1 Materials and Methods

1.1 Fossil occurrence information

The dataset analyzed here was sourced from the Paleobiology Database (http://www.paleodb.org) which was then filtered based on taxonomic, temporal, stratigraphic, and other occurrence information that was necessary for this analysis. These filtering criteria are very similar to those from **?**] with an additional constraint of being present in the body size data set from **?**]. Epicontinental versus open-ocean assignments for each fossil occurrence are partially based on those from **?**], with additional occurrences assigned similarly (Miller and Foote, personal communication).

Fossil occurrences were analyzed at the genus level which is common for paleobiological, macroevolution, or macroecological studies of marine invertebrates [**? ? ? ? ? ? ? ? ? ?**]. While species diversity dynamics are of much greater interest than those of higher taxa, the nature of the fossil record makes accurate and precise taxonomic assignments at the species level for all occurrences. In particular, the simplicity of brachiopod external morphology and the quality of preservation makes it very difficult for assignments below the genus level. As such, the choice to analyze genera as opposed to species was in order to assure a minimum level of confidence and accuracy in the data analyzed here.

Sampled occurrences were restricted to those with paleolatitude and paleolongitude coordinates, assignment to either epicontinental or open-ocean environment, and belonging to a genus present in the body size dataset [**?**].

Genus duration was calculated as the number of geologic stages from first appearance to last appearance, inclusive. Durations were based on geologic stages as opposed to millions of years because of the inherently discrete nature of the fossil record; dates are not assigned to fossils themselves but instead fossils are known from a geological interval which represents some temporal range. Stages are effectively irreducible temporal intervals in which taxa may occur.

Genera with a last occurrence in or after Changhsingian stage were right censored at the Changhsingian. Genera with a duration of only one stage were left censored (Appendix **??**). The covariates used to model genus duration

were geographic range size (r), environmental preference (v, v^2), and body size (m).

Geographic range was calculated using an occupancy approach. First, all occurrences were projected onto an equal-area cylindrical map projection. Each occurrence was then assigned to one of the cells from a 70×34 regular raster grid placed on the map. Each grid cell represents approximately 250,000 km^2. The map projection and regular lattice were made using shape files from http://www.naturalearthdata.com/ and the `raster` package for R [?].

For each stage, the total number of occupied grid cells, or cells in which a fossil occurs, was calculated. Then, for each genus, the number of grid cells occupied by that genus was calculated. Dividing the genus occupancy by the total occupancy gives the relative occupancy of that genus. Mean relative genus occupancy was then calculated as the mean of the per stage relative occupancies of that genus.

Body size data was sourced directly from **?**]. Because those measurements are presented without error, a measurement error model similar to the one for environmental affinity could not be implemented (Appendix **??**).

Environmental preference v was defined as the probability of observing the ratio of epicontinental occurrences to total occurrences (e_i/E_i) is represented by θ_i, given the probability of an epicontinental environment occurring θ_i'. This second probability is estimated from the observed occurrences (e_i', E_i') of all other taxa that occur during the duration of taxon i, which is given a flat Beta prior. This measure of environmental preference is expressed.

$$
\begin{aligned}
p\left(\theta_i' \left| \frac{e_i'}{E_i'} \right.\right) &\propto \text{Beta}(e_i', E_i' - e_i')\text{Beta}(1, 1) \\
&= \text{Beta}(e_i' + 1, E_i' - e_i' + 1) \\
v &= p(\theta_i | \theta_i')
\end{aligned}
\tag{1}
$$

Prior to analysis, geographic range r and body size m were transformed and standardized in order to improve interpretability of the results. Geographic range, which can only vary between 0 and 1, was logit transformed. Body size, which is defined for all positive real values, was natural log transformed. These covariates were then standardized by mean centering and dividing by two times their standard deviation following **?**]. Environmental preference v was

rescaled in the same manner. This standardization allows for all regression coefficients to be comparable as the estimate represents the change in risk per change in standard deviation of the covariate.

1.2 Analytical approach

Hierarchical modelling is a statistical approach which explicitly takes into account the structure of the observed data in order to model both the within and between group variance [? ?]. The units of study (e.g. genera) each belong to a single grouping (e.g. origination cohort). These groups are considered draws from a shared probability distribution (e.g. all cohorts, observed and unobserved). The group-level parameters are then estimated simultaneously as the other parameters of interest (e.g. covariate effects) [?]. The subsequent estimates are partially pooled together, where parameters from groups with large samples or effects remain large while those of groups with small samples or effects are pulled towards the overall group mean.

This partial pooling is one of the greatest advantages of hierarchical modeling. By letting the groups "support" each other, parameter estimates then better reflect our statistical uncertainty. Additionally, this partial pooling helps control for multiple comparisons and possibly spurious results as effects with little support are drawn towards the overall group mean [? ?].

All covariate effects (regression coefficients), as well as the intercept term (baseline extinction risk), were allowed to vary by group (origination cohort). The covariance/correlation between covariate effects was also modeled. This hierarchical structure allows inference for how covariates effects may change with respect to each other while simultaneously estimating the effects themselves, propagating our uncertainty through all estimates.

1.3 Survival model

Genus durations were modeled as time-till-event data [?], with covariate information used in estimates of extinction risk as a hierarchical regression model. Genus durations were assumed to follow a Weibull distribution. While the exponential distribution assumes that extinction risk is independent of duration, the Weibull distribution allows for age dependent extinction [?

]. The Weibull distribution has two parameters: a scale σ, and a shape α. When $\alpha = 1$, σ is equal to the expected duration of any taxon. α acts as a time dilation effect where values greater than 1 indicate that extinction risk increases with age, and values less than 1 indicate that extinction risk decreases with age. Note that the Weibull distribution is equivalent to the exponential distribution when $\alpha = 1$.

The scale parameter σ was modeled as a regression with both varying intercept and varying slopes. The following variables are defined: y_i is the duration of genus i in geologic stages, X is the matrix of covariates including a constant term, B_j is the vector of regression coefficients for origination cohort j, μ is the vector of means of each regression coefficient, Σ is the covariance matrix of the regression coefficients, τ is the vector of scales the standard deviations of the between-cohort variation in regression coefficient estimates, and Ω is the correlation matrix of the regression coefficients. The correlation matrix Ω is then given an almost flat LKJ distributed prior.

The shape parameter α was also modeled as a regression with intercept α' and standard deviation σ^α. The effect of origination cohort a_j is modeled as draws from a shared normal distribution with mean 0 and standard deviation σ^a. γ is the regression coefficient for the effect of the rescaled logarithm of the number of samples s of taxon i.

Except where noted, regression coefficients were given a weakly informative normally distributed prior, scale (e.g. standard deviation) parameters were given a weakly informative half-Cauchy prior.

The full sampling statement, along with all necessary priors, is expressed as

$$
\begin{aligned}
y_i &\sim \text{Weibull}(\alpha_i, \sigma_i) \\
\sigma_i &= \exp\left(\frac{-(\mathbf{X}_i B_{j[i]})}{\alpha_i}\right) \\
B_j &\sim \text{MVN}(\mu, \Sigma) \\
\mu_0 &\sim \mathcal{N}(0, 5) \\
\mu_r &\sim \mathcal{N}(-1, 1) \\
\mu_v &\sim \mathcal{N}(0, 1) \\
\mu_{v^2} &\sim \mathcal{N}(1, 1) \\
\mu_m &\sim \mathcal{N}(0, 1) \\
\Sigma &= \text{Diag}(\tau)\Omega\text{Diag}(\tau) \\
\tau &\sim \text{C}^+(1) \\
\Omega &\sim \text{LKJ}(2) \\
\alpha_i &= \exp\left(\mathcal{N}(\alpha' + a_{j[i]} + \gamma s_i, \sigma^\alpha)\right) \\
\alpha' &\sim \mathcal{N}(0, 1) \\
a_j &\sim \mathcal{N}(0, \sigma^a) \\
\sigma^a &\sim \text{C}^+(1) \\
\gamma &\sim \mathcal{N}(0, 1) \\
\sigma^\alpha &\sim \text{C}^+(1).
\end{aligned}
\tag{2}
$$

1.4   Parameter estimation

The joint posterior was approximated using a Markov-chain Monte Carlo routine that is a variant of Hamiltonian Monte Carlo called the No-U-Turn Sampler [?] as implemented in the probabilistic programming language Stan [?]. The posterior distribution was approximated from four parallel chains run for 10,000 draws each, split half warm-up and half sampling and thinned to every 10th sample for a total of 5000 posterior samples. Chain convergence was assessed via the scale reduction factor $\hat{R}$ where values close to 1 ($\hat{R} < 1.1$) indicate approximate convergence. Convergence means that the chains are approximately stationary and the samples are well mixed [?].

1.5 Model evaluation

Models were evaluated using both posterior predictive checks and an estimate of out-of-sample predictive accuracy. The motivation behind posterior predictive checks as tools for determining model adequacy is that replicated data sets using the fitted model should be similar to the original data [**?**]. Systematic differences between the simulations and observations indicate weaknesses of the model fit. An example of a technique that is very similar would be inspecting the residuals from a linear regression.

The strategy behind posterior predictive checks is to draw simulated values from the joint posterior predictive distribution, $p(y^{rep}|y)$, and then compare those draws to the empirically observed values [**?**]. To accomplish this, for each replicate, a single value is drawn from the marginal posterior distributions of each regression coefficient from the final model as well as α for the Weibull model (Eq. 2). Then, given the covariate information $\mathbf{X}$, a new set of n genus durations are generated giving a single replicated data set y^{rep}. This is repeated 1000 times in order to provide a distribution of possible values that could have been observed given the model.

In order to compare the fitted model to the observed data, various graphical comparisons or test quantities need to be defined. The principal comparison used here is a comparison between non-parameteric approximation of the survival function $S(t)$ as estimated from both the observed data and each of the replicated data sets. The purpose of this comparison is to determine if the model approximates the same survival/extinction pattern as the original data.