Appendix for: Tempotal variation and correlation of effects of biological traits on taxonomic survival of brachiopods

A Uncertainty in environmental preference

The calculation and inclusion of environmental affinity in the survival model is a statistical procedure that takes into account our uncertainty based on where fossils tend to occur. Because we cannot directly observe if a fossil taxon had occurrences restricted to only a single environment, instead we can only estimate its affinity with uncertainty. One advantage of using a Bayesian analytical approach is that both parameters and data are considered random samples from some underlying distribution, which means that is is possible to model the uncertainty in our covariates of interest [5]. My approach is conceptually similar to Simpson and Harnik [11] but instead of obtaining a single point estimate, an entire posterior distribution is estimated.

The first step is to determine the probability θ at which genus i occurs in an epicontinental settings based on its own pattern of occurrences. Define e_i as the number of occurrences of genus i in an epicontiental sea and o_i as the number of occurrences of genus i not in an epicontinental sea (e.g. open ocean). Because the value of interest is the probability of occurring in an epicontinental environment, given the observed fossil record, I assume that probability follows a Bernoulli distribution. We can then define our sampling statement as

$$e_i \sim \text{Bernoulli}(e_i + o_i, \theta_i). \tag{S1}$$

I used a flat prior for θ_i defined as $\theta_i \sim \text{Beta}(1, 1)$. Because the beta distribution is the conjugate prior for the Bernoulli distribution, the posterior is easy

to compute in closed form. The posterior probability of θ is then

$$\theta_i \sim \text{Beta}(e_i + 1, o_i + 1) \tag{S2}$$

It is extremely important, however, to take into account the overall environmental occurrence probability of all other genera present at the same time as genus i. This is incorporated as an additional probability Θ. Define E_i as the total number of other fossil occurrences (exceptfor genus i) in epicontinental seas during stages where i occurs and O_i as the number of other fossil occurrences not on epicontinental seas. We can then define the sampling statement as

$$E_i \sim \text{Bernoulli}(E_i + O_i, \Theta_i). \tag{S3}$$

Again, I used a flat prior of Θ_i defined as $\Theta_i \sim \text{Beta}(1, 1)$. The posterior of Θ is then simply defined as

$$\Theta_i \sim \text{Beta}(E_i + 1, O_i + 1) \tag{S4}$$

I then define the environmental affinity of genus i as $v_i = \theta_i - \Theta_i$. v_i is a value that can range between -1 and 1, where negative values indicate that genus i tends to occur more frequently in open ocean environments than background while positive values indicate that genus i tends to occur in epicontiental environments.

While this approach is noticeably more complicated than previous ones [3, 8, 10, 11] there are some important benefits to both using a continuous measure of affinity as well directly modeling our uncertainty. In order to show some of these benefits, I performed a simulation analysis of how modal/maximum *a posteriori* (MAP) estimates versus full posterior estimates.

In this simulation, I first defined the "background" epicontinental occurrence θ_b as 0.50 with a small amount of noise. This was represented as a beta distribution

$$\Theta_b = \text{Beta}(\alpha = 2500, \beta = 2500). \tag{S5}$$

This choice of parameters for the distribution reflects the average number of background occurrences for either epicontinental or open ocean environments per genus.

Using this background occurrence ratio, I randomly generated the occurrence patterns of 1000 simulated taxa. This was done at multiple sample sizes (1, 2, 3, 4, 5, 10, 25, 50, 100) in order to demonstrate the effects of increasing sample size on the confidence of environmental affinity. For each simulated taxon I calculated the full posterior distribution while assuming a flat Beta prior (Beta$(1, 1)$). Using the full posterior I calculated the MAP probability of occurring in epicontinental environments. The environmental affinity was calculated for each of the simulated taxa using both the full posterior and the MAP estimate. In this toy example, environmental affinity can range between -0.5 and 0.5.

As should be expected, as sample size increases the distribution of MAP estimates converge on the true value (Fig. S1). For taxa with less than 10 occurrences, the MAP estimate is biased towards extreme values. Note that the mode of the beta distribution is not defined for situations where there were 0 draws of one of the environmental conditions. Instead, the vertical line is based entirely on the observed occurrences which are technically the modal estimates because they are the most frequently occurring/highest density.

In contrast, we can compare the true occurrence probability distribution versus the posterior estimate for a given sample (Fig. S2). When sample sizes are low, posterior estimates are flat and represent a compromise between the likelihood and the flat prior (Eq. S2). Because of this, estimates from small sizes are less likely to be overly biased towards the extremes. This is further emphasized by inspection of the estimates of environmental affinity for the simulated taxa (Fig. S3). Posterior estimates from simulated taxa with small sample size have a much broader distribution that both allows for the extreme observation but still captures the "true" value (0).

By defining environmental preference as the difference in full posterior estimates of occurrence probability, it is possible to include taxa with low sample sizes that are normal discarded [3, 8, 10, 11]. Additionally, 55+% of observed Paleozoic brachiopod genera have less than 10 occurrences which is the range of sample sizes where MAP (or ML) estimates would be potentially most biased. This is preferable to finding the difference between the MAP estimates (blue line; Fig. S3).

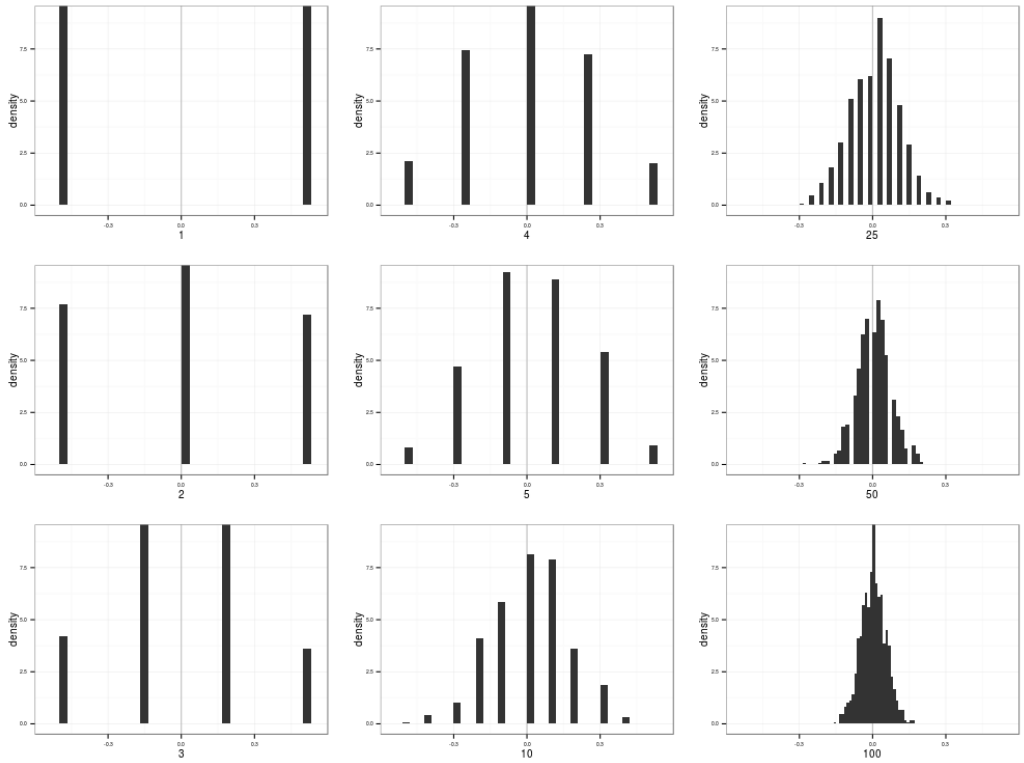

Figure S1: Histograms of the distributions of from the beta distribution defined in Eq. S5. As to be expected, as sample size increases the draws better resemble the underlying true distribution. Sample size is indicated as the label of the x-axis, increasing in column major order.

B Survival model

The simplest model of genus duration includes no covariate or structural information. Define y_i as the duration in stages of genus i, where $i = 1, \ldots, n$ and n is the number of observed genera. These two models are them simply defined as

$$\begin{aligned} y_i &\sim \text{Exponential}(\lambda) \\ y_i &\sim \text{Weibull}(\alpha, \sigma). \end{aligned} \tag{S6}$$

$\lambda, \alpha, \text{and} \sigma$ are all defined for all positive reals. Note that λ is a "rate" or inverse-scale while σ is a scale parameter, meaning that $\frac{1}{\lambda} = \sigma$.

These simple models can then be expanded to include covariate information as predictors by reparameterizing λ or σ as a regression [9]. Each of the covariates of interest is given its own regression coefficient (e.g. β_r) along with an intercept term β_0. There are some additional complications to the parameterization of σ associated with the inclusion of α as well as for interpretability [9]. Both of these are then written as

$$\begin{aligned} \lambda_i &= \exp(\beta_0 + \beta_r r_i + \beta_v v_i + \beta_{v^2} v_i^2 + \beta_m m_i) \\ \sigma_i &= \exp\left(\frac{-(\beta_0 + \beta_r r_i + \beta_v v_i + \beta_{v^2} v_i^2 + \beta_m m_i)}{\alpha}\right). \end{aligned} \tag{S7}$$

The quadratic term for environmental affinity v is to allow for the possible nonlinear relationship between environmental affinity and extinction risk.

The models which incorporate both equations S6 and S7 can then be further expanded to allow all of the β coefficients, including β_0, to vary with origination cohort while also modeling their covariance and correlation. This is called a varying-intercepts, varying-slopes model [6]. It is much easier to represent and explain how this is parameterized using matrix notation. First, define $\mathbf{B}$ as $k \times J$ matrix of the k coefficients including the intercept term ($k = 5$) for each of the J cohorts. Second, define $\mathbf{X}$ as a $n \times k$ matrix where each column is one of the covariates of interest. Importantly, $\mathbf{X}$ includes a columns of all 1s which correspond to the constant term β_0. Third, define $j[i]$ as the origination cohort of genus i, where $j = 1, \ldots, J$ and J is the total number of observed cohorts. We then rewrite λ and σ of equation S7 in matrix notation as

$$\begin{aligned} \lambda_i &= \exp(\mathbf{X}_i B_{j[i]}) \\ \sigma_i &= \exp\left(\frac{-(\mathbf{X}_i B_{j[i]})}{\alpha}\right). \end{aligned} \tag{S8}$$

Because B is a matrix, I use a multivariate normal prior with unknown vector of means μ and covariance matrix Σ. This is written as

$$B \sim \text{MVN}(\vec{\mu}, \Sigma) \tag{S9}$$

where $\vec{\mu}$ is length k vector representing the overall mean of the distributions of β coefficients. Σ is a $k \times k$ covariance matrix of the β coefficients.

What remains is assigning priors the elements of $\vec{\mu}$ and the covariance matrix Σ. All elements of $\vec{\mu}$ except for μ_r were given horseshoe priors [1, 2] while μ_r

was given an informative normal prior ($\mu_r \sim \mathcal{N}(-1, 1)$. Horseshoe priors are a strong regularizing priors with effectively infinite density at 0 and heavy, Cauchy-like tails [1, 2] which allow weakly inferred effects to be strongly drawn towards 0 while truly strong effects can remain large. The horseshoe prior consists of a normal distribution with scale term that is the product between a global shrinkage parameter ν and a locak shrinkage parameter ψ unique to each of the parameters of interest. These parameters are themselves given half-Cauchy priors (Eq. 1 and 2).

The prior for Σ is a bit more complicated due to its multivariate nature. Following the Stan Development Team [12], I modeled the scale terms separate from the correlation structure of the coefficients. This is possible because of the relationship between a covariance and a correlation matrix, defined as

$$\Sigma_B = \text{Diag}(\vec{\tau})\Omega\text{Diag}(\vec{\tau}) \tag{S10}$$

where $\vec{\tau}$ is a length k vector of variances and $\text{Diag}(\tau)$ is a diagonal matrix.

I used a LKJ prior distribution for correlation matrix Ω as recommended by Stan Development Team [12]. The LKJ distribution is a single parameter multivariate distribution where values of the parameter η greater than 1 concentrate density at the unit correlation matrix, which corresponds to no correlation between the β coefficients. The scale parameters, $\vec{\tau}$, are given weakly informative half-Cauchy (C^+) priors following Gelman [4].

C Censored observations

A key aspect of survival analysis is the inclusion of censored, or incompletely observed, data points [7, 9]. The two classes of censored observations encountered in this study were right and left censored observations. Right censored genera are those that did not go extinct during the window of observation, or genera that are still extant. Left censored observations are those taxa for which we know only an upper limit on their duration.

In the context of this study, I considered all genera that had a duration of only one geologic stage to be left censored as we do not have a finer degree of resolution.

The key function for modeling censored observations is the survival function, or $S(t)$. $S(t)$ corresponds to the probability that a genus having existed for t stages will not have gone extinct while $h(t)$ corresponds to the instantaneous extinction rate at taxon age t [9]. For an exponential model, $S(t)$ is defined as

$$S(t) = \exp(-\lambda t), \tag{S11}$$

and for the Weibull distribution $S(t)$ is defined as

$$S(t) = \exp\left(-\left(\frac{t}{\sigma}\right)^{\alpha}\right). \tag{S12}$$

$S(t)$ is equivalent to the complementary cumulative distribution function, $1 - F(t)$ [9].

For right censored observations, instead of calculating the likelihood as normal (Eq. S8) the likelihood of an observation is evaluated using $S(t)$. Conceptually, this approach calculates the likelihood of observing a taxon that existed for at least that long. For left censored data, instead the likelihood is calculated using $1 - S(t)$ which corresponds to the likelihood of observing a taxon that existed no longer than t.

The full likelihood statements incorporating fully observed, right censored, and left censored observations are then

$$\begin{aligned}
\mathcal{L} &\propto \prod_{i \in C} \text{Exponential}(y_i|\lambda) \prod_{j \in R} S(y_j|\lambda) \prod_{k \in L} (1 - S(y_k|\lambda)) \\
\mathcal{L} &\propto \prod_{i \in C} \text{Weibull}(y_i|\alpha, \sigma) \prod_{j \in R} S(y_j|\alpha, \sigma) \prod_{k \in L} (1 - S(y_k|\alpha, \sigma))
\end{aligned} \tag{S13}$$

where C is the set of all fully observed taxa, R the set of all right censored taxa, and L the set of all left-censored taxa.

D Widely applicable information criterion

WAIC can be considered a fully Bayesian alternative to the Akaike information criterion, where WAIC acts as an approximation of leave-one-out cross-validation which acts as a measure of out-of-sample predictive accuracy

[5]. WAIC is calculated starting with the log pointwise posterior predictive density calculated as

$$\text{lppd} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\Theta^S) \right), \tag{S14}$$

where n is sample size, S is the number posterior simulation draws, and Θ represents all of the estimated parameters of the model. This is similar to calculating the likelihood of each observation given the entire posterior. A correction for the effective number of parameters is then added to lppd to adjust for overfitting. The effective number of parameters is calculated, following the recommendations of Gelman et al. [5], as

$$p_{\text{WAIC}} = \sum_{i=1}^{n} V_{s=1}^{S} (\log p(y_i|\Theta^S)). \tag{S15}$$

where V is the sample posterior variance of the log predictive density for each data point.

Given both equations S14 and S15, WAIC is then calculated

$$\text{WAIC} = \text{lppd} - p_{\text{WAIC}}. \tag{S16}$$

When comparing two or more models, lower WAIC values indicate better out-of-sample predictive accuracy. Importantly, WAIC is just one way of comparing models. When combined with posterior predictive checks it is possible to get a more complete understanding of a model's fit to the data.

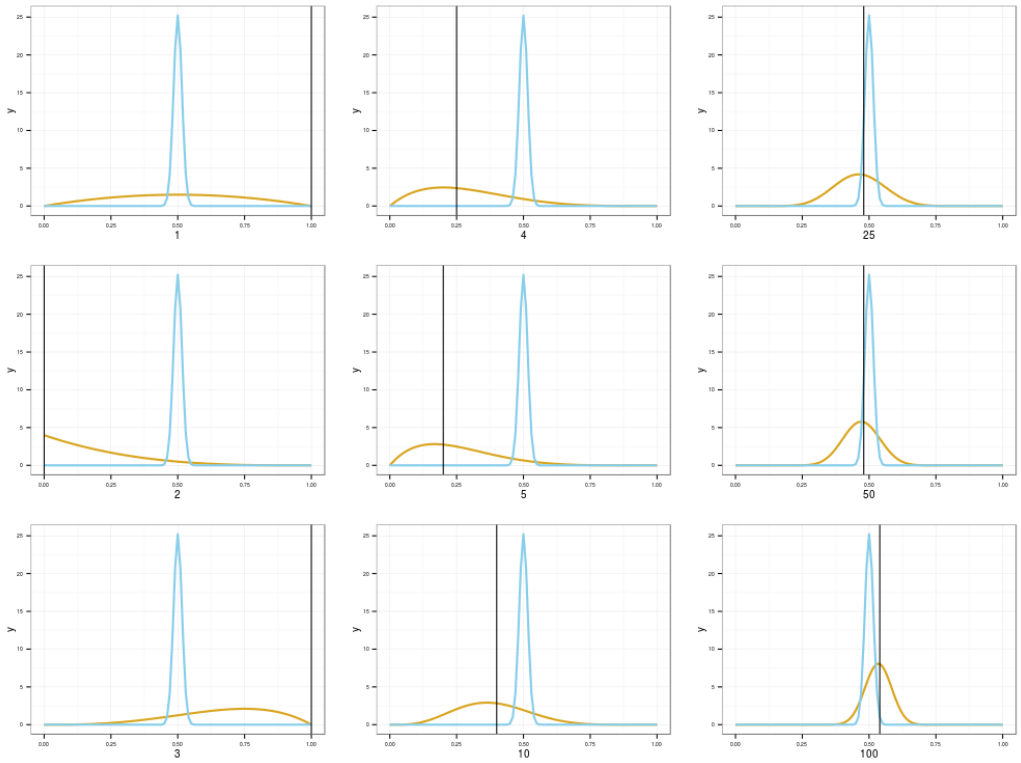

Figure S2: Comparisons of the underlying distribution (blue) to posterior estimates based on increasing sample size (gold). Each posterior estimate is represented for only a single realization of draws, each with sample size indicated as the x-axis label (increasing in column major order). Black vertical lines correspond to the MAP estimate of the simulated taxon's affinity. This stands in contrast to the posterior distribution of expected affinity in gold.

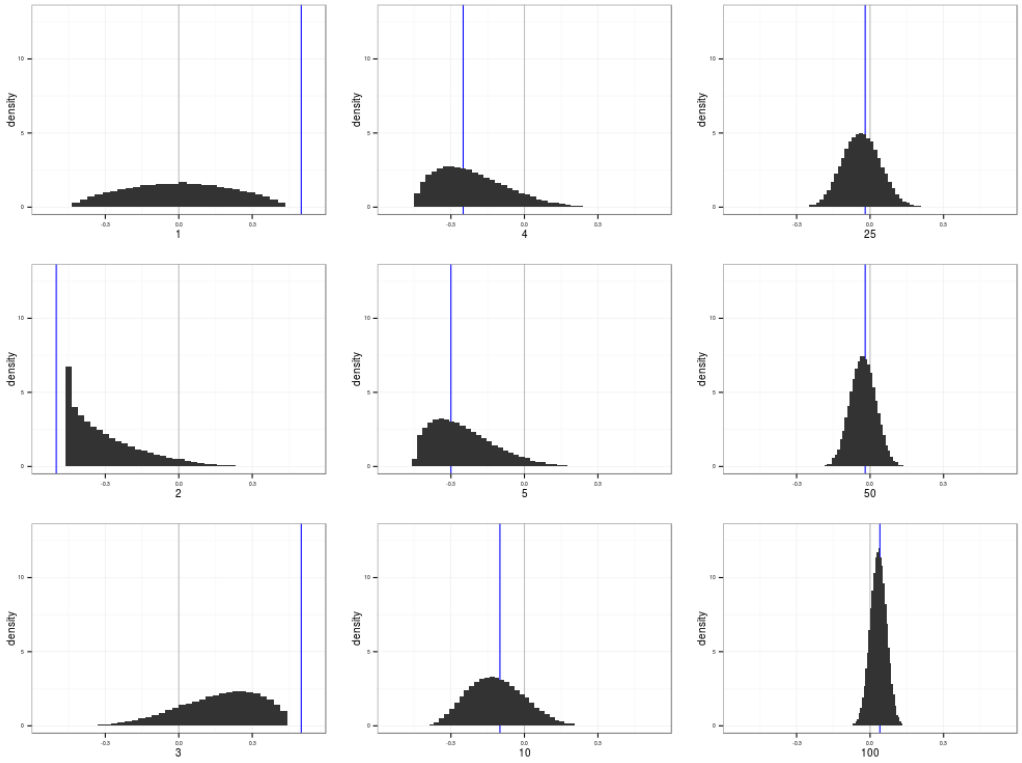

Figure S3: Histograms of the difference in the underlying occurrence distribution and the posterior distribution estimates from the previous graph (Fig. S2). The "true" value is included in all distributions of environmental affinities. Each affinity estimate is represented for only a single realization of draws, each with sample size indicated as the x-axis label (increasing in column major order). Blue vertical lines correspond to the difference in MAP estimates between the underlying distribution and the simulated taxon's draws. This stands in contrast to the distribution of the differences between the simulated taxon and background.

References

[1] Carvalho, C. M., N. G. Polson, and J. G. Scott, 2009. Handling Sparsity via the Horseshoe. *in* Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, vol. 5, Pp. 73–80.

[2] ———, 2010. The horseshoe estimator for sparse signals. Biometrika 97:465–480.

[3] Foote, M., 2006. Substrate affinity and diversity dynamics of Paleozoic marine animals. Paleobiology 32:345–366.

[4] Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1:515–533.

[5] Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, 2013. Bayesian data analysis. 3 ed. Chapman and Hall, Boca Raton, FL.

[6] Gelman, A. and J. Hill, 2007. Data Analysis using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York, NY.

[7] Ibrahim, J. G., M.-H. Chen, and D. Sinha, 2001. Bayesian Survival Analysis. Springer, New York.

[8] Kiessling, W. and M. Aberhan, 2007. Environmental determinants of marine benthic biodiversity dynamics through Triassic Jurassic time. Paleobiology 33:414–434.

[9] Klein, J. P. and M. L. Moeschberger, 2003. Survival Analysis: Techniques for Censored and Truncated Data. 2nd ed. Springer, New York.

[10] Miller, A. I. and S. R. Connolly, 2001. Substrate affinities of higher taxa and the Ordovician Radiation. Paleobiology 27:768–778.

[11] Simpson, C. and P. G. Harnik, 2009. Assessing the role of abundance in marine bivalve extinction over the post-Paleozoic. Paleobiology 35:631–647.

[12] Stan Development Team, 2014. Stan Modeling Language Users Guide and Reference Manual, Version 2.5.0. URL `http://mc-stan.org/`.