this distribution. This approach is an example of imputing data missing completely at random CITATION. Because log mass values are rescaled as a part of this study, the body mass distribution is already known ($\mathcal{N}(0, 0.5)$); augmented species body mass just simply drawn from this distribution.

These augmented species are ecotypically classified as augmented, an additional grouping indicating their unknown biology.

Observation process

The type of hidden Markov model used in this study has three characteristic probabilities: probability p of observing a species given that it is present, probability ϕ of a species surviving from one time to another, and probability π of a species first appearing CITATION. In this formulation, the probability

Table 1: Observation parameters

Parameter	dimensions	explanation
y	$N \times T$	observed species presence/absence
z	$N \times T$	"true" species presence/absence
p	T	probability of observing a species that is present at time t
m	N	species log mass, rescaled
α_0	1	average log-odds of p
α_1	1	change in average log-odds of p per change mass
r	T	difference from α_0 associated with time t
σ	1	standard deviation of r

of a species going extinct is $1 - \pi$. For the pure-presence model $\phi = \pi$, while for the birth-death model $\phi \neq \pi$.

The probability of observing a species that is present p is modeled as a logistic regression was a time-varying intercept and species mass as a covariate. The effect of species mass on p was assumed linear and constant over time and given a prior reflecting a possible positive relationship; these assumptions are reflected in the structure of the model Equation 1. The parameters associated with this part of the model are described in Table 1.

$$\begin{aligned} y_{i,t} &\sim \text{Bernoulli}(p_{i,t} z_{i,t}) \\ p_{i,t} &= \text{logit}^{-1}(\alpha_0 + \alpha_1 m_i + r_t) \\ r_t &\sim \mathcal{N}(0, \sigma) \end{aligned} \tag{1}$$

Pure-presence process

For the pure-presence model there is only a single probability dealing with the presence of a species θ. This probability was modeled as multi-level logistic regression with both species-level and group-level covariates CITATION. The parameters associated with pure-presence model are presented in Table 2 and the full sampling statement in Equation 2.

The species-level of the model (Eq. 2) is a varying-intercept model where the intercept varies by ecotype. Additionally, species mass was included as a covariate associated with two regression coefficients allowing a quadratic

Table 2: Parameters for the model of presence in the pure-presence model

Parameter	dimensions	explanation
z	$N \times T$	"true" species presence/absence
θ	$N \times T - 1$	probability of $z = 1$
a	$T - 1 \times D$	ecotype-varying intercept; mean value of log-odds of θ
m	N	species log mass, rescaled
b_1	1	effect of species mass on log-odds of θ
b_2	1	effect of species mass, squared, on log-odds of θ
U	$T \times D$	matrix of group-level covariates
γ	$U \times D$	matrix of group-level regression coefficients
Σ	$D \times D$	covariance matrix of a
Ω	$D \times D$	correlation matrix of a
τ	D	vector of standard deviations for each ecotype a_d

relationship with log-odds of occurrence. This assumption is based on the known distribution of mammal body masses where species with intermediate mass values are more common than either small or large bodied species. These assumptions are also reflected in the choice of priors for these regression coefficients.

The values of each ecotype's intercept are themselves modeled as regressions using the group-level covariates associated with environmental context. Each of these regressions has an associated variance of possible values of each ecotype's intercept CITATION. In addition, the covariances between ecotype intercepts, given this group-level regression, are modeled CITATION.

Following recommendations from CITATION, CITATION, CITATION all parameters not modeled elsewhere were given weakly informative priors. Weakly informative means that priors do not necessarily encode actual prior information but instead help regularize or weakly constrain posterior estimates. These priors have a concentrated probability density around and near zero; this has the effect of tempering our estimates and help prevent overfitting the model to the data CITATION.

$$
\begin{aligned}
y_{i,t} &\sim \text{Bernoulli}(p_{i,t} z_{i,t}) \\
p_{i,t} &= \text{logit}^{-1}(\alpha_0 + \alpha_1 m_i + r_t) \\
r_t &\sim \mathcal{N}(0, \sigma) \\
z_{i,t} &\sim \text{Bernoulli}(\theta_{i,t}) \\
\theta_{i,t} &= \text{logit}^{-1}(a_{t,j[i]} + b_1 m_i + b_2 m_i^2) \\
a &\sim \text{MVN}(u\gamma, \Sigma) \\
\Sigma &= \text{diag}(\tau)\Omega\text{diag}(\tau) \\
\alpha_0 &\sim \mathcal{N}(0, 1) \\
\alpha_1 &\sim \mathcal{N}(1, 1) \\
\sigma &\sim \mathcal{N}^+(1) \\
b_1 &\sim \mathcal{N}(0, 1) \\
b_2 &\sim \mathcal{N}(-1, 1) \\
\gamma &\sim \mathcal{N}(0, 1) \\
\tau &\sim \mathcal{N}^+(1) \\
\Omega &\sim \text{LKJ}(2)
\end{aligned}
\tag{2}
$$

Birth-death process

In the birth-death model, $\phi \neq \pi$ and so each of these probabilities are modeled separately but in a similar manner to how θ is modeled in the pure-presence model (Eq. 2). The parameters associated with the birth-death presence model are presented in Table 3 and the full sampling statement, including observation (Eq. 1), is described in Equation 3.

Table 3: Parameters for the model of presence in the pure-presence model

Parameter	dimensions	explanation
z	$N \times T$	"true" species presence/absence
ϕ	$N \times T - 1$	probability of $z_{\_,t} = 1 \vert z_{\_,t-1} = 0$
π	$N \times T - 1$	probability of $z_{\_,t} = 1 \vert z_{\_,t-1} = 1$
a^{ϕ}	$T - 1 \times D$	ecotype-varying intercept; mean value of log-odds of θ
a^{π}	$T - 1 \times D$	ecotype-varying intercept; mean value of log-odds of θ
m	N	species log mass, rescaled
b_1^{ϕ}	1	effect of species mass on log-odds of ϕ
b_1^{π}	1	effect of species mass on log-odds of π
b_2^{ϕ}	1	effect of species mass, squared, on log-odds of ϕ
b_2^{π}	1	effect of species mass, squared, on log-odds of π
U	$T \times D$	matrix of group-level covariates
γ^{ϕ}	$U \times D$	matrix of group-level regression coefficients
γ^{π}	$U \times D$	matrix of group-level regression coefficients
Σ^{ϕ}	$D \times D$	covariance matrix of a^{ϕ}
Σ^{π}	$D \times D$	covariance matrix of a^{π}
Ω^{ϕ}	$D \times D$	correlation matrix of a^{ϕ}
Ω^{π}	$D \times D$	correlation matrix of a^{π}
τ^{ϕ}	D	vector of standard deviations for each ecotype a_d^{ϕ}
τ^{π}	D	vector of standard deviations for each ecotype a_d^{π}

$$
\begin{aligned}
y_{i,t} &\sim \text{Bernoulli}(p_{i,t} z_{i,t}) \\
p_{i,t} &= \text{logit}^{-1}(\alpha_0 + \alpha_1 m_i + r_t) \\
r_t &\sim \mathcal{N}(0, \sigma) \\
\alpha_0 &\sim \mathcal{N}(0, 1) \\
\alpha_1 &\sim \mathcal{N}(1, 1) \\
\sigma &\sim \mathcal{N}^+(1) \\
z_{i,1} &\sim \text{Bernoulli}(\rho) \\
z_{i,t} &\sim \text{Bernoulli}\left(z_{i,t-1}\pi + \sum_{x=1}^{t} (1 - z_{i,x})\phi_t \right) \\
\phi_{i,t} &= \text{logit}^{-1}(a^{\phi}_{t,j[i]} + b^{\phi}_1 m_i + b^{\phi}_2 m_i^2) \\
\pi_{i,t} &= \text{logit}^{-1}(a^{\pi}_{t,j[i]} + b^{\pi}_1 m_i + b^{\pi}_2 m_i^2) \\
a^{\phi} &\sim \text{MVN}(u\gamma^{\phi}, \Sigma^{\phi}) \\
a^{\pi} &\sim \text{MVN}(u\gamma^{\pi}, \Sigma^{\pi}) \\
\Sigma^{\phi} &= \text{diag}(\tau^{\phi})\Omega^{\phi}\text{diag}(\tau^{\phi}) \\
\Sigma^{\pi} &= \text{diag}(\tau^{\pi})\Omega^{\pi}\text{diag}(\tau^{\pi}) \\
\rho &\sim \text{U}(0, 1) \\
b^{\phi}_1 &\sim \mathcal{N}(0, 1) \\
b^{\pi}_1 &\sim \mathcal{N}(0, 1) \\
b^{\phi}_2 &\sim \mathcal{N}(-1, 1) \\
b^{\pi}_2 &\sim \mathcal{N}(-1, 1) \\
\gamma^{\phi} &\sim \mathcal{N}(0, 1) \\
\gamma^{\pi} &\sim \mathcal{N}(0, 1) \\
\tau^{\phi} &\sim \mathcal{N}^+(1) \\
\tau^{\pi} &\sim \mathcal{N}^+(1) \\
\Omega^{\phi} &\sim \text{LKJ}(2) \\
\Omega^{\pi} &\sim \text{LKJ}(2)
\end{aligned}
\tag{3}
$$

Posterior inference and model adequacy

Programs that implement joint posterior inference for the above models (Eqs. 2, 3) were implemented in the probabilistic programming language Stan CITATION. The models used here both feature latent discrete parameters in the large matrix z (Tables 1, 2, 3; Eqs. 1, 2, 3). All methods for posterior inference implemented in Stan are derivative based which causes complications for actually implementing the above models because integers do not have derivatives. Instead of implementing a latent discrete parameterization, the posterior probabilities of all possible states of the latent parameters z were estimated (i.e. marginalized).

Species durations at minimum range-through from the FAD to the LAD, but the incompleteness of all observations means that the actual time of origination or extinction is unknown. The marginalization approach used here means that the probabilities all possible histories for a species are calculated, from the end members of the species having existed for the entire study interval and the species having only existed between the directly observed FAD and LAD to all possible intermediaries CITATION.

The combined size of the dataset and large number of parameters in both models (Eqs. 2, 3), specifically the total number of latent parameters that are the matrix z, means that stochastic approximate posterior inference is computationally very slow even using HMC. Instead, an approximate Bayesian approach was used: variational inference. A recently developed automatic variational inference algorithm called "automatic differention variational inference" (ADVI) is implemented in Stan and was used here CITATION. ADVI assumes that the posterior is Gaussian but still yields a true Bayesian posterior; this assumption is similar to quadratic approximation of the likelihood function used in maximum likelihood inference CITATION. The principal limitation of assuming the joint posterior is Gaussian is that the true topology of the log-posterior isn't represented; this is a particular burden for scale parameters which are bound to be positive (e.g. standard deviation).

After fitting both models (Eqs. 2, 3) using ADVI, model adequacy and quality of fit was assessed using a series of posterior predictive checks CITATION CITATION. Because all Bayesian models are inherently generative, simulations of new data sets is "free" CITATION. By simulating many theoretical data sets using the observed covariate information the congruence between predic-

tions made by the model and the observed empirical data can be assessed. By combining multiple posterior predictive tests of congruence between empirical and simulated values of interest, the holistic adequacy of the model can be analyzed CITATION.

An example posterior predictive check used in this study was comparing the observed average number of observations per species to a distribution of simulated averages; if the empirically observed value sits in the middle of the distribution than the model is adequate in reproducing the observed number of occurrences per species.