Title: Known unknowns and model selection in ecological evidence synthesis

Author: Shane A. Blowes, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany; Department of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Salle), Germany. shane.blowes@idiv.de

Corresponding author: Shane Blowes, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, Leipzig 04103, Germany. shane.blowes@idiv.de.

Open research: All data and code used in this work are in the public domain. Code to reproduce results are available at: https://github.com/sablowes/heterogeneity-evidence-synthesis, and will be archived upon acceptance.

Key words: meta-analysis, quantitative evidence synthesis, heterogeneity, cross-validation, model selection, heteroscedasticity, biodiversity, scale-dependence.

Abstract (200 word limit)

Quantitative evidence synthesis is a prominent path towards generality in ecology. Generality is typically discussed in terms of central tendencies, such as an average effect across a compilation of studies, and the role of heterogeneity for advancing understanding is not as well developed. Heterogeneity quantifies the consistency of ecological effects across studies, and between-study heterogeneity is typically assumed as constant. Here, I relax the assumption of constant heterogeneity, and show how cross validation can further the generality goals of quantitative evidence synthesis. First, I examine scale-dependent heterogeneity for a meta-analysis of plant native-exotic species richness relationships, quantifying the relationships among unexplained effect size variation, spatial grain and extent. Second, I examine relationships among patch size, study-level covariates and unexplained variation in species richness using a database of fragmentation studies. Both case studies show that relaxing the assumption of constant unexplained variation can provide a more detailed description of available evidence - for example, where effects can be transferred with more or less certainty. Moreover, cross validation shows for both case studies that assuming constant heterogeneity can constrain model predictive performance, and in particular, the ability of model predictions to generalise to new studies.

Introduction

Quantitative evidence synthesis aims for general insights into the direction, magnitude and variability of ecological effects (Gurevitch et al. 2018, Fox 2019, Spake et al. 2022). Two common forms of quantitative synthesis are meta-analysis (i.e., analyses of effect sizes collated or calculated from existing studies), and analyses of raw data compilations (referred to as full data analysis; Spake et al. 2022). For both approaches, generalities typically come in the form of central tendencies, such as an average effect size across studies. While both approaches

frequently quantify the heterogeneity of effect sizes across studies to examine how consistent ecological effects are (Senior et al. 2016, Gurevitch et al. 2018, Spake et al. 2022, Nakagawa et al. 2023a), any role for heterogeneity in assessing generality is less well developed. Critically, quantitative syntheses typically assume either constant between-study heterogeneity (meta-analysis) or residual variation (i.e., homoscedasticity for full data analyses). This assumption is equivalent to assuming that included studies sample effect sizes from a single population (Williams et al. 2021). Despite the rapid rise of meta-analysis in ecology and conservation biology (Gurevitch et al. 2018, Anderson et al. 2021), neither the consequences of this assumption, nor the opportunities available when it is relaxed, have received much attention from quantitative ecologists doing evidence synthesis.

Ecological meta-analyses often fit "random effects" or multilevel meta-analytic models (Gurevitch et al. 2018, Nakagawa et al. 2023a). These models incorporate two forms of uncertainty. First, sampling error or variance is represented by uncertainty in the effect size (i.e., response variable), and assumed known (with methods available for inferring missing values for some commonly used effect sizes, e.g., Nakagawa et al. 2023b). The second form of uncertainty, between-study heterogeneity or variance, is typically modelled as a varying (or random) intercept, and assumed to be constant across studies; i.e., estimated with a single standard deviation across all studies. Multilevel meta-analytic models estimate additional within study heterogeneity (via an additional varying intercept), and are recommended when included studies provide multiple (non-independent) effect sizes (Nakagawa et al. 2023a). Recently introduced (location-scale) statistical models for meta-analysis relax the assumption of constant between-study uncertainty with an extension to allow predictors of heterogeneity (i.e., of the standard deviation of the varying intercepts; Williams et al. 2021, Viechtbauer & Lopez-Lopez 2022,

Rodriguez et al. 2023). Simulations show little is lost in terms of power and type-I error rates (false rejection of the null hypothesis) when the assumption of constant between-study variances is relaxed (Rodriguez et al. 2023). Evidence syntheses using full data compilations typically make qualitatively similar assumptions: effects are allowed to vary between studies (e.g., varying slopes and intercepts), and most often estimated assuming constant residual variation across studies (i.e., homoscedasticity). Importantly, for both meta-analyses (Williams et al. 2021) and full data analyses (Pinherio & Bates 2000, Zuur et al. 2009), relaxing the assumption of constant heterogeneity or heteroscedasticity can impact estimates of other parameters, such as the overall mean or study-level effects.

Examining unexplained variation can open up new research questions (Williams et al. 2021, Viechtbauer & Lopez-Lopez 2022). Quantifying known unknowns - for example, by using covariates to describe patterns in unexplained variation – represents an extension to the majority of existing ecological evidence syntheses, and can help identify promising directions for both empirical and theoretical research. For example, covariates strongly associated with unexplained variation could also help to more clearly define limits to the generality of quantitative evidence syntheses (Spake et al. 2022). If different groups, such as taxon groups or geographic regions are associated with more or less unexplained variation, researchers could better communicate where effect sizes can be transferred with more or less uncertainty.

To compare models assuming homogeneous versus heterogeneous unexplained variation I use cross validation. To date, model selection in evidence synthesis has largely focused on the amount of variance explained (e.g., R^2, Nakagawa et al. 2023a), or on predictive performance assessed with information criteria (Cinar et al. 2021). R^2 describes how much variation is accounted for by different model components (e.g., meta-regression, or hierarchical partitions of

heterogeneity), and allow for comparisons of the total amount of variation explained by different covariates or model components. In contrast, cross validation and information criteria assess model predictive performance (Hastie et al. 2009, Cinar et al. 2021). However, explicit consideration of the data used to train versus test models (Hastie et al. 2009), and the need to define a function to be optimized (Hastie et al. 2009, Yates et al. 2022) make cross validation more flexible than information criteria. For example, data splitting can be designed to compare the ability of models to make different types of predictions (e.g., Merkle et al. 2019, Yates et al. 2022). In the context of quantitative evidence synthesis, I consider two types of predictions: (1) to a new datum or data within the existing studies; or (2), for a new study, i.e., data outside the sample of studies used for model fitting.

I show how location-scale models and cross-validation for model selection can be used in quantitative evidence synthesis for ecology using two biodiversity synthesis case studies. First, I extend a recent meta-analysis of spatial scale-dependence in plant native-exotic species richness relationships (Peng et al. 2019) to quantify relationships among unexplained variation, grain size and spatial extent. The second case study uses a data compilation of habitat fragment diversity studies (Chase et al. 2019). I focus on the relationship between fragment size and local (i.e., patch-scale) species richness, examine whether residual variation is related to fragment size and other study-level predictors. For both case studies, the assumption that unexplained variation comes from a single population limits model predictive performance, particularly when making predictions to new studies.

Cross validation for model comparison in evidence synthesis

Cross validation uses data splitting techniques to test model predictive performance. Models are fit to a ‘training’ data set and assessed on their ability to predict the ‘test’ data set

(Hastie et al 2009, Yates et al. 2022). To quantify predictive performance, I used pointwise *expected log predictive density* (*elpd*; see Yates et al. 2022 for further discussion of other common loss functions). To minimize an inherent bias to support overfit models, I used the modified one-standard-error rule (Yates et al. 2021) that aims to select the least complex model that is comparable to the best scoring model. I evaluate model performance for making two different types of predictions (Merkle et al. 2019, Yates et al. 2022): (i) conditional predictions that examine model performance when making predictions to existing studies, and (ii) marginal predictions that assess predictive performance to new studies. Conditional predictions are qualitatively similar (and quantitively the same for some loss functions, Yates et al. 2022) to those made by other information criteria methods (e.g., AIC). However, if generality is considered as the focus for a quantitative evidence synthesis, it might be more valuable for models to make good predictions to new studies (i.e., outside the sample of studies used to parameterize the model). To illustrate this, I use model selection based on comparisons of marginal predictions for a left-out study as an example in the two case studies.

I assess conditional and marginal predictions using two different data splitting techniques. For conditional predictions, approximate or exact leave-one-out (loo) cross validation is considered the gold standard (Yates et al. 2022). However, approximate loo diagnostics (Vehtari et al. 2017) showed many influential observations for the models I fit in the case studies (which reduces the reliability of performance estimates, Vehtari et al. 2017), and because exact loo can be computationally expensive for large data compilations (due to the need to refit the model as many times as there are data points), I approximated loo using *k*-fold cross validation. *k*-fold cross validation converges on loo when the number of blocks (*k*) into which the data are split equals the number of data points. The majority of quantitative evidence syntheses,

including the case studies presented here, use mixed or multilevel models, and so I use a stratified *k*-fold approximation of loo. Stratified *k*-fold cross validation balances data splitting among subgroups when creating the *k* blocks of data to which models are refit, ensuring that the relative group frequencies (here, the data coming from individual studies included in my analyses) are maintained. Similar to loo (and other information criteria such as AIC), *k*-fold cross validation is focused on making predictions to new data points within existing studies conditional on the model parameters. However, the most common goals of evidence synthesis – intervention evaluation and generality (Gurevitch et al. 2018) - might also be assessed by examining how well models make predictions to new studies. To assess the predictive performance of models to new studies, I use leave-one-group-out cross validation, where individual studies are removed one at a time, and models are assessed on their ability to predict the held-out data.

Models and model fitting

Location-scale statistical models allowing predictor variables for both the mean (location) and residual variation (scale) are well studied (Pinheiro & Bates 2000 and references therein). To relax the assumption of constant heterogeneity, recent meta-analytic models fit predictors for the standard deviation of varying (random) intercepts (e.g., Williams et al. 2021, Rodriguez et al. 2022, Viechtbauer & Lopez-Lopez 2022), covariates that appear as predictors for the location parameter. These models admit only the known sampling error associated with individual effect sizes to contribute to the scale component of the likelihood. Here, I describe an alternative location-scale model for meta-analysis, where unexplained variation is modelled by covariates for the scale parameter (that are added to the known error variances; see model 1.2 - 1.4 below; a brief comparison of the two approaches shows that they have very similar parameter estimates,

but differ slightly in the model comparisons for the meta-analytic case study presented here: Appendix S1). Questions being addressed with analyses of full data compilations can also model heteroscedasticity in multiple ways. Models can be fit with predictors for the scale parameter (e.g., when using Gaussian or log-normal likelihoods), or for parameters related to the scale (e.g., negative binomial or beta distributions). Many R packages commonly used by ecologists allow for mixed (or multilevel) location-scale models to be fit using either frequentist (e.g., *nlme*, Pinheiro & Bates 2024; *glmmTMB*, Brooks et al. 2017) or Bayesian methods (e.g., *brms*, Bürkner 2017), which I adopt for the case-studies presented here.

To robustly build increasingly complex models of heterogeneity and heteroscedasticity I used simulation-based calibration (Gelman et al. 2018, Talts et al. 2020, Modrák et al. 2023). This process uses simulated data to check model calibration, and I focus on whether models can recover known parameter values using simulated data sets of the same size, shape and structure as that found in each of the case studies. Simulation-based calibration requires a data generating model, and I followed to same procedure for each model in both case studies. First, the model is fit to the empirical (i.e., real) data. Then, the same model with priors informed by the fit to the empirical data is used to simulate many new data sets (with the same size, shape and structure as the empirical data). Then, based on the results of the calibration (i.e., are known parameters of interest recovered with reasonable coverage?), I proceed with model building towards the goal of models that characterize heterogeneity and heteroscedasticity. All models were fit using the Hamiltonian Monte Carlo (HMC) sampler Stan (Carpenter et al. 2017) due to the availability of tools for diagnosing model pathologies (Betancourt 2016, Monnahan et al. 2017), as well as for cross validation (Vehtari et al. 2017). Models fit to empirical data used weakly regularizing priors, four chains each with 4000 iterations, with 2000 used for warmup (details for case study

one in Appendix S2, and case study two Appendix S3). Models were coded using the 'brms' package (Bürkner 2017). Code for all analyses is available at https://github.com/sablowes/heterogeneity-evidence-synthesis (that will be archived following acceptance).

Case study one: Scale-dependent uncertainty in plant native-exotic richness relationships

Communities with more species are often thought to be more resistant to invasion by exotic (or non-native) species than communities with fewer species (Elton 1958), an idea referred to as the 'biotic resistance hypothesis'. However, the spatial scale-dependence of biodiversity complicates any straightforward interpretation of this hypothesis. Negative relationships between the numbers of native and non-native species are often thought more likely at small scales, with a switch to positive relationships expected at larger spatial scales (Levine 2000). Such scale-dependence has been linked to niche opportunities (Shea & Chesson 2002), whereby at relatively small scales, more species rich native communities leave fewer niche opportunities for exotics to invade. As spatial scale increases, so too does heterogeneity in resources, natural enemies and the physical environment, which creates greater niche availability for exotic species to become established (Shea & Chesson 2002).

Recently, Peng et al. (2019) used multilevel meta-regressions to synthesize evidence for relationships between spatial scale (both grain and extent) and the correlation of native and exotic species richness. Using 101 observational (i.e., unmanipulated, non-experimental) studies and 204 effect sizes, Peng et al. (2019) found that the average effect size was positive, and that the positive correlations between native and exotic species richness got stronger with increasing (log) grain size. No relationship was found between spatial extent and the native-exotic richness relationship. All of the models examined by Peng et al. (2019) assumed between-study variation

in the effect size associated with the native-exotic richness relationship, and adjusted for the non-independence of multiple effect sizes coming from some studies with (nested) multilevel varying (random) intercepts. Here I relax the assumption that between study heterogeneity is constant, and examine whether unexplained variation in the effect size is related to either grain size or extent. I compare how all models perform when making predictions to existing and new studies using cross validation.

Peng et al. (2019) transformed all correlations to Pearson product moment correlation coefficient (*r*), and calculated effect sizes using the Fisher *z*-transformation: $z = \frac{1}{2} log\left(\frac{1+r}{1-r}\right)$, with variance estimated as $Var(z) = \frac{1}{n-3}$, *log* refers to the natural logarithm (as it does hereafter), and *n* is the sample size (i.e., number of native-exotic species pairs; effect size data sourced from Peng et al. 2019). I start by reproducing the main result from Peng et al. (2019) using a multilevel linear meta-regression model for grain size that assumes the effect sizes, z_{ij}, are normally distributed with known within-case variance (s_{ij}^2) and constant between-study variance (τ^2), which can be expressed as following:

$$z_{ij} \sim N(\mu_{ij}, s_{ij}),$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + \beta_{0ij} + \beta_1 X_{ij},$$

$$\beta_{0i} \sim N(0, \tau),$$

$$\beta_{0ij} \sim N(0, \omega), \qquad \text{(Model 1.1)}$$

where cases (*j*) are nested within studies (*i*), and have among-case (within study) variance, ω^2; study-level heterogeneity, β_{0i}, has variance τ^2, and varies around the overall linear relationship with intercept, β_0, slope β_1, and predictor X_i (here the natural logarithm of grain size in study *i*). In this model, both τ and ω quantify variation not accounted for by the population (fixed) linear

predictor ($\beta_0 + B_1 X_{ij}$), and are sometimes referred to residual heterogeneity and interpreted as residuals (e.g., Viechtbauer & Lopez-Lopez 2022, Nakagawa et al. 2023a). Given the strong relationship between effect size and grain size (Peng et al. 2019), I retain (log) grain size as a predictor in all subsequent models.

Peng et al. (2019) observed that variation around the average (overall positive) relationship between grain size and effect sizes decreased with increasing grain size. That is heterogeneity decreased as a function of grain size, though they did not explicitly model this relationship. Indeed, the most widely used statistical computing software for meta-analysis has only recently been extended to allow predictors for residual heteroscedasticity (Viechtbauer & Lopez-Lopez 2022). Here I start by fitting a model that estimates additional parameters for residual variation (i.e., the scale parameter of the likelihood) from the data. Recall that s_{ij} is the known within-study standard deviation of the effect size estimate, and unexplained variation in existing meta-analytic location-scale models include predictors for a varying intercept (e.g., τ or ω in model 1.1). An alternative is to include another parameter to be estimated from the data for residual variation, and I fit a model where an additional parameter (σ_{ij}) for residual variation is a function of grain size:

$$z_{ij} \sim N(\mu_{ij}, s_{ij} + \sigma_{ij}),$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + \beta_1 X_i,$$

$$\beta_{0i} \sim N(0, \tau),$$

$$log(\sigma_{ij}) = \beta_0^{\sigma} + \beta_1^{\sigma} X_{ij}, \qquad \text{(Model 1.2)}$$

where X_{ij} is grain size on a log-scale for the jth case in study i. I note that this model did not produce reliable estimates when I included the case-level random intercept (ω), likely due to

identification problems associated with multiple parameters (i.e., σ and ω) describing the same level of variation in the data, and the case-level random intercept was dropped from the model.

To model extent, Peng et al. (2019) discretized spatial extent into bins with a range of one order of magnitude in km^2, i.e., (0, 10), [10, 100), [10^2, 10^3), …, [10^6,). Peng et al. (2019) fit multiple meta-regression models to examine for an effect of spatial extent on effect sizes, including models with an interaction between grain and extent. While Peng et al. (2019) did not report a strong influence of spatial extent on effect sizes, they did describe a complex interaction between extent and grain. Here, I examine whether unexplained variation (i.e., after adjusting for the relationship between effect size and grain size) was related to extent using a model where residual variation was a function of spatial extent. I fit extent as a categorical predictor of the residual variation (σ_{ij}, i.e., same as previous model, but without the intercept (β_0^{ω}), and extent categories coded as indicator variables for the predictor, X_{ij} [Model 1.3]).

Finally, I fit a model that allowed study-level residual variation. I note that where data are limited to one effect size per study this model will have identification problems. Here, the data have some replication within studies (mean = 2; median = 1; range = 1-17 effect sizes per study), and simulation-based calibration suggested reasonable coverage for parameters in the order of those found for the empirical data with the observed sample sizes (Appendix S2: Fig. Sx). Again, I fit a model fit with (log) grain size as a predictor for the mean effect size, and estimated residual variation with studies varying around an overall average:

$$z_{ij} \sim N(\mu_{ij}, s_{ij} + \sigma_{ij}),$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + \beta_1 X_i,$$

$$\beta_{0i} \sim N(0, \tau),$$

$$log(\sigma_{ij}) = \beta_0^{\sigma} + \beta_{0i}^{\sigma},$$

$$\beta_{0i}^{\sigma} \sim N(0, \zeta), \qquad \text{(Model 1.4)}$$

where β_0^{σ} is the average residual variation (on a log-scale), and β_{0i}^{σ} is a normally distributed study-level varying intercept with zero mean and ζ standard deviation.

Case study one: results and discussion

All models were able to reproduce the observed data well, and simulation-based calibration showed that all models had reasonable coverage for parameters of interest (Appendix S2: Figs. S1-S16). Model selection identified the same model (1.4) as best for making predictions within existing studies (Fig. 1a) and to new studies (Fig. 1b), though subsequent ordering among the other models depended on predictive task (Fig. 1). All models produced similar estimates for the parameters they shared (i.e., β_0, β_1, τ; Appendix S1: Fig. S3), and support for model 1.4 is notable because this model is atypical of meta-analyses. The more widely used constant heterogeneity meta-regression (model 1.1) performed worst for making predictions within the observed data (Fig. 1a), and was tied for second (with all other models) when making predictions for a new study (Fig. 1b). On average, unexplained variation decreased with increasing grain size (as was observed qualitatively by Peng et al. 2019), though there is considerable uncertainty surrounding the relationship for the smallest and largest grains (Fig. 1c), and the 90% credible interval overlapped zero (model 1.2: β_1^{σ}= -0.08 – 0.015). Finally, the model that included spatial extent (i.e., the total geographic area covered by the samples) showed that after adjusting for the positive relationship between effect sizes and grain size, unexplained variation was on average the greatest at the smallest extents (Fig. 1d).

Case study two: Variation in patch-scale effects of ecosystem decay is largest for the smallest patches

Habitat loss is a major driver of biodiversity loss (Diaz et al. 2019). As habitat is removed from landscapes, biodiversity reductions in remaining habitat fragments can be thought of as arising from one of two processes (Chase et al. 2020). First, because fewer individuals and species can exist in smaller habitat fragments, the species found small habitat fragments might simply be a random (or passive) sample of those found in large habitat fragments. Alternatively, if smaller habitat fragments negatively impact demographic rates relative to larger habitat fragments, then ecosystem decay might result in greater diversity declines compared to those expected from a solely passive sampling process (Chase et al. 2020).

To test these competing hypotheses for (patch-scale) diversity fragment size relationships, Chase et al. (2019) compiled assemblage data (counts of individuals of each species) in effort-standardized (or standardizable) samples across habitat fragments of different sizes within landscapes. Across 123 studies comprising 1509 habitat fragments, neither the number of individuals nor the species found in small fragments were simply a random sample of those in large habitat fragments (Chase et al. 2020). Instead, altered demographic rates (e.g., due to edge effects, reduced dispersal or demographic stochasticity, and collectively referred to as ecosystem decay) in smaller habitat fragments, resulted on average in fewer individuals, fewer species, and less even communities than expected from a passive sample of larger fragments (Chase et al. 2020). Here, I revisit this analysis to examine for heteroscedasticity, and to see whether modelling heteroscedasticity impacts model predictive performance. For simplicity, I focus on species richness, and start by refitting the same multilevel model used by Chase et al. (2020) that assumed residual variation was constant across all of the studies. I then fit a series of models of increasing complexity that relax the assumption of homoscedasticity.

The model fit by Chase et al. (2020) to effort-standardized for species richness (S_{ij}) in fragment j from study i took the form:

$$S_{ij} \sim lognormal(\mu_{ij}, \sigma)$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i})X_{ij},$$

$$[\beta_{0i}, \beta_{1i}]' \sim MVN(0, \boldsymbol{SRS}),$$

$$\boldsymbol{S} = \begin{bmatrix} \sigma_{0i} & 0 \\ 0 & \sigma_{1i} \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}\sigma_{1i}} \\ \rho_{\sigma_{0i}\sigma_{1i}} & 1 \end{bmatrix}, \qquad \text{(Model 2.1)}$$

where X_{ij} is the fragment size on a (natural) log-scale, which was centered by subtracting the overall mean from each observation before modelling; β_{0i} and β_{1i} are study-level departures from the overall intercept and slope, respectively, drawn from a multivariate normal (*MVN*) distribution with standard deviations, σ_{0i} and σ_{1i}, that allowed for correlations ($\boldsymbol{R}$ matrix) between the varying intercepts and slopes.

The first heteroscedastic model is motivated similarly to the varying study-level intercepts and slopes. Just as average richness and its relationship with fragment size might be expected to varying among studies around an overall average, so to might residual variation. Therefore, I allowed varying study-level residuals around an overall average residual standard deviation. First, I estimated study-level residual variation independently of other varying study-level parameters for the mean:

$$S_{ij} \sim lognormal(\mu_{ij}, \sigma_i)$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i})X_{ij},$$

$$[\beta_{0i}, \beta_{1i}]' \sim MVN(0, \boldsymbol{SRS}),$$

$$\boldsymbol{S} = \begin{bmatrix} \sigma_{0i} & 0 \\ 0 & \sigma_{1i} \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}\sigma_{1i}} \\ \rho_{\sigma_{0i}\sigma_{1i}} & 1 \end{bmatrix},$$

$$log(\sigma_i) = \beta_0^{\sigma} + \beta_{0i}^{\sigma},$$

$$\beta_{0i}^{\sigma} \sim N(0, \sigma_{0i}^{\sigma}), \qquad \text{(Model 2.2)}$$

where β_0^{σ} is the overall average of residual variation (on a log-scale), and β_{0i}^{σ} are the varying study-level departures (for the residual variation) drawn from a normal distribution with zero mean and standard deviation σ_{0i}^{σ}.

Next, I extend this model to include (log) fragment size as a predictor of residual variation, and again allow it to varying between studies independently of varying study-level parameters for the mean:

$$S_{ij} \sim lognormal(\mu_{ij}, \sigma_i)$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i})X_{ij},$$

$$[\beta_{0i}, \beta_{1i}]' \sim MVN(0, \boldsymbol{SRS}),$$

$$\boldsymbol{S} = \begin{bmatrix} \sigma_{0i} & 0 \\ 0 & \sigma_{1i} \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}\sigma_{1i}} \\ \rho_{\sigma_{0i}\sigma_{1i}} & 1 \end{bmatrix},$$

$$log(\sigma_i) = \beta_0^{\sigma} + \beta_{0i}^{\sigma} + (\beta_1^{\sigma} + \beta_{1i}^{\sigma})X_i,$$

$$[\beta_{0i}^{\sigma}, \beta_{1i}^{\sigma}]' \sim MVN(0, \boldsymbol{S^{\sigma}R^{\sigma}S^{\sigma}}),$$

$$\boldsymbol{S^{\sigma}} = \begin{bmatrix} \sigma_{0i}^{\sigma} & 0 \\ 0 & \sigma_{1i}^{\sigma} \end{bmatrix},$$

$$\boldsymbol{R^{\sigma}} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}^{\sigma}\sigma_{1i}^{\sigma}} \\ \rho_{\sigma_{0i}^{\sigma}\sigma_{1i}^{\sigma}} & 1 \end{bmatrix}, \qquad \text{(Model 2.3)}$$

where β_0^σ is the overall average residual variation, and β_1^σ is the overall average slope of residual variation with fragment size; β_{0i}^σ and β_{1i}^σ are the varying study-level departures from the intercept and slope, respectively, and were drawn from a multivariate normal distribution with zero mean and standard deviation σ_{0i}^σ and σ_{1i}^σ, with correlations estimated in matrix $\boldsymbol{R^\sigma}$.

Finally, I estimate models with study-level residual variation with and without fragment size as a predictor that allow for correlations between varying (study-level) parameters for the mean (μ) and residual variation (σ). The model without fragment size is:

$$S_{ij} \sim lognormal(\mu_{ij}, \sigma_i)$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i})X_{ij},$$

$$log(\sigma_i) = \beta_0^\sigma + \beta_{0i}^\sigma,$$

$$[\beta_{0i}, \beta_{1i}, \beta_{0i}^\sigma]' \sim MVN(0, \boldsymbol{SRS}),$$

$$\boldsymbol{S} = \begin{bmatrix} \sigma_{0i} & 0 & 0 \\ 0 & \sigma_{1i} & 0 \\ 0 & 0 & \sigma_{0i}^\sigma \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}\sigma_{1i}} & \rho_{\sigma_{0i}\sigma_{0i}^\sigma} \\ \rho_{\sigma_{0i}\sigma_{1i}} & 1 & \rho_{\sigma_{1i}\sigma_{0i}^\sigma} \\ \rho_{\sigma_{0i}\sigma_{0i}^\sigma} & \rho_{\sigma_{1i}\sigma_{0i}^\sigma} & 1 \end{bmatrix}. \qquad \text{(Model 2.4)}$$

The model with fragment size as a predictor of both the mean and residuals that allows for correlations among the varying study-level parameters is:

$$S_{ij} \sim lognormal(\mu_{ij}, \sigma_i)$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i})X_{ij},$$

$$log(\sigma_i) = \beta_0^\sigma + \beta_{0i}^\sigma + (\beta_1^\sigma + \beta_{1i}^\sigma)X_i,$$

$$[\beta_{0i}, \beta_{1i}, \beta_{0i}^\sigma, \beta_{1i}^\sigma]' \sim MVN(0, \boldsymbol{SRS}),$$

$$\boldsymbol{S} = \begin{bmatrix} \sigma_{0i} & 0 & 0 & 0 \\ 0 & \sigma_{1i} & 0 & 0 \\ 0 & 0 & \sigma_{0i}^{\sigma} & 0 \\ 0 & 0 & 0 & \sigma_{1i}^{\sigma} \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}\sigma_{1i}} & \rho_{\sigma_{0i}\sigma_{0i}^{\sigma}} & \rho_{\sigma_{0i}\sigma_{1i}^{\sigma}} \\ \rho_{\sigma_{0i}\sigma_{1i}} & 1 & \rho_{\sigma_{1i}\sigma_{0i}^{\sigma}} & \rho_{\sigma_{1i}\sigma_{1i}^{\sigma}} \\ \rho_{\sigma_{0i}\sigma_{0i}^{\sigma}} & \rho_{\sigma_{1i}\sigma_{0i}^{\sigma}} & 1 & \rho_{\sigma_{0i}^{\sigma}\sigma_{1i}^{\sigma}} \\ \rho_{\sigma_{0i}\sigma_{1i}^{\sigma}} & \rho_{\sigma_{1i}\sigma_{1i}^{\sigma}} & \rho_{\sigma_{0i}^{\sigma}\sigma_{1i}^{\sigma}} & 1 \end{bmatrix}. \qquad \text{(Model 2.5)}$$

Case study two: results and discussion

All models were able to reproduce the observed data well, and simulation-based calibration showed that all models had reasonable coverage for parameters of interest (Appendix S3: Fig. S1-S20). Models that allowed for heteroscedasticity outperformed the model with constant residual variation for making predictions of species richness in fragments for existing studies (Fig. 2a), and to new studies (Fig. 2b). Modelling heteroscedasticity, however, did not qualitatively impact the support for the ecosystem decay hypothesis (Chase et al. 2020) for species richness (Appendix S3: Fig. S21). Yet, the best model for predictions to existing studies differed from that favored for new studies. For making predictions to new data in existing studies, cross validation supported the most complex model (model 2.5; Fig. 2a); this model shows that residual variation was a decreasing function of fragment size (Fig. 2c), meaning unexplained variation in the fragment size-species richness relationship is greatest for small patches, and suggesting considerable potential for context-dependency in the response of taxonomic diversity in small habitat fragments. Moreover, this model suggests that there were some relationships between varying study-level effects and residual variation (Appendix S3: Fig. S22). For example, the fragment size study-level slopes for the mean effect and residual variation were negatively correlated (Appendix S3: Fig. S22a), meaning that the strongest ecosystem

386 decay effects were associated with the strongest decline in residual variation with increasing
387 fragment size (Appendix S3: Fig. S23).

388 For predictions to new studies leave-one-group-out cross validation supported a simpler,
389 multilevel model for residual variation (Fig. 2b), where study-level residuals varied around an
390 overall mean (model 2.2), meaning each study gets its own estimate of residual (unexplained)
391 variation. Because study encodes many other covariates (e.g., taxon group, matrix quality, time
392 since fragmentation), this model can be used to further examine for systematic variation in the
393 unexplained variation among studies by plotting posterior samples of study-level residual
394 variation (σ_i) against study-level covariates (see Chase et al. 2020 for a similar approach to
395 examining predictors of ecosystem decay effect size estimates). For example, studies of
396 amphibians and reptiles had more unexplained variation than average, while studies focused on
397 birds had less unexplained variation than average (Fig. 2d). The relatively small sample sizes (n
398 = 14 and n = 15 studies, respectively) precludes strong inference, and I highlight this result to
399 show how examining residual variation can provide a more complete description in a
400 quantitative evidence synthesis. For example, I might hypothesize less variation in connectivity
401 between patches for birds compared to amphibians and reptiles could drive this difference in
402 unexplained patch-scale diversity variation, and design either an empirical or theoretical study to
403 further examine the role of varying connectivity for fragmentation effects on patch-scale
404 diversity. More generally, this shows how explicit models of unexplained variation can help
405 communicate where the evidence for patch-scale fragmentation effects is more or less uncertain.

406 **Discussion**

407 When will it be reasonable to assume that studies compiled for evidence synthesis come
408 from single (statistical) population? And on what criteria should ecologists compare models for

evidence synthesis? Here, I relaxed the assumption that data compilations sampled a single statistical population for two case studies, and compared model performance for two different prediction goals. Statistical models that relaxed the assumption of homoscedastic residuals performed best for making both types of predictions - to new data within the included studies, and to unobserved data from a new study. Moreover, I found that model performance (i.e., the comparative rank of candidate models) can depend on whether predictions are being made to existing studies or to new studies, highlighting the flexibility of cross-validation for assessing models based on different goals in an evidence synthesis context.

Meta-analysis and quantitative syntheses of compiled data are a prominent path to generality in ecology and evolution (Gurevitch et al. 2018, Fox 2019, Anderson et al. 2021). While heterogeneity of effect sizes in meta-analyses are frequently quantified, they are often not reported (Senior et al. 2016, Nakagawa et al. 2023a), and the role of heterogeneity in assessing claims of generality is not well developed (Spake et al. 2022, Yang et al. 2023). Relaxing the assumption of constant heterogeneity in meta-analytical statistical models is relatively new (Williams et al. 2021, Rodriguez et al. 2022, Viechtbauer & Lopez-Lopez 2022), and here I examined an alternative to existing models with predictors for the scale parameter of the likelihood. For the meta-analysis re-examined here, this new model produced qualitatively similar estimates of the shared parameters with the other models examined (Appendix S1: Fig. S3). Moreover, the estimated relationship between unexplained variation and grain size was essentially indistinguishable between the model with the additional scale parameter (i.e., model 1.2 above) and a model comparable to existing models (Williams et al. 2021, Rodriguez et al. 2022, Viechtbauer & Lopez-Lopez 2022; Appendix S1: Fig. S2). The two models performed similarly for predictions to new data within studies (Appendix S1: Fig. S4a), but the new model

was favored for making predictions to a new study (Appendix S1: Fig. S4b). Indeed, among the models examined for each case study, the best predictions to a new study were made by models with very similar structures.

Quantitative evidence syntheses in ecology and evolution routinely estimate variation in effect sizes among studies (Gurevitch et al. 2018, Spake et al. 2024). However, unexplained variation is most often assumed as constant, and therefore, as coming from a single (statistical) population. In the frequently heterogeneous data that is compiled for evidence synthesis, this assumption warrants careful examination. Here, for predictions to new data from an unobserved study, both case studies favored a model where the scale (or residual variation) was modelled similarly to the varying effects for the location (mean). In effect, these models admit greater uncertainty, and this conservatism is likely what led to them being favored when making predictions to unobserved studies.

Models that relaxed the assumption of constant heterogeneity (or homoscedasticity) yielded new quantitative insights for both case studies presented here. First, I quantified the pattern of decreasing unexplained variation in the number native and exotic species with increasing grain size observed by Peng et al. (2019). I found support for the negative relationship on average, though the model revealed considerable uncertainty in the relationship (especially for large spatial grains where there is little data, Fig. 1c), and the 90% credible interval of the effect overlapped zero. Similar to Peng et al. (2019), I also found some evidence for an association between unexplained variation (i.e., after adjusting for grain size) and spatial extent, indicating that there is considerable context dependency in the relationship between numbers of native and exotic species, particularly at small spatial scales (i.e., small grains and small extents).

The second case study relaxed the assumption of homoscedasticity for the relationship between patch-scale (standardized) species richness and fragment size. I found a negative relationship between fragment size and residual variation, and that this negative relationship was strongest for landscapes that had the steepest (positive) relationship between standardized species richness and fragment size. Taken together, this means that variation in the patch-scale richness is greatest for the smallest patches in fragmented landscapes, and that unexplained variation decreases most strongly in landscapes with the strongest (patch-scale) ecosystem decay effects. In effect, relaxing the assumption of homoscedasticity has identified another empirical pattern that theory describing processes associated with fragmentation (e.g., Zhang et al. 2024) might explore. I also used a simpler heteroscedastic model with varying study-level residuals to show how quantitative descriptions of unexplained variation can reveal limitations to the generality of an evidence synthesis (e.g., do some groups have more unexplained variation than others), as well as providing promising new new paths for empirical and theoretical research (e.g., *why* do some groups have more or less unexplained variation than others?). Importantly, model selection in both case studies showed that relaxing the assumption of constant heterogeneity or homoscedasticity can result in models with greater predictive ability.

Model selection in evidence synthesis has typically used either variance explained (R^2, Nakagawa et al. 2023a) or information criterion methods (Cinar et al. 2021). Here, I showed how cross-validation can be used to compare model performance for different predictive tasks. Specifically, I compared models based on their ability to predict new data within existing studies (conditional predictions), and on their ability to predict data for a new (unobserved) study (refs). Cross-validation for conditional prediction (e.g., approximate or exact leave-one-out, and different k-fold approximations to leave-one-out) are similar to information criteria in their goal

of making predictions to new data within the existing studies. However, it is in the assessment of predictions to new studies (marginal predictions) that cross-validation can provide new insights into model predictive performance for quantitative evidence synthesis. Here I broadly focused on a new study in general, and defined individual studies as the group in the leave-one-group-out cross validation, but other choices could be made depending on a particular research question or goal for evidence synthesis. For example, limits beyond which effect size estimates or relationships become unreliable (sometime called the transfer domain, Spake et al. 2022) of an evidence synthesis could be assessed using cross-validation to partition studies into different groups, e.g., based on taxonomy, geography, biogeography, phylogeny, etc. Moreover, the ranking of models was prediction task-dependent for both case studies presented here (though the same model was best for both predictions in case study one), which suggests that using cross-validation to assess different prediction goals will be a useful addition to the evidence synthesis toolbox.

Author contributions

SAB conceived and did the study.

Acknowlegements

I thank the Biodiversity Synthesis lab for constructive discussions, and for providing much fodder for thinking about models for evidence synthesis. SAB was supported by the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, funded by the German Research Foundation (FZT 118), and by the European Union. Views and opinions expressed are however those of the author only, and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Conflict of interest

I declare no conflicts of interest.

References

Anderson SC, Elsen PR, Hughes BB, Tonietto RK, Bletz MC, Gill DA, et al. Trends in ecology and conservation over eight decades. Frontiers in Ecol & Environ. 2021 Jun;19(5):274–82.

Betancourt, M. 2016, April 3. Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo. arXiv.

Brooks ME, Kristensen K, Van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R journal. 2017;9(2):378–400.

Bürkner PC. brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software. 2017;80(1):1–28.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of statistical software. 2017;76(1).

Chase JM, Blowes SA, Knight TM, Gerstner K, May F. Ecosystem decay exacerbates biodiversity loss with habitat loss. Nature. 2020;584(7820):238–43.

Chase JM, Liebergesell M, Sagouis A, May F, Blowes SA, Berg \AAke, et al. FragSAD: A database of diversity and species abundance distributions from habitat fragments. Ecology. 2019;100(12).

Cinar, O., J. Umbanhowar, J. D. Hoeksema, and W. Viechtbauer. 2021. Using information-theoretic approaches for model selection in meta-analysis. Research Synthesis Methods 12:537–556.

Díaz S, Settele J, Brondízio ES, Ngo HT, Agard J, Arneth A, et al. Pervasive human-driven decline of life on Earth points to the need for transformative change. Science. 2019;366(6471).

Elton CS. The ecology of invasions by animals and plants. London, UK.: Methuen and Company; 1958.

Fox JW. The many roads to generality in ecology. Philosophical Topics. 2019;47(1):83–104.

Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature. 2018;555(7695):175–82.

Hastie, T., R. Tibshirani, and J. Friedman. (2009). The Elements of Statistical Learning. Second edition. Springer New York, NY.

Levine JM. Species Diversity and Biological Invasions: Relating Local Process to Community Pattern. Science. 2000 May 5;288(5467):852–4.

Merkle EC, Furr D, Rabe-Hesketh S. Bayesian Comparison of Latent Variable Models: Conditional Versus Marginal Likelihoods. Psychometrika. 2019 Sep;84(3):802–29.

Modrák, M., A. H. Moon, S. Kim, P. Bürkner, N. Huurre, K. Faltejsková, A. Gelman, and A. Vehtari. 2023. Simulation-based calibration checking for Bayesian computation: The choice of test quantities shapes sensitivity. Bayesian Analysis 1:1–28.

Monnahan, C. C., J. T. Thorson, and T. A. Branch. 2017. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. Methods in Ecology and Evolution 8:339–348.

Nakagawa, S., Y. Yang, E. L. Macartney, R. Spake, and M. Lagisz. 2023a. Quantitative evidence synthesis: a practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. Environmental Evidence 12:8.

Nakagawa, S., D. W. A. Noble, M. Lagisz, R. Spake, W. Viechtbauer, and A. M. Senior. 2023b. A robust and readily implementable method for the meta-analysis of response ratios with and without missing standard deviations. Ecology Letters 26:232–244.

Peng S, Kinlock NL, Gurevitch J, Peng S. Correlation of native and exotic species richness: a global meta-analysis finds no invasion paradox across scales. Ecology. 2019 Jan;100(1):e02552.

Pinheiro J, Bates D. Mixed-effects models in S and S-PLUS [Internet]. Springer science & business media; 2006.

Rodriguez JE, Williams DR, Bürkner P. Heterogeneous heterogeneity by default: Testing categorical moderators in mixed-effects meta-analysis. Brit J Math & Statis. 2023 May;76(2):402–33.

Schmid C H, Mengersen K. Bayesian Meta-analysis. In: Koricheva J, Gurevitch J, Mengersen K, editors. Handbook of Meta-analysis in Ecology and Evolution [Internet]. Princeton University Press; 2013 p. 145–73.

Senior AM, Grueber CE, Kamiya T, Lagisz M, O'Dwyer K, Santos ESA, et al. Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. Ecology. 2016 Dec;97(12):3293–9.

Spake R, O'dea RE, Nakagawa S, Doncaster CP, Ryo M, Callaghan CT, et al. Improving quantitative synthesis to achieve generality in ecology. Nature Ecology & Evolution. 2022;6(12):1818–28.

Talts, S., M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. 2020, October 21. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv.

Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. 2017 Sep;27(5):1413–32.

Viechtbauer, W., and J. A. López-López. 2022. Location-scale models for meta-analysis. Research Synthesis Methods 13:697–715.

Williams DR, Rodriguez JE, Bürkner PC. Putting variation into variance: modeling between-study heterogeneity in meta-analysis. 2021.

Yang, Y., D. W. Noble, R. Spake, A. M. Senior, M. Lagisz, and S. Nakagawa. 2023. A pluralistic framework for measuring and stratifying heterogeneity in meta-analyses. ecoevoRxiv. https://doi.org/10.32942/X2RG7X

Yates LA, Aandahl Z, Richards SA, Brook BW. Cross validation for model selection: A review with examples from ecology. Ecological Monographs. 2023 Feb;93(1):e1557.

Zhang, H., J. M. Chase, and J. Liao. 2024. Habitat amount modulates biodiversity responses to fragmentation. Nature Ecology & Evolution 8:1437–1447.

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed effects models and extensions in ecology with R. Springer; 2009. 574 p.

Figure 1: **Heterogeneous heterogeneity outperforms constant heterogeneity for predicting plant native-exotic species richness relationships.** Model selection for predictions (a) within existing studies using stratified k-fold cross-validation, and (b) for new studies using leave-one-group-out cross validation. Scale-dependent unexplained variation for (c) grain, and (d) extent. Thick lines on (c) show median expectation for grain-size dependent unexplained variation from model 1.2 ($log\ [\sigma_{ij}] = \beta_0^{\sigma} + \beta_1^{\sigma} X_{ij}$); thin lines show 100 draws from the posterior distribution to visualise uncertainty. Violin plots (d) show the distribution of 1000 draws of the posterior distribution of unexplained variation (σ) for each of the extent categories; boxplots show the median (bar), 25% and 75% quantiles (box), whiskers show 1.5 times the interquartile range, points show observations beyond 1.5 times the interquartile range; n = number of effect sizes for each extent category (cases).

Figure 2: **Patch-scale species richness in habitat fragments is predicted best by models with heteroscedastic residual variation.** Model selection for predictions (a) within existing studies using stratified k-fold cross-validation, and (b) for predictions to new studies using leave-one-group-out cross validation; (c) fragment size-dependent residual variation, and (d) study-level residual variation grouped in taxon groups. Bold line on (c) shows median predicted average relationship between fragment size and residual variation from model 2.5, thin lines show 100 draws from the posterior distribution. Density plots (d) for taxon groups show study-level variation (1000 draws from the posterior) of residual variation from model 2.2 (i.e., $\beta_0^{\sigma} + \beta_{0i}^{\sigma}$ grouped into taxon groups); black triangle shows median σ for each taxon group, black dashed line and surrounding shading are the overall mean (β_0^{σ}) and 95% credible interval.

Figure 1

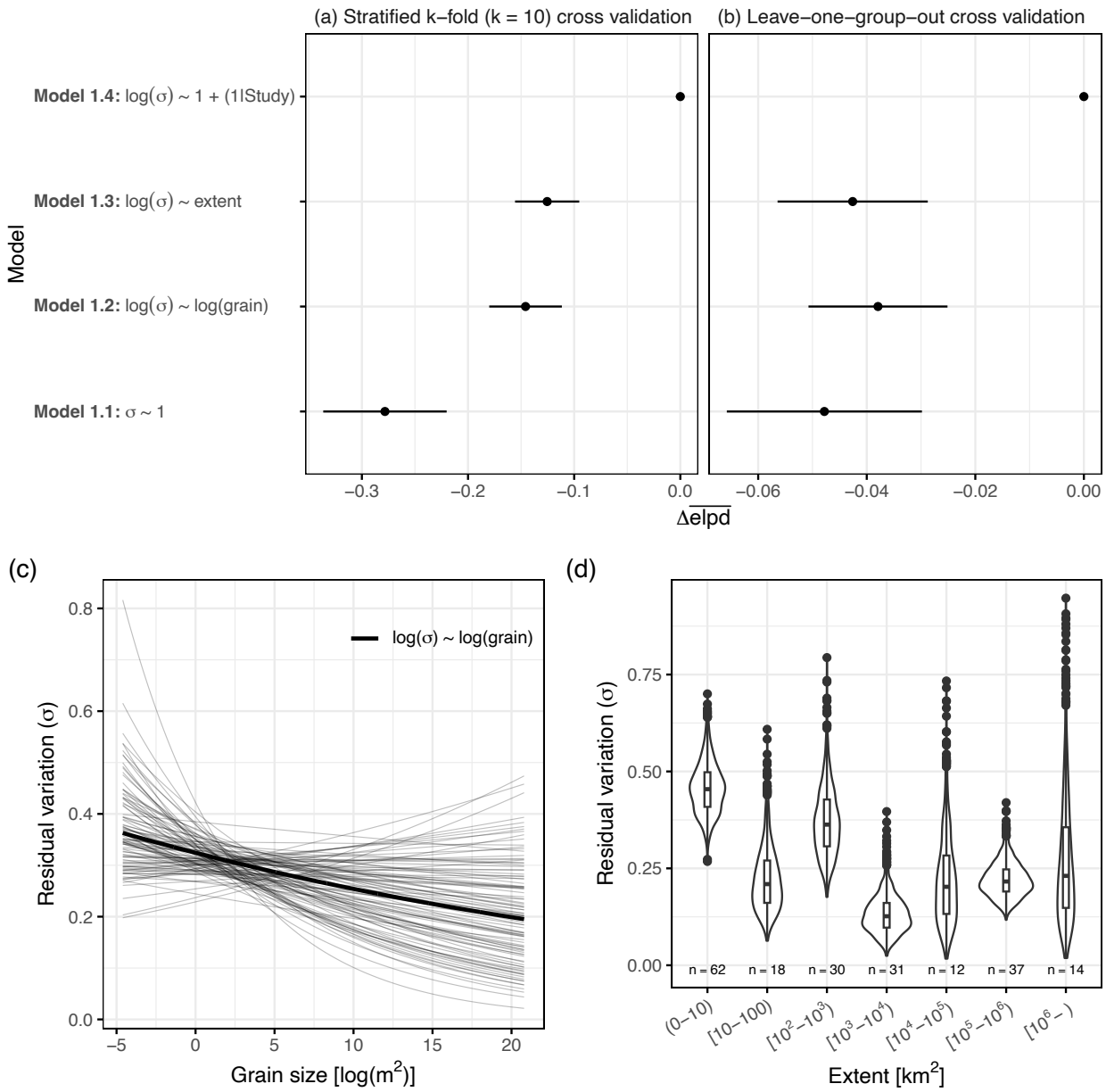
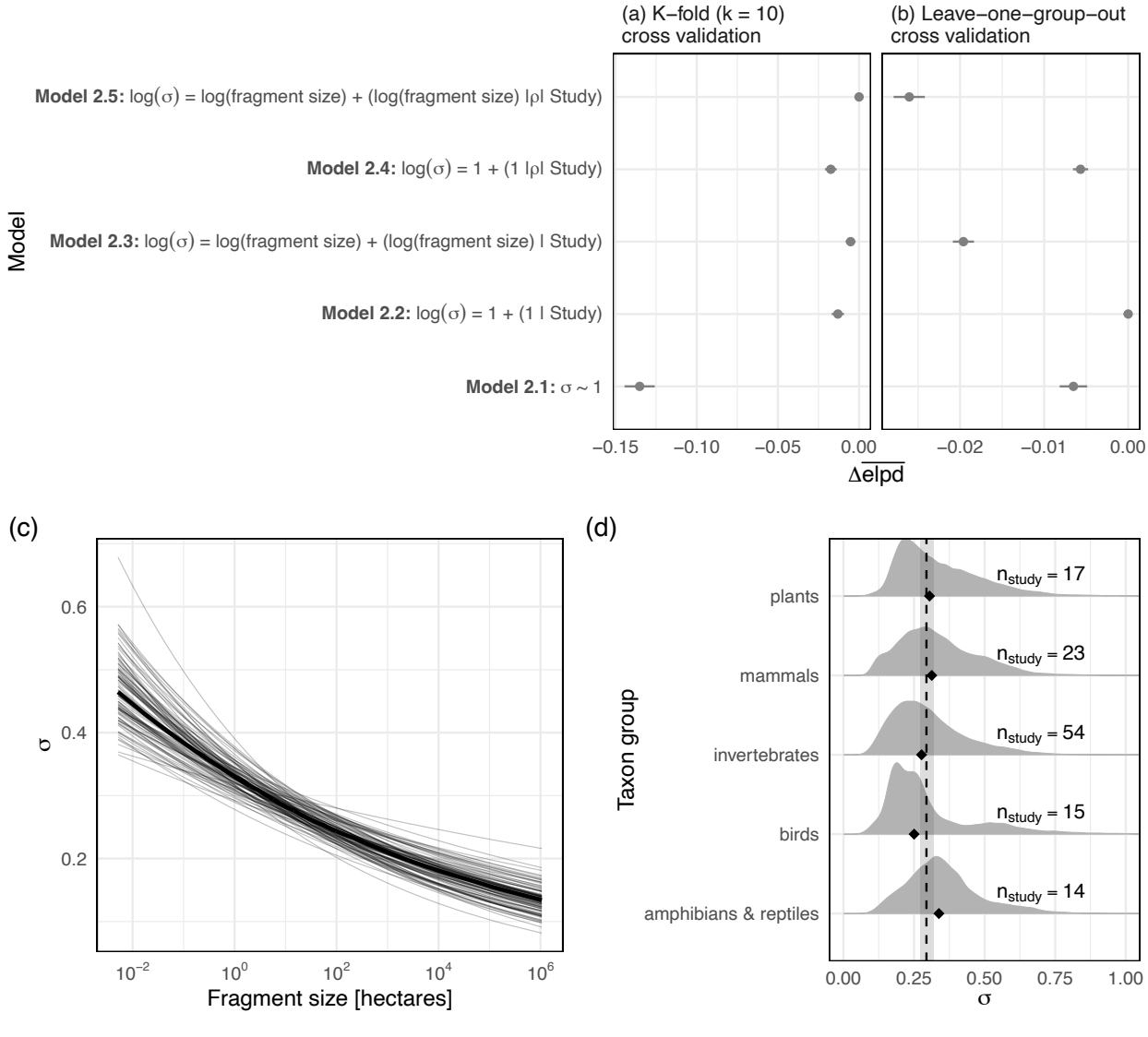

(a) Stratified k–fold (k = 10) cross validation
(b) Leave–one–group–out cross validation
Model 1.4: log(σ) ~ 1 + (1|Study)
Model 1.3: log(σ) ~ extent
Model 1.2: log(σ) ~ log(grain)
Model 1.1: σ ~ 1
Model
−0.3
−0.2
−0.1
0.0
−0.06
−0.04
−0.02
0.00
Δelpd
(c)
log(σ) ~ log(grain)
Residual variation (σ)
0.0
0.2
0.4
0.6
0.8
−5
0
5
10
15
20
Grain size [log(m2)]
(d)
Residual variation (σ)
0.00
0.25
0.50
0.75
n = 62
n = 18
n = 30
n = 31
n = 12
n = 37
n = 14
(0–10)
[10–100)
[10^2–10^3)
[10^3–10^4)
[10^4–10^5)
[10^5–10^6)
[10^6 –)
Extent [km2]

Figure 2


(a) K−fold (k = 10) cross validation
(b) Leave−one−group−out cross validation
Model
Model 2.5: log(σ) = log(fragment size) + (log(fragment size) |ρ| Study)
Model 2.4: log(σ) = 1 + (1 |ρ| Study)
Model 2.3: log(σ) = log(fragment size) + (log(fragment size) | Study)
Model 2.2: log(σ) = 1 + (1 | Study)
Model 2.1: σ ~ 1
−0.15
−0.10
−0.05
0.00
−0.02
−0.01
0.00
Δelpd
(c)
σ
0.6
0.4
0.2
10^−2
10^0
10^2
10^4
10^6
Fragment size [hectares]
(d)
Taxon group
plants
mammals
invertebrates
birds
amphibians & reptiles
n_study = 17
n_study = 23
n_study = 54
n_study = 15
n_study = 14
0.00
0.25
0.50
0.75
1.00
σ

Appendix S1: A (case-study) comparison of two location-scale models for meta-analysis

My first extension to include heteroscedasticity for the meta-analysis case-study modelled unexplained variation (i.e., the scale parameter, σ_{ij}) as a function of (log) grain size, which was added to the known sampling error (s_{ij}) of the effect size. Specifically,

$$z_{ij} \sim N(\mu_{ij}, s_{ij} + \sigma_{ij}),$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + \beta_1 X_i,$$

$$\beta_{0i} \sim N(0, \tau),$$

$$log(\sigma_{ij}) = \beta_0^{\sigma} + \beta_1^{\sigma} X_{ij}, \qquad \text{(Model 1.2)}$$

where X_{ij} is grain size on a log-scale for the jth case in study i; see Appendix S2 for the weakly regularizing priors that were used. This model is atypical of meta-analyses, which normally only include the known (or imputed if unknown) sampling error (s_{ij}) as the sole contributor to the scale component of the likelihood. Accordingly, the recent location-scale meta-analytic models instead include covariates for varying (random) intercepts in meta-regressions (e.g., Williams et al. 2021, Rodriguez et al. 2022, Viechtbauer & Lopez-Lopez 2022). To compare the two approaches I fit a model that includes grain size as a predictor the case-level varying intercept. Specifically, as each case had a grain, and some studies had multiple cases (i.e., more than one grain size), I modelled case-level variation (ω_j) as a linear function of (log) grain size (X_{ij}) with a log-link to ensure positive values for ω:

$$z_{ij} \sim N(\mu_{ij}, s_{ij}),$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + \beta_{0ij} + \beta_1 X_i,$$

$$\beta_{0i} \sim N(0, \tau),$$

$$\beta_{0ij} \sim N(0, \omega_{ij}),$$

$$log(\omega_{ij}) = \beta_0^{\omega} + \beta_1^{\omega} X_{ij}. \qquad \text{(Model 1.2a)}$$

I fit this model with similar weakly regularizing priors to that of model 1.2:

$$\beta_0 \sim N(0.3, 1),$$

$$\beta_0^{\omega}, \beta_1^{\omega}, \beta_1 \sim N(0, 1),$$

$$\tau \sim N(0, 1).$$

Model 1.2a had good convergence (all Rhats < 1.01), and was able to reproduce the observed effect sizes well (Appendix S1: Fig. S1).

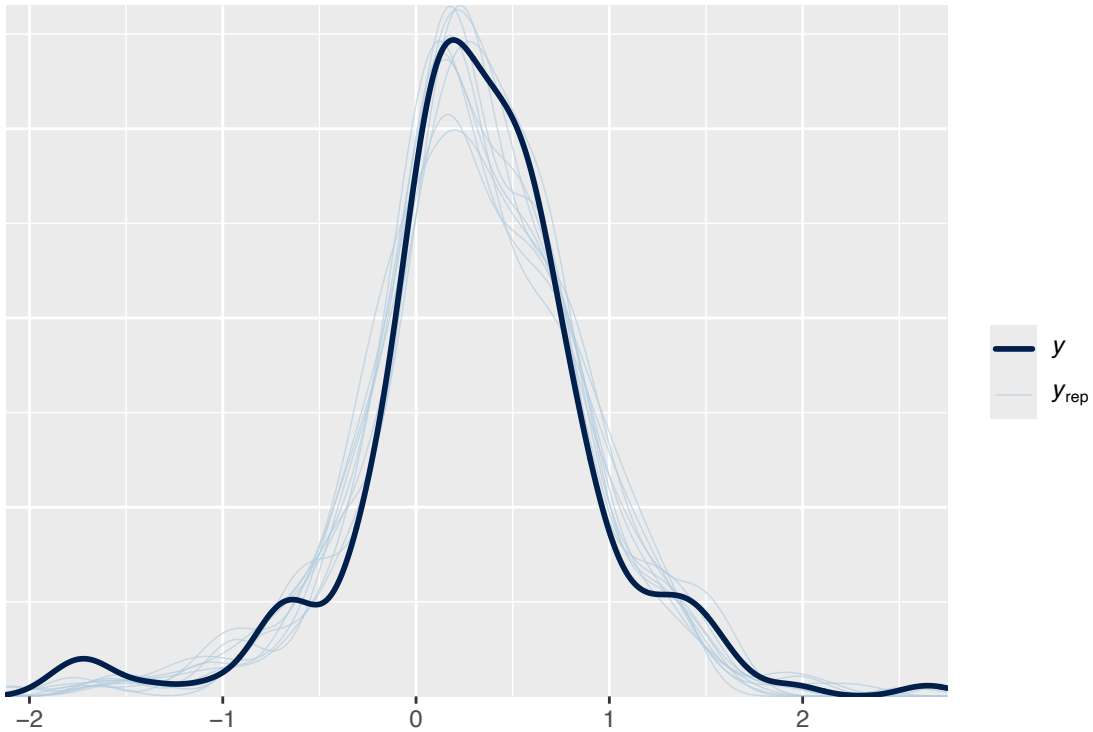


Figure S1: Posterior predictions from model 1.2a were able to reproduce the observed effect sizes relatively accurately.

The two models (1.2 and 1.2a) predicted similar relationships between unexplained variation and grain size (Appendix S1: Fig. S2), and had indistinguishable estimates for shared parameters (Appendix S1: Fig. S3). Model 1.2 and 1.2a had similar performance when making predictions within existing studies (Appendix S1: Fig. S4a), however model 1.2 was favored when making predictions to new studies (Appendix S1: Fig. S4b).

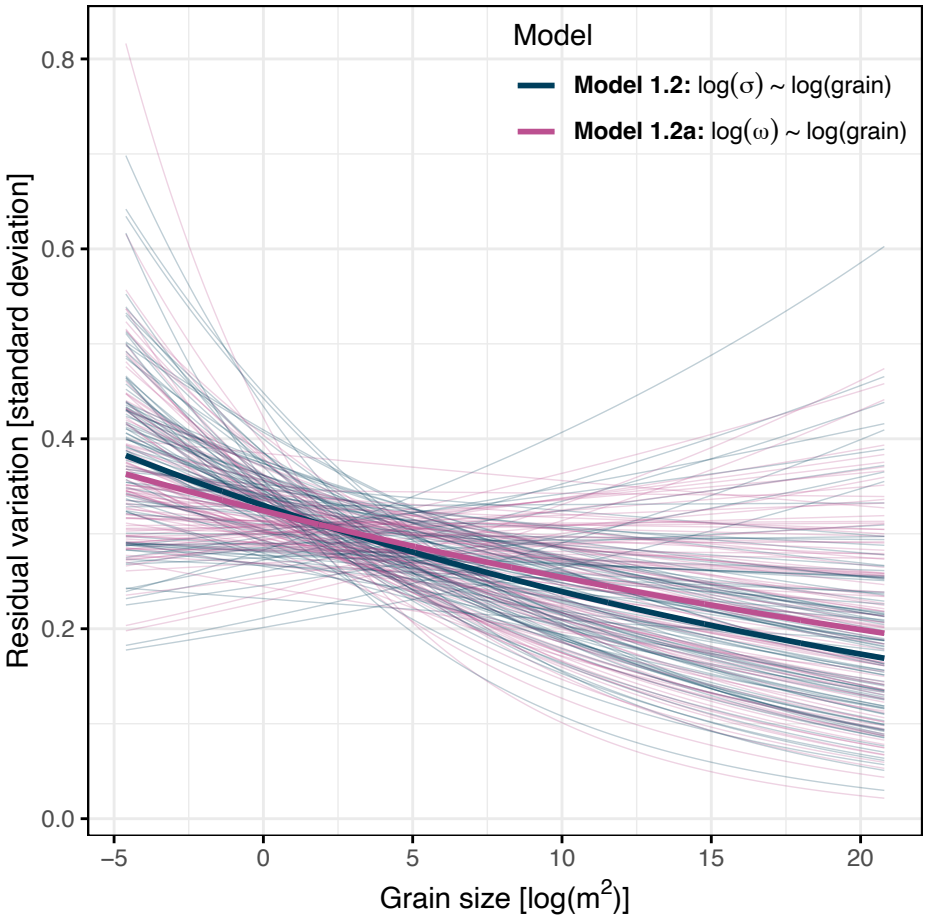


Figure S2: Models 1.2 and 1.2a predicted very similar relationships between unexplained variation and grain size.

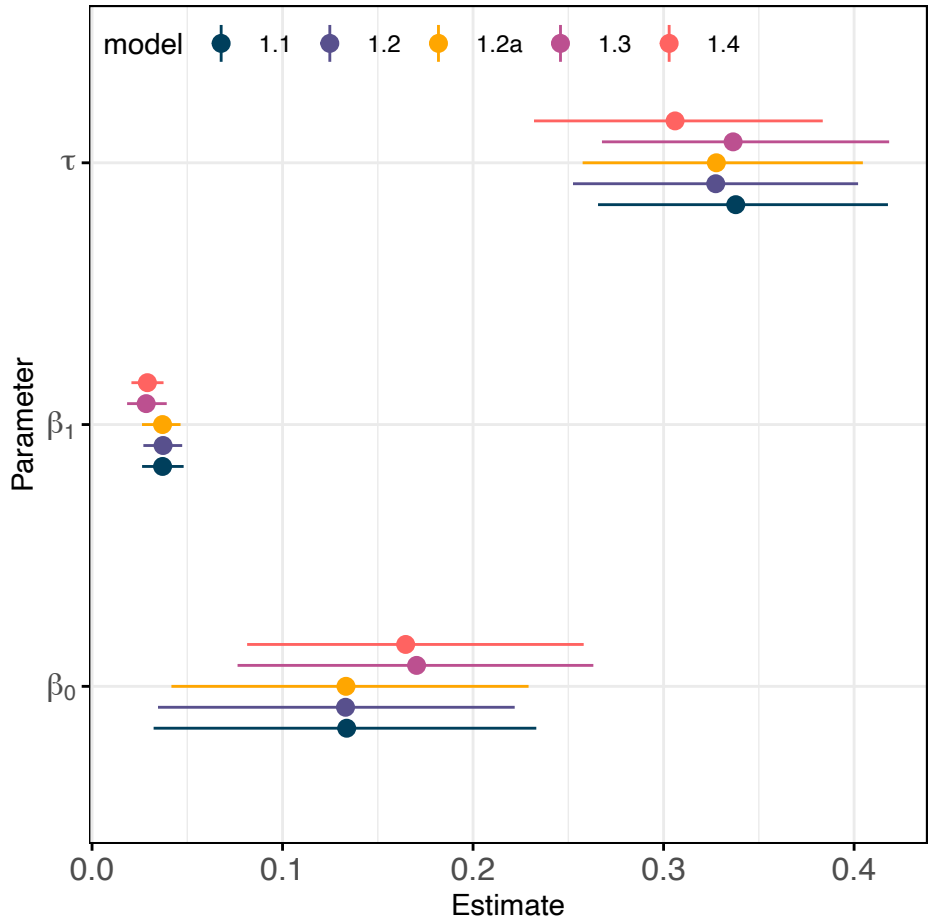


Figure S3: All models produced qualitative similar estimates for shared parameters (β_0, β_1, τ).

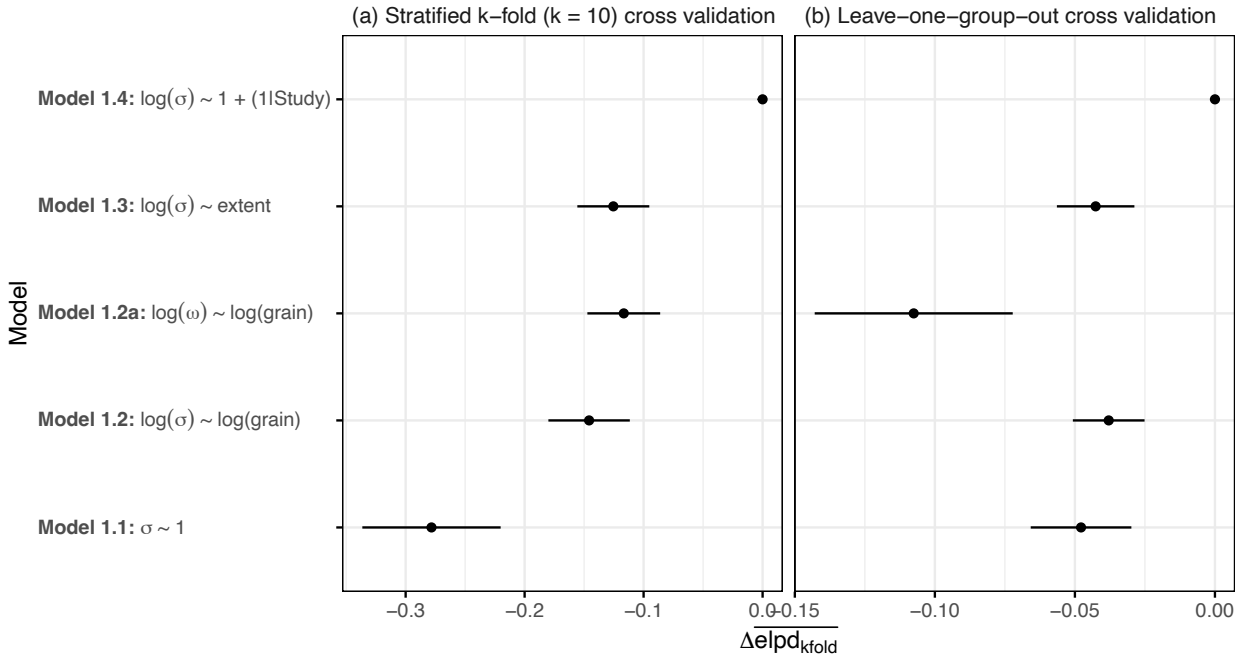


Figure S4: Heterogeneous heterogeneity outperforms constant heterogeneity for predicting plant native-exotic species richness relationships. Model selection for predictions (a) within existing studies using stratified k-fold cross-validation, and (b) for new studies using leave-one-group-out cross validation.

References

Rodriguez, J. E., D. R. Williams, and P. Bürkner. 2023. Heterogeneous heterogeneity by default: Testing categorical moderators in mixed-effects meta-analysis. British Journal of Mathematical and Statistical Psychology 76:402–433.

Viechtbauer, W., and J. A. López-López. 2022. Location-scale models for meta-analysis. Research Synthesis Methods 13:697–715.

Williams, D. R., J. E. Rodriguez, and P.-C. Bürkner. 2021. Putting variation into variance: modeling between-study heterogeneity in meta-analysis.

Appendix S2

Case-study one: Simulation-based calibration and supplemental figures

I used Simulation-Based Calibration (Talts et al. 2020, Modrák et al. 2023) to check whether models could recover known parameter values with reasonable coverage (i.e., the probability that a constructed [e.g., credible] interval contains the true value). Briefly, this involves: (1) fitting a model to the empirical data; (2) using the same model with priors informed by the fit to the empirical data to simulate many new (fake) datasets with the same size, shape and structure as the empirical data (i.e., the same number of observations, and the same number of groups for each level in the hierarchical structure); (3) refitting the model to each simulated data set; and, (4) calculating SBC diagnostics and plotting results of the model fits to simulated data to check for reasonable coverage (Talts et al. 2020, Modrák et al. 2023).

Model 1.1

I begin case study one by reproducing the main result from Peng et al. (2019) using a multilevel linear meta-regression model for grain size that assumes the effect sizes, z_{ij}, are normally distributed with known within-case variance (s_{ij}^2) and constant between-study variance (τ^2), which can be expressed as:

$$z_{ij} \sim N(\mu_{ij}, s_{ij}),$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + \beta_{0ij} + \beta_1 X_{ij},$$

$$\beta_{0i} \sim N(0, \tau),$$

$$\beta_{0ij} \sim N(0, \omega),$$

where cases (j) are nested within studies (i), and have among-case (within study) variance, ω^2; study-level heterogeneity, β_{0i}, has variance τ^2, and varies around the overall linear relationship

with intercept, β_0, slope β_1, and predictor X_i (here the natural logarithm of grain size in study *i*). The model was fit to empirical data with weakly regularizing priors:

$$\beta_0 \sim N(0.3, 1),$$

$$\beta_1 \sim N(0, 1),$$

$$[\tau, \omega] \sim N(0, 1), \qquad \text{(Model S1.1).}$$

This model had good convergence (all Rhat < 1.01), and showed a reasonable fit to the empirical data (Appendix S2: Fig. S1).

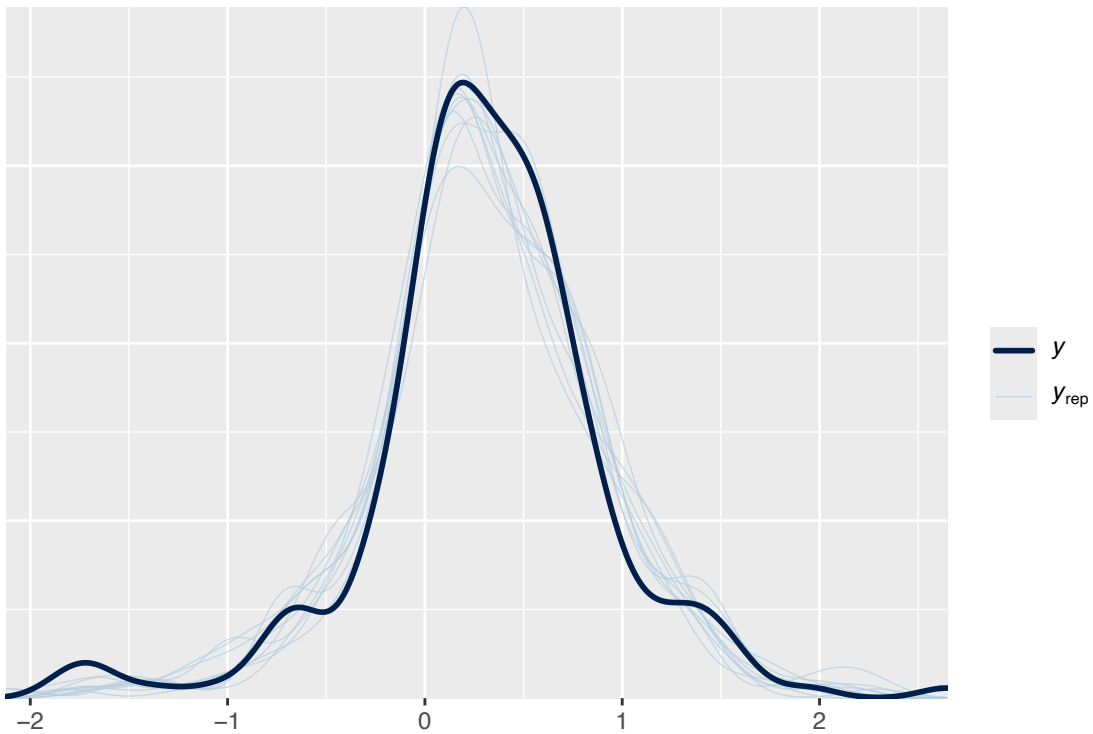


Figure S1: Posterior predictive check for model 1.1 fit to the empirical data (*z*-scores).

The parameter estimates from the fit of model S1.1 to empirical data (β_0: 0.13 [95% credible interval: 0.01 – 0.24]; β_1: 0.04 [95% credible interval: 0.02 – 0.05]; τ: 0.34 [95% credible interval: 0.25 – 0.43]; ω: 0.3 [95% credible interval: 0.24 – 0.35]) were used to inform the following priors:

$$\beta_0 \sim N(0.13, 0.06),$$

$$\beta_1 \sim N(0.04, 0.02),$$

$$\tau \sim N(0, 0.25),$$

$$\omega \sim N(0, 0.25),$$

which were combined with model S1.1 to simulate (fake) data sets. To ensure as much realism as possible in the simulated data, each simulated data set retained the characteristics of the empirical data: 204 observations distributed across 101 studies (with the same balance, i.e., effect sizes per study, and the known standard errors [s_{ij}] from the empirical data). I refit model S1.1 to each simulated data set, and identified models fit to fake data with poor diagnostics (e.g., divergent transitions and Rhats > 1.05), likely due to the simulated data have extreme values for the simulated data; I report the number of simulated data sets used to calculate simulation diagnostics in figure captions.

Simulation-based calibration for model 1.1 showed that the rank statistics were approximately uniformly distributed (Appendix S2: Fig. S2), and that the coverage of parameters was reasonable (Appendix S2: Fig. S3); known parameters were approximately recovered (Appendix S2: Fig. S4).

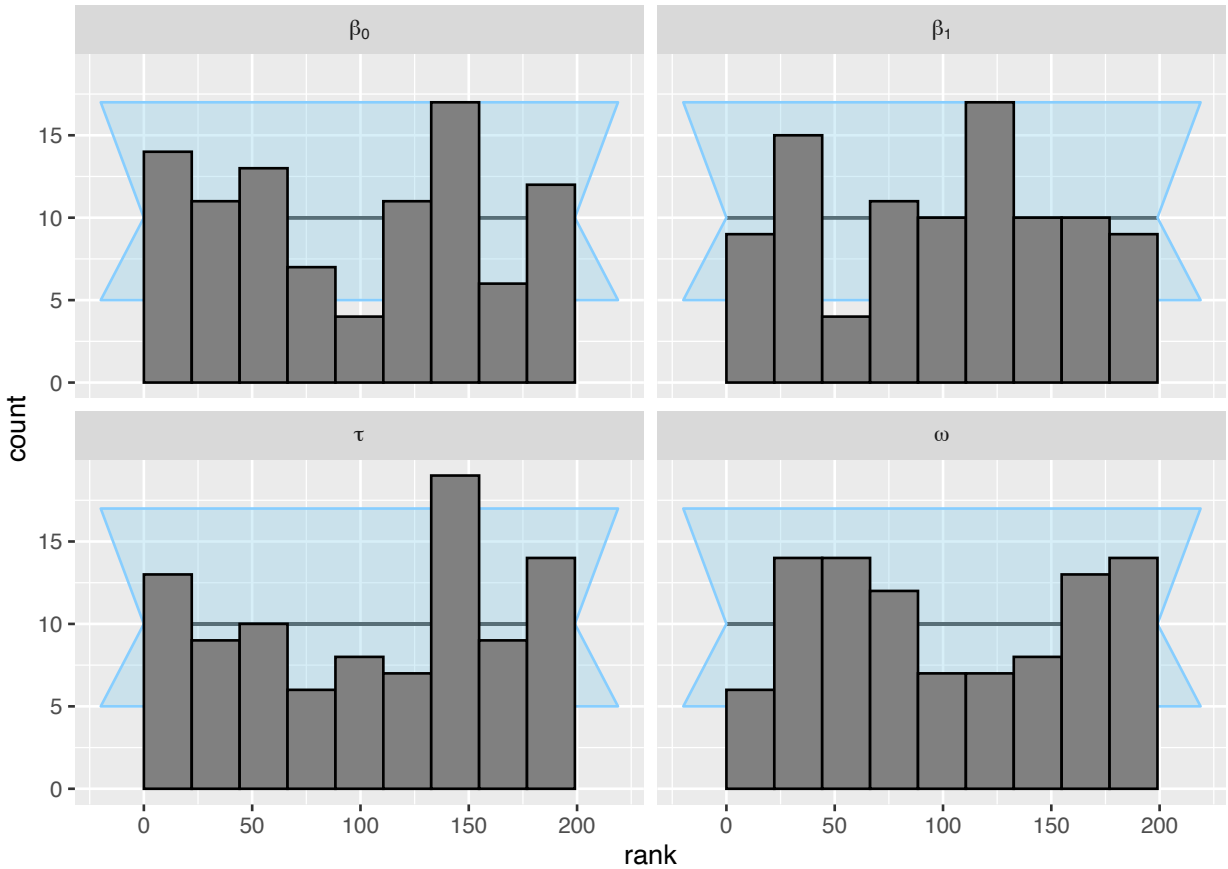


Figure S2: The posterior ranks of the prior draws were approximately normally distributed for the parameters of interest in model S1.1. Results are shown for n = 95 simulated data sets.

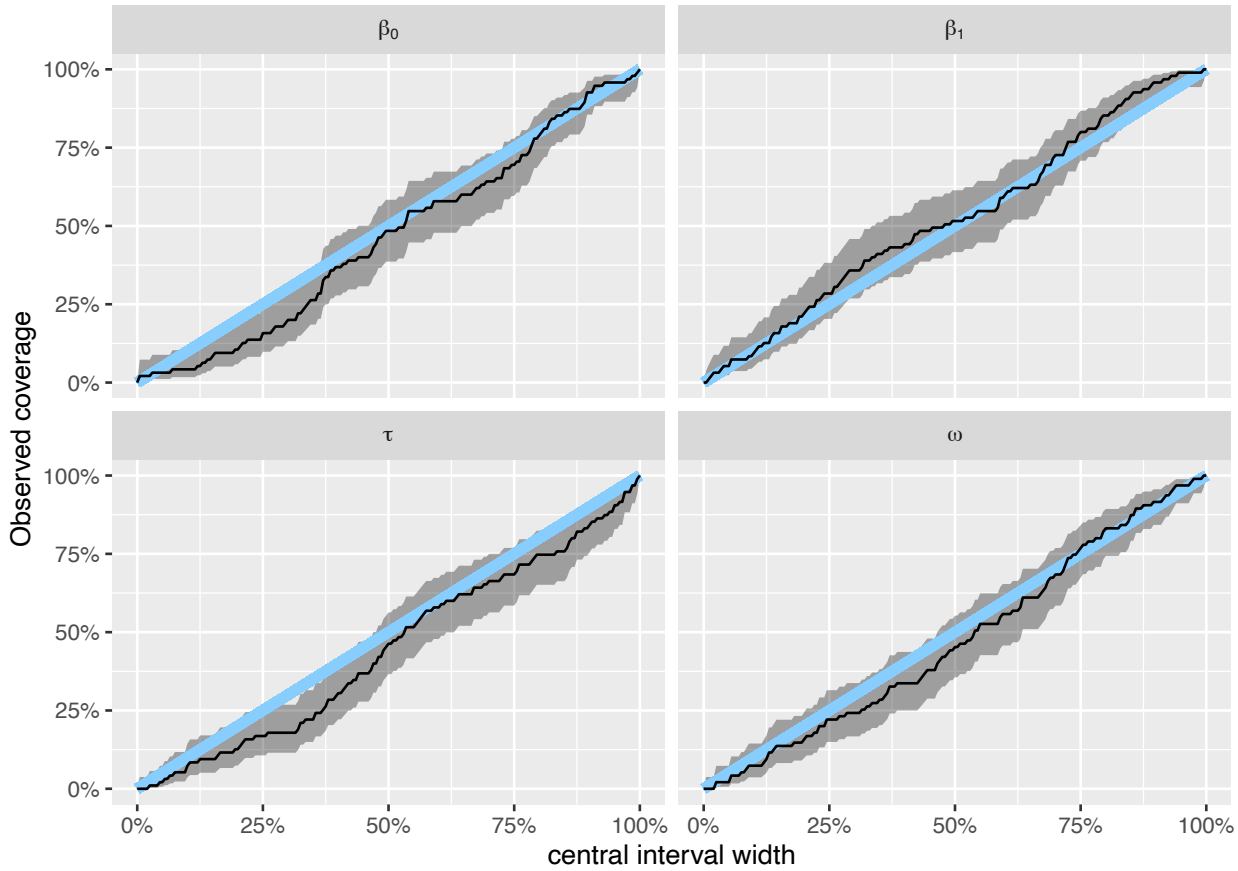


Figure S3: Model S1.1 had reasonable coverage for the parameters of interest. Results are shown for n = 95 simulated data sets.

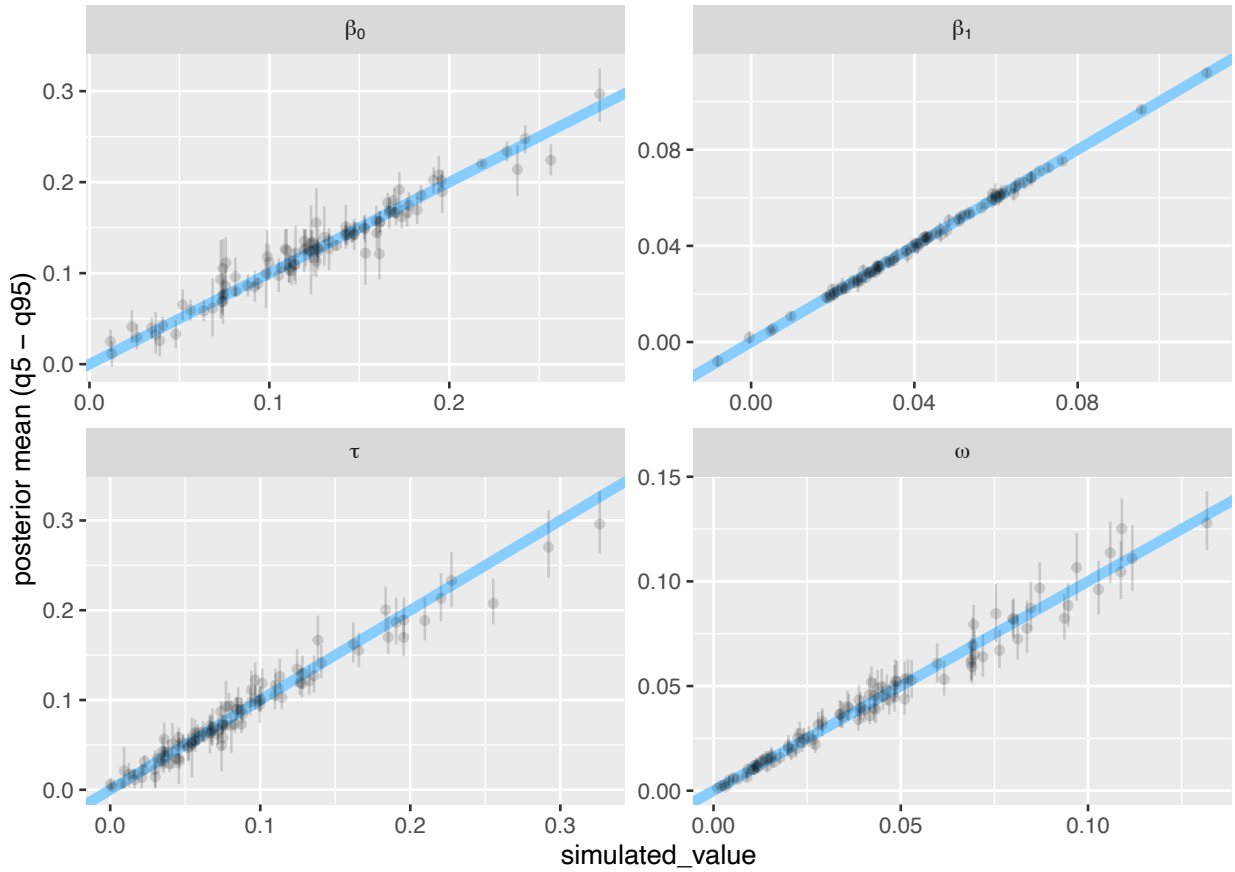


Figure S4: Model S1.1 was largely able to recover known parameter values. Results are shown for n = 95 simulated data sets.

Model 1.2

The second model in case study one extended model S1.1 to include an additional parameter (σ_{ij}) for residual variation as a function of grain size:

$$z_{ij} \sim N(\mu_{ij}, s_{ij} + \sigma_{ij}),$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + \beta_1 X_i,$$

$$\beta_{0i} \sim N(0, \tau),$$

$$log(\sigma_{ij}) = \beta_0^\sigma + \beta_1^\sigma X_{ij},$$

where X_{ij} is grain size on a log-scale for the *j*th case in study *i*. I fit this model to the empirical data with weakly regularizing priors:

$$\beta_0 \sim N(0.3, 1),$$

$$\beta_0^\sigma, \beta_1^\sigma, \beta_1 \sim N(0, 1),$$

$$\tau \sim N(0, 1), \qquad \text{(Model S1.2).}$$

Model 1.2 had good convergence (all Rhats < 1.01), and showed a good fit to the empirical data (Appendix S2: Fig. S5).

The parameter estimates from the fit of model S1.2 to empirical data (β_0: 0.13 [95% credible interval: 0.01 – 0.24]; β_1: 0.04 [95% credible interval: 0.02 – 0.05]; τ: 0.33 [95% credible interval: 0.24 – 0.42]; β_0^σ: -1.12 [95% credible interval: -1.4 – -0.86]; β_1^σ: -0.03 [95% credible interval: -0.1 – 0.03]) were used to inform the following priors:

$$\beta_0 \sim N(0.13, 0.06),$$

$$\beta_1 \sim N(0.04, 0.02),$$

$$\tau \sim N(0, 0.2),$$

$$\beta_0^\sigma \sim N(-1.1, 0.2),$$

$$\beta_1^\sigma \sim N(-0.03, 0.03),$$

which were combined with model S1.2 to simulate (fake) data sets.

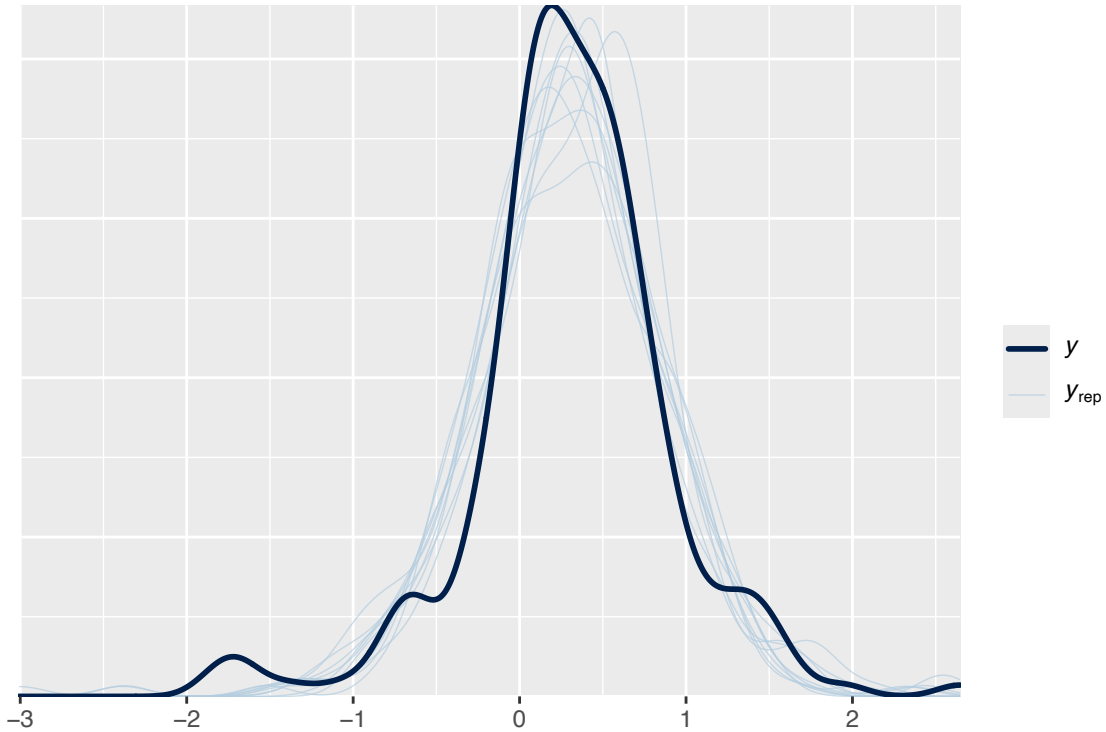


Figure S5: Posterior predictions from model 1.2 were largely consistent with the observed data.

Simulation-based calibration for model 1.2 showed that the rank statistics were approximately uniformly distributed (Appendix S2: Fig. S6), and that the coverage of parameters was reasonable (Appendix S2: Fig. S7); known parameters were approximately recovered, though with a fair amount of uncertainty for the intercept (β_0), the varying intercept for studies (τ), and the slope of relationship between unexplained variation and grain size (β_1^{σ}; Appendix S2: Fig. S8).

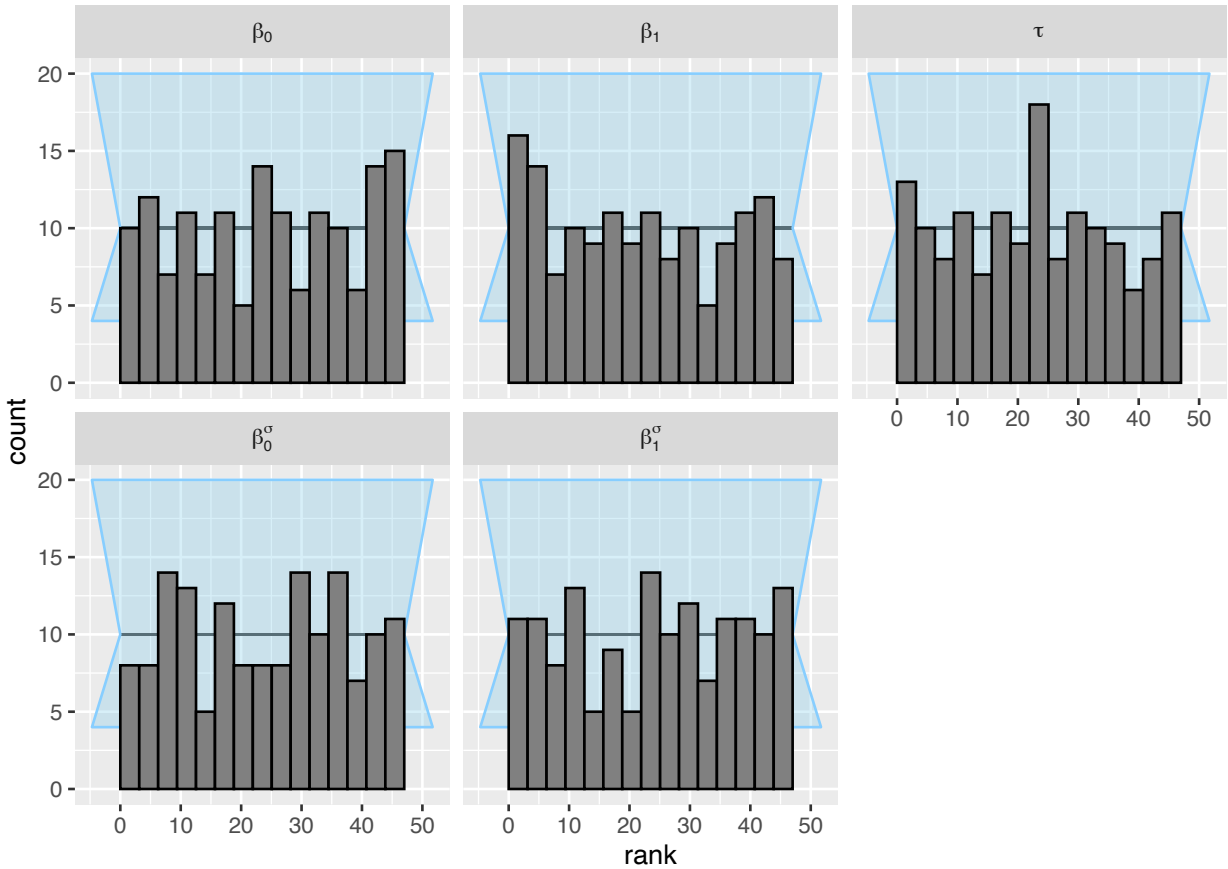


Figure S6: The posterior ranks of the prior draws were approximately normally distributed for the parameters of interest in model S1.2. Results are shown for n = 150 simulated data sets.

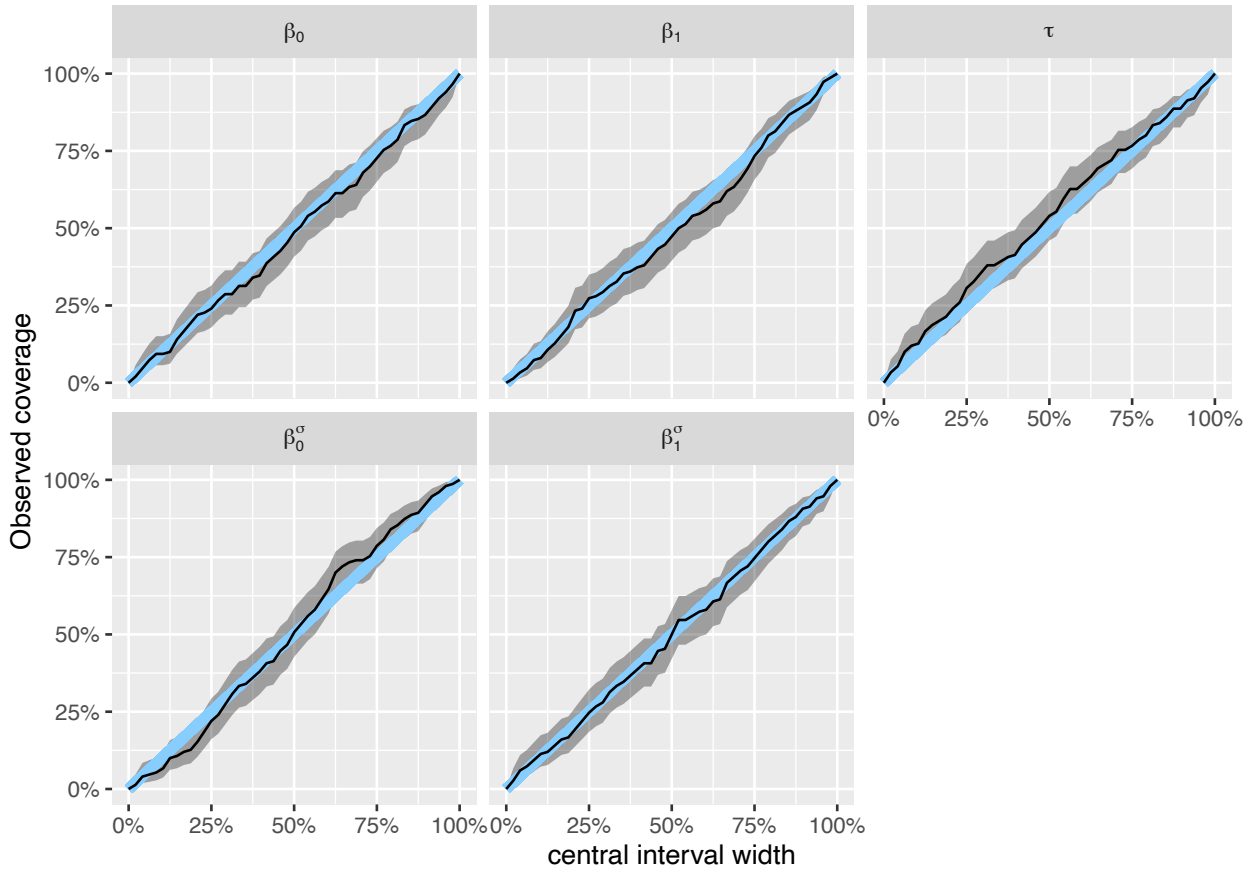


Figure S7: Model S1.2 had reasonable coverage for the parameters of interest. Results are shown for n = 150 simulated data sets.

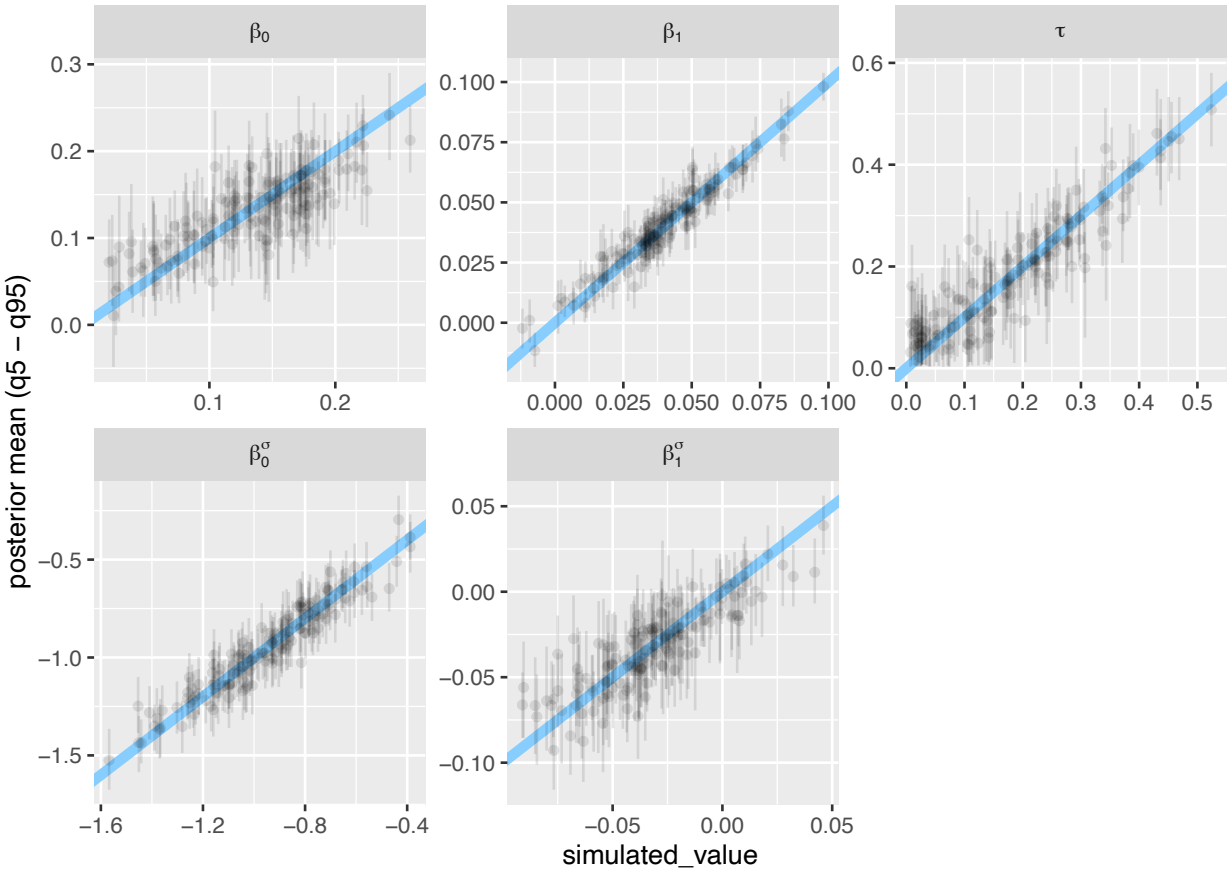


Figure S8: Model S1.2 was able to recover known parameter values, though with some uncertainty (especially for β_0, τ, and β_1^{σ}). Results are shown for n = 150 simulated data sets.

Model 1.3

The third model replaced grain size as a predictor of unexplained variation with spatial extent, which was estimated with indicator variables for each category of spatial extent [i.e., no intercept, see main text]). I fit this model to the empirical data with the same weakly regularizing priors as S1.2 (with the exception of β_0^{σ}, which was dropped from the model).

Model 1.3 had good convergence (all Rhats < 1.01), and showed a reasonable fit to the empirical data (Appendix S2: Fig. S9).

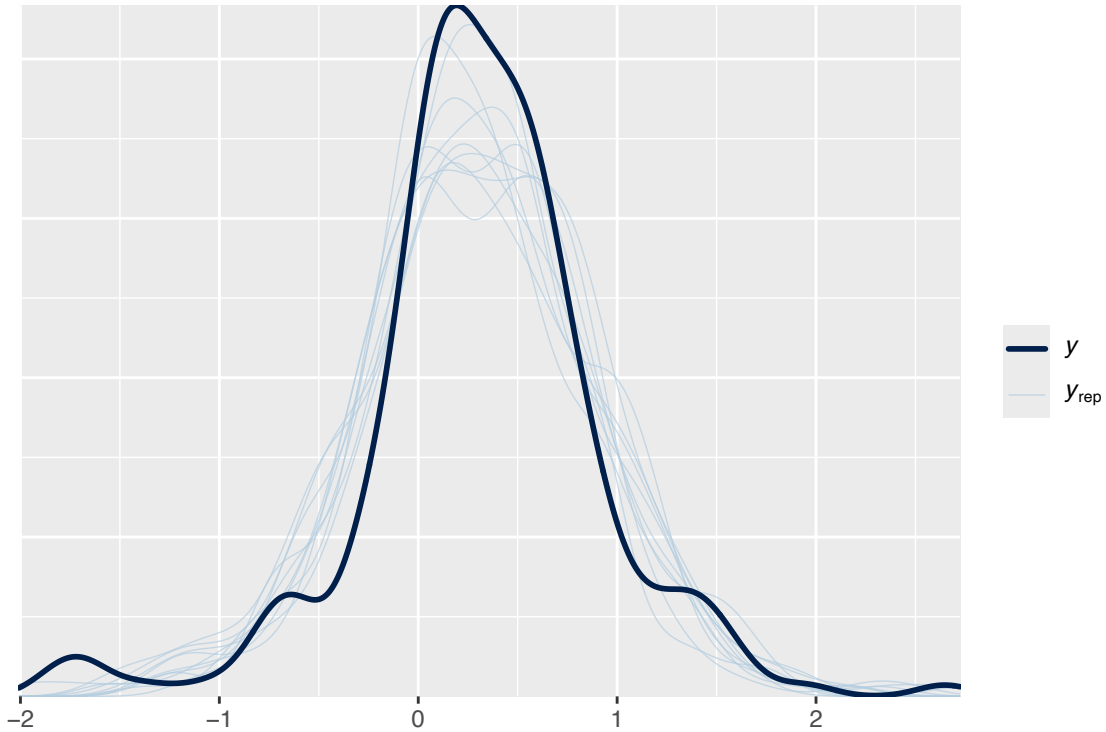


Figure S9: Posterior predictions for model 1.3 showed reasonable fidelity to the empirical data.

The parameter estimates from the fit of model S1.3 to empirical data (β_0: 0.17 [95% credible interval: 0.06 – 0.28]; β_1: 0.03 [95% credible interval: 0.02 – 0.04]; τ: 0.34 [95% credible interval: 0.25 – 0.43]; β^{σ}_{0-10}: -0.79 [95% credible interval: -1.07 - -0.52]; β^{σ}_{10-100}: -1.58 [95% credible interval: -2.3 - -0.83]; $\beta^{\sigma}_{10\text{^}2-10\text{^}3}$: -1.03 [95% credible interval: -1.5 - -0.57]; $\beta^{\sigma}_{10\text{^}3-10\text{^}4}$: -2.09 [95% credible interval: -2.95- -1.39]; $\beta^{\sigma}_{10\text{^}4-10\text{^}5}$: -1.66 [95% credible interval: -2.93 - -0.69]; $\beta^{\sigma}_{10\text{^}5-10\text{^}6}$: -1.53 [95% credible interval: -1.91 - -1.13]; $\beta^{\sigma}_{10\text{^}6-}$: -1.51 [95% credible interval: -2.87- -0.38]) were used to inform the following priors:

$$\beta_0 \sim N(0.13, 0.05),$$

$$\beta_1 \sim N(0.04, 0.02),$$

$$\tau \sim N(0, 0.2),$$

$$\beta^{\sigma}_1 \sim N(-1,1),$$

which were combined with model S1.3 to simulate (fake) data sets.

Simulation-based calibration for model 1.3 showed that the rank statistics were approximately uniformly distributed (Appendix S2: Fig. S10), and that the coverage of parameters was reasonable (Appendix S2: Fig. S11); known parameters were approximately recovered, though with a fair amount of uncertainty for the intercept (β_0), and the varying intercept for studies (τ, Appendix S2: Fig. S12).

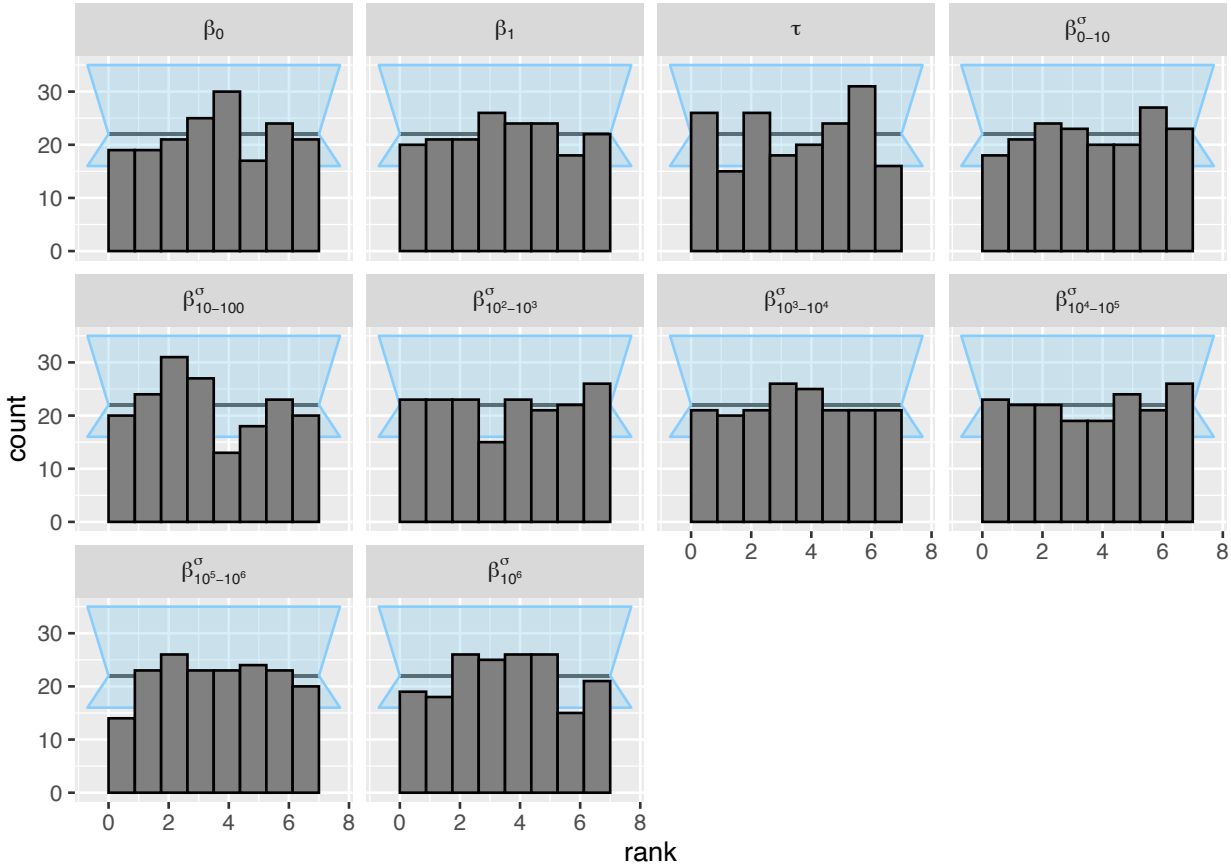


Figure S10: The posterior ranks of the prior draws were approximately normally distributed for the parameters of interest in model S1.3. Results are shown for n = 176 simulated data sets.

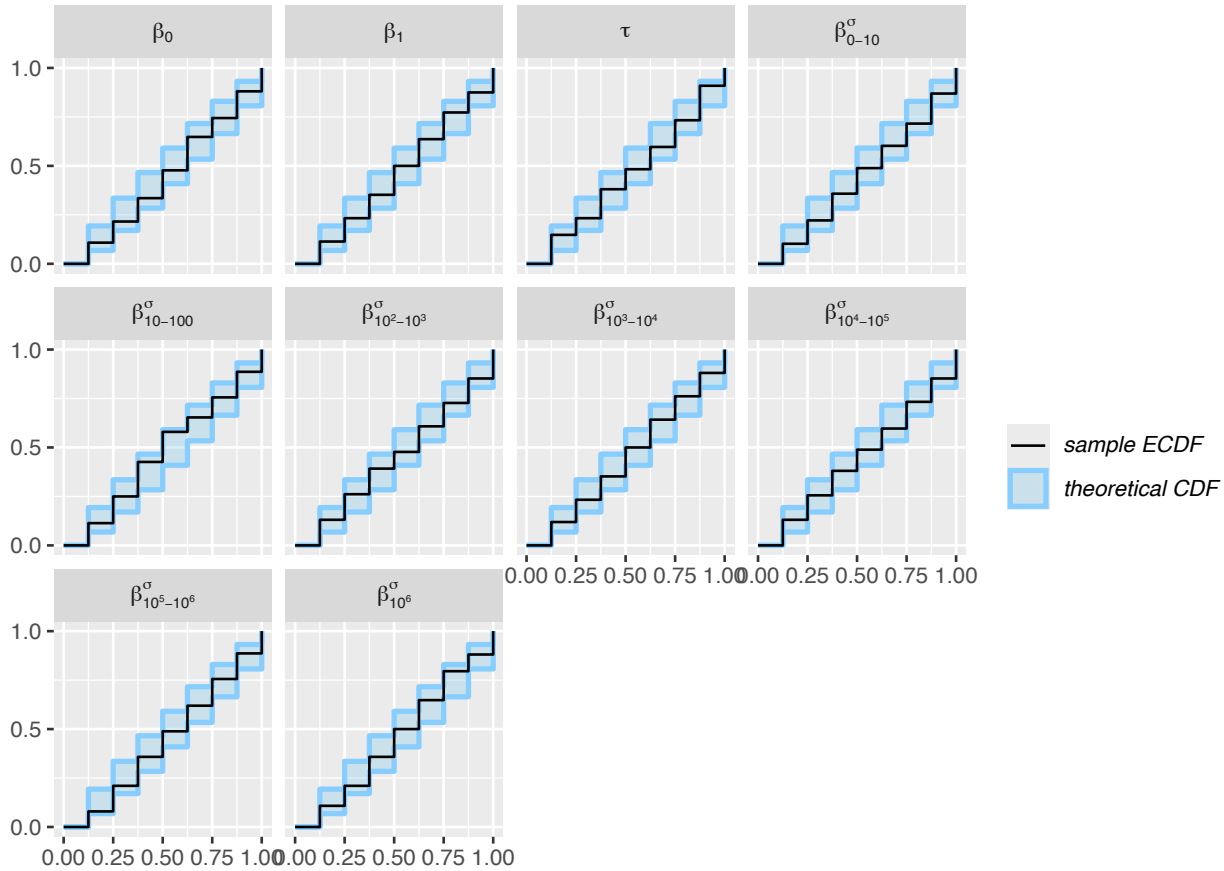


Figure S11: Model S1.3 had reasonable coverage for the parameters of interest. Results are shown for n = 176 simulated data sets.

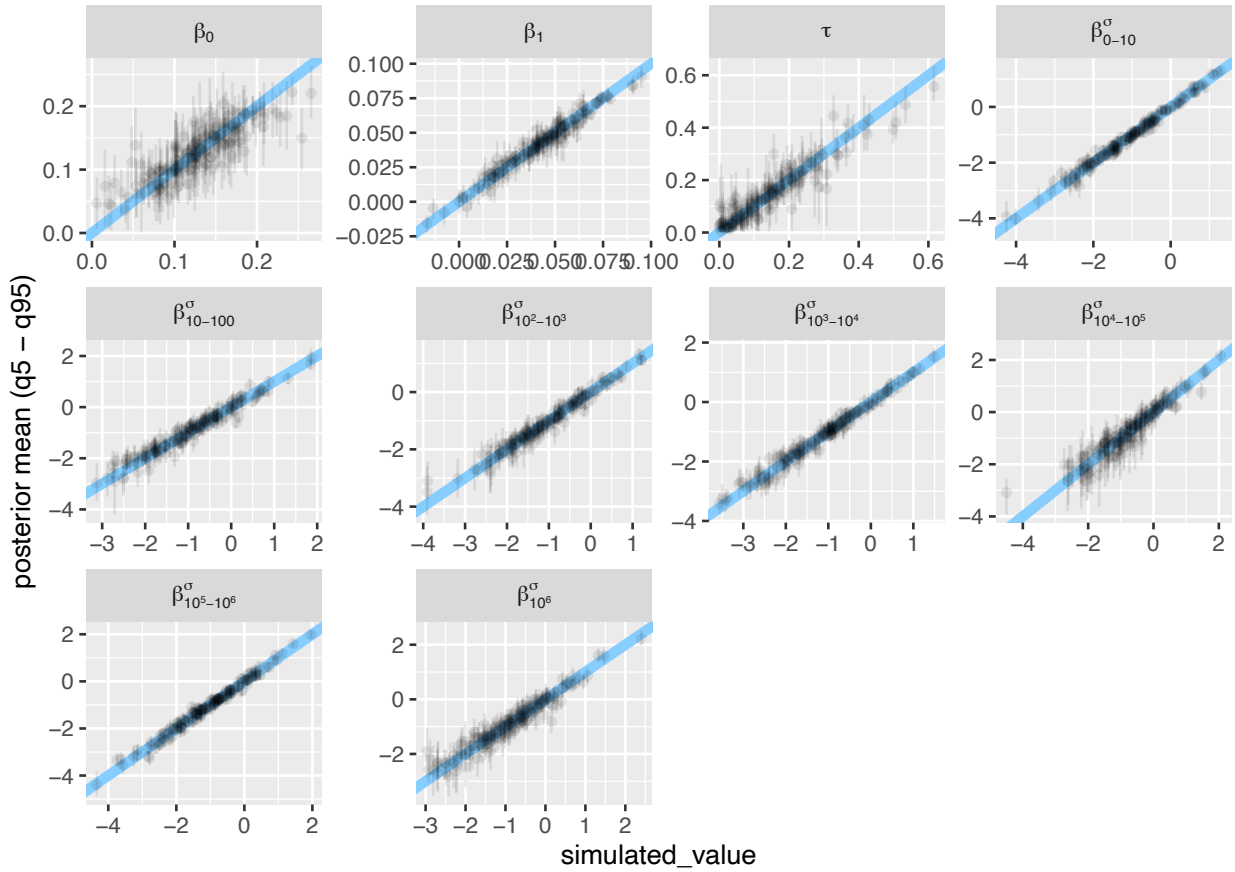


Figure S12: Model S1.3 was able to recover known parameter values, though with some uncertainty (most notably for β_0, and τ). Results are shown for n = 176 simulated data sets.

Model 1.4

The final model examined in the main text included study-level residual variation:

$$z_{ij} \sim N(\mu_{ij}, s_{ij} + \sigma_{ij}),$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + \beta_1 X_i,$$

$$\beta_{0i} \sim N(0, \tau),$$

$$log(\sigma_{ij}) = \beta_0^{\sigma} + \beta_{0i}^{\sigma},$$

$$\beta_{0i}^{\sigma} \sim N(0, \zeta),$$

where β_0^σ is the average residual variation (on a log-scale), and β_{0i}^σ is a normally distributed study-level varying intercept with zero mean and ζ standard deviation. The model was fit to the empirical data with weakly regularizing parameters:

$$\beta_0 \sim N(0.3, 1),$$

$$\beta_1, \beta_0^\sigma \sim N(0, 1),$$

$$[\tau, \zeta] \sim N(0, 1). \qquad \text{(Model S1.4)}$$

Model 1.4 had good convergence (all Rhats < 1.01), and showed a reasonable fit to the empirical data, albeit with some very low predicted z-scores (Appendix S2: Fig. S13).

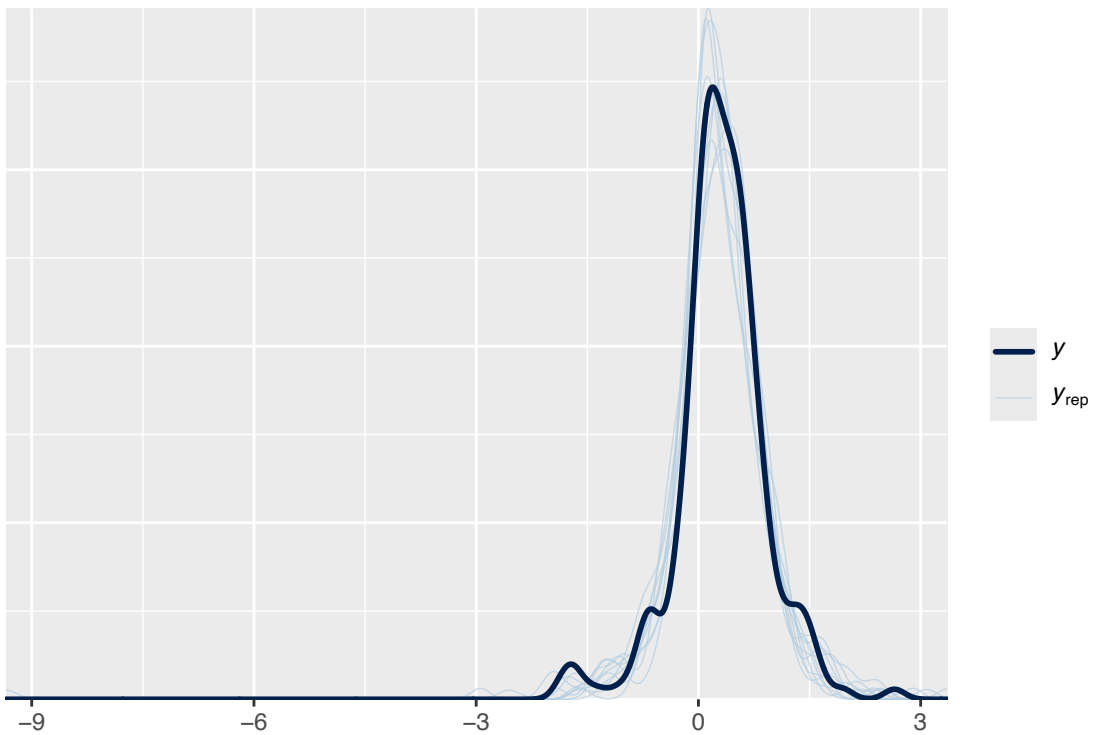


Figure S13: Posterior predictions from model 1.4 showed a good fit to the empirical data, though it did predict some extremely low z-scores (not seen in the observed data).

The parameter estimates from the fit of model S1.4 to empirical data (β_0: 0.17 [95% credible interval: 0.06 – 0.27]; β_1: 0.03 [95% credible interval: 0.02 – 0.05]; τ: 0.31 [95%

credible interval: 0.22 – 0.40]; β_0^σ: -1.65 [95% credible interval: -2.15 – -1.26]; ζ: 0.89 [95% credible interval: 0.56 – 1.31]) were used to inform the following priors:

$$\beta_0 \sim N(0.17, 0.05),$$

$$\beta_1 \sim N(0.04, 0.02),$$

$$\tau \sim N(0, 0.2),$$

$$\beta_0^\sigma \sim N(-1.6, 0.25),$$

$$\zeta \sim N(0, 0.7),$$

which were combined with model S1.4 to simulate (fake) data sets.

Simulation-based calibration for model 1.4 showed that the rank statistics were approximately uniformly distributed (Appendix S2: Fig. S14), and that the coverage of parameters was reasonable (Appendix S2: Fig. S15). Finally, most of the known parameters were approximately recovered, though with a fair amount of uncertainty for the intercept (β_0), and particularly the varying intercept for study-level unexplained variation (ζ; Appendix S2: Fig. S16). The relatively poor (and highly uncertain) estimate of study-level unexplained variation (ζ) is likely due to the limited replication within studies (the median number of effect sizes per study was one; 74 of 101 studies had a single effect size).

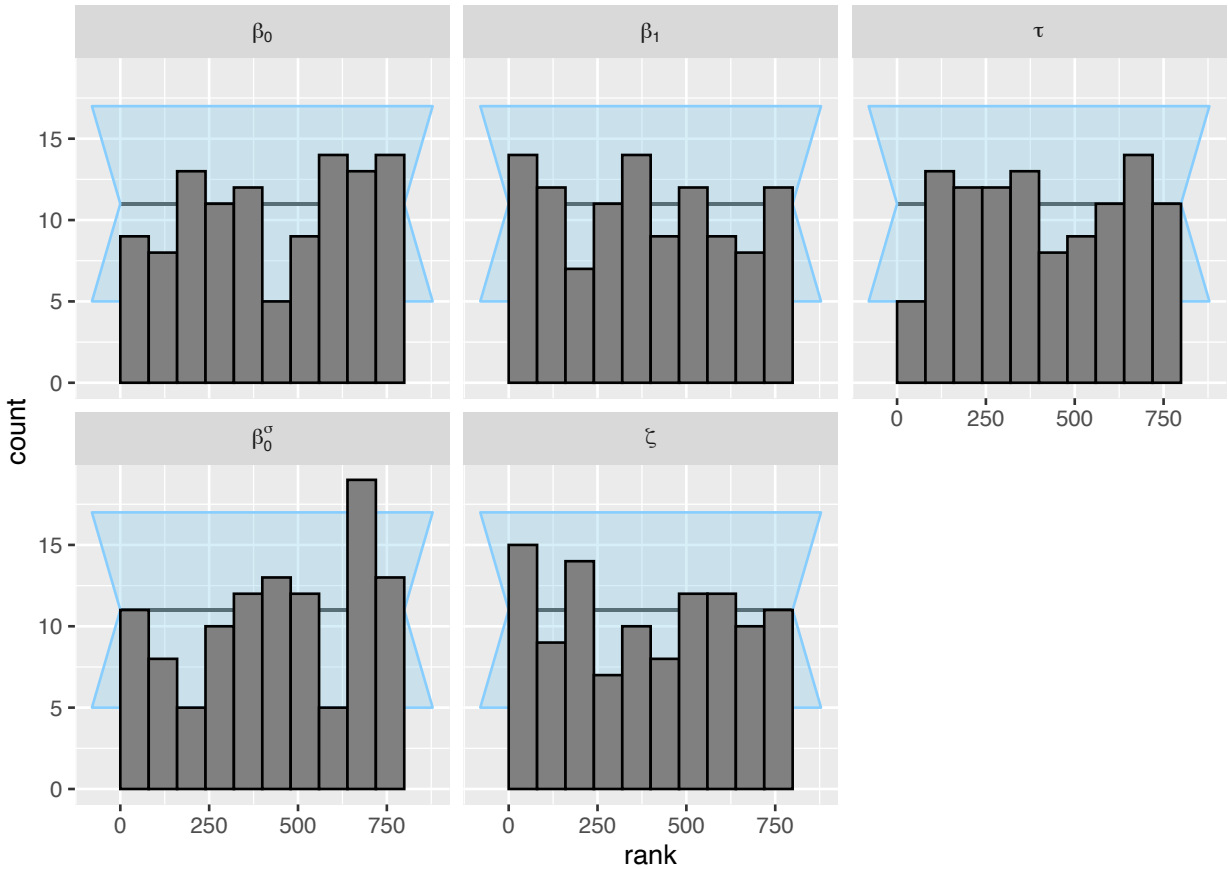


Figure S14: The posterior ranks of the prior draws were approximately normally distributed for the parameters of interest in model S1.4. Results are shown for n = 108 simulated data sets.

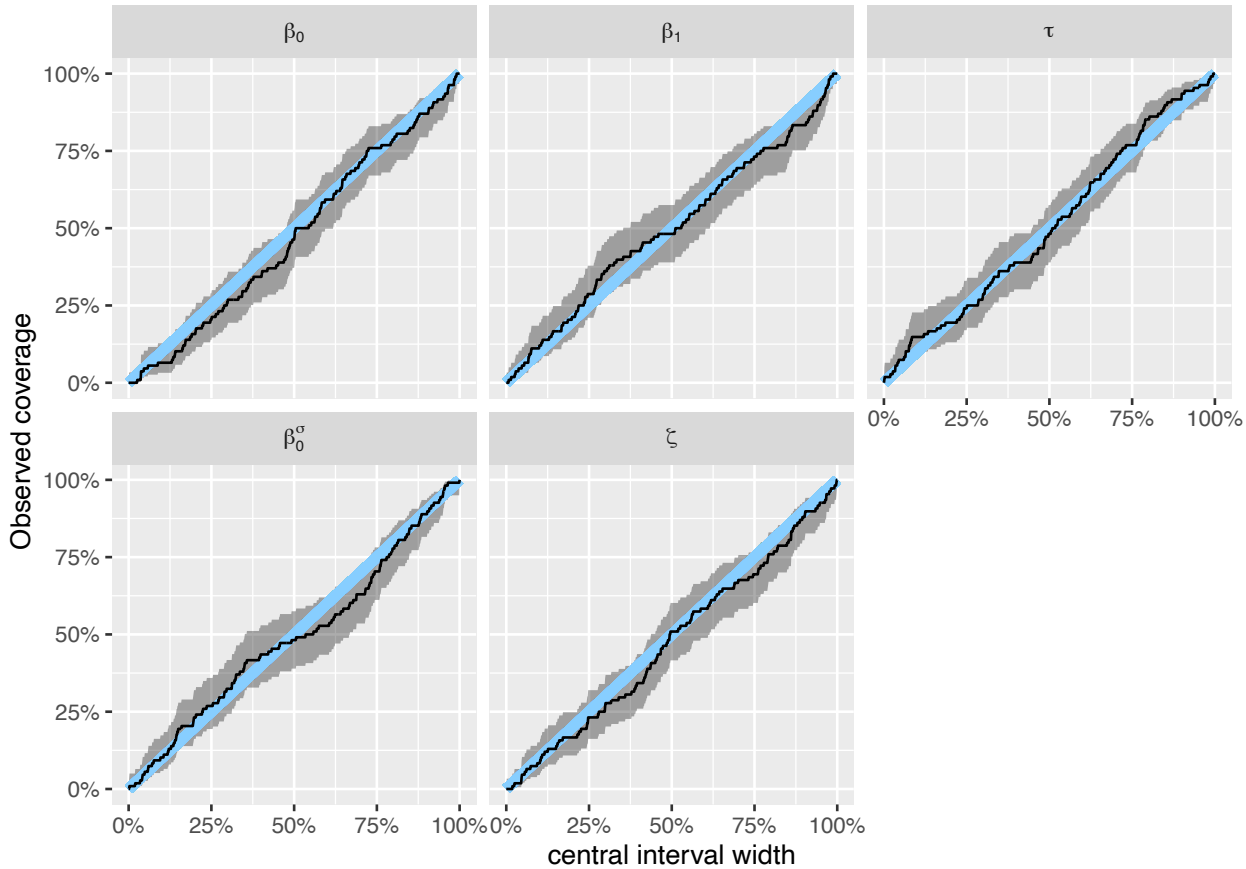


Figure S15: Model S1.4 had reasonable coverage for the parameters of interest, though the intercept (β_0) showed some discrepancies. Results are shown for n = 108 simulated data sets.

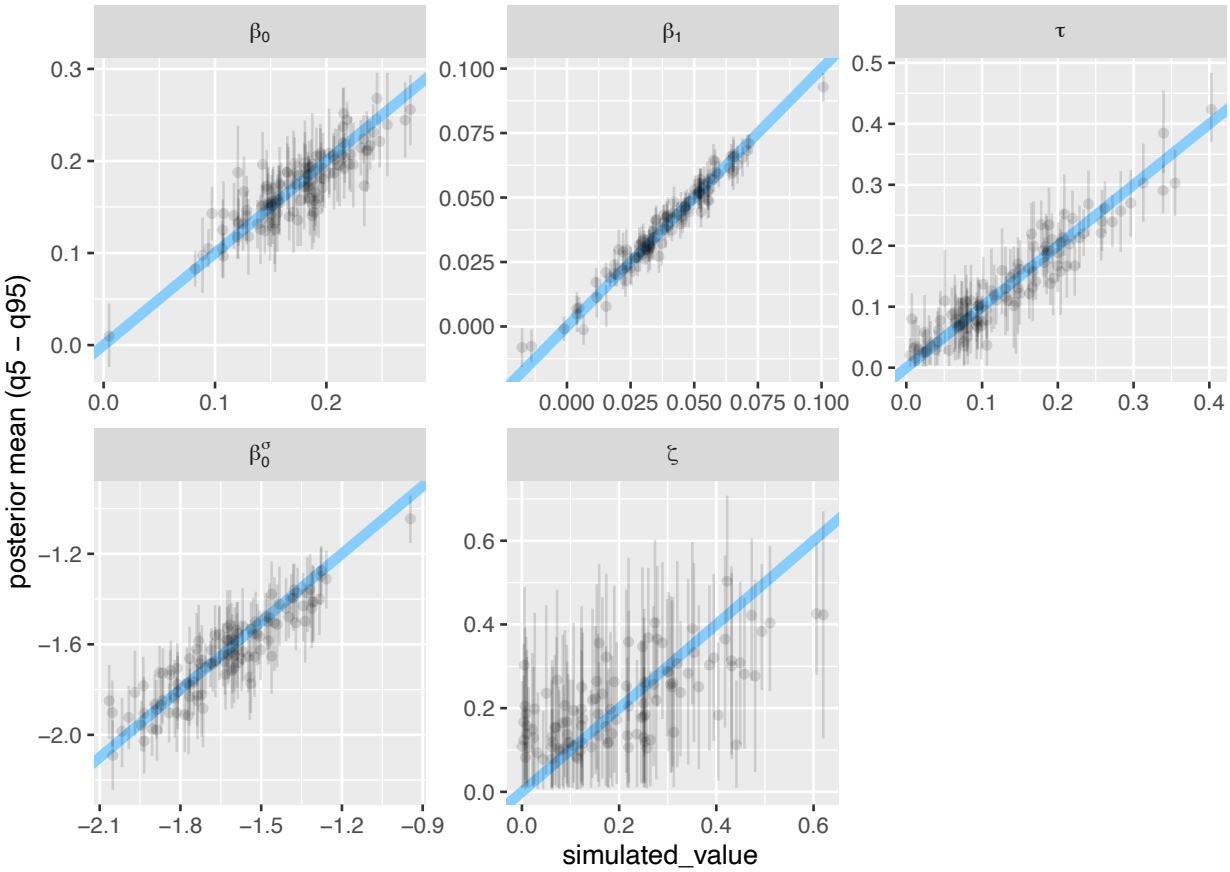


Figure S16: Model S1.3 was able to recover known parameter values, though with some uncertainty (most notably ζ). Results are shown for n = 108 simulated data sets.

References

Modrák, M., A. H. Moon, S. Kim, P. Bürkner, N. Huurre, K. Faltejsková, A. Gelman, and A. Vehtari. 2023. Simulation-based calibration checking for Bayesian computation: The choice of test quantities shapes sensitivity. Bayesian Analysis 1:1–28.

Peng, S., N. L. Kinlock, J. Gurevitch, and S. Peng. 2019. Correlation of native and exotic species richness: a global meta-analysis finds no invasion paradox across scales. Ecology 100:e02552.

Talts, S., M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. 2020, October 21. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv.

Appendix S3

Case-study two: Simulation-based calibration and supplemental figures

I used Simulation-Based Calibration (Talts et al. 2020, Modrák et al. 2023) to check whether models could recover known parameter values with reasonable coverage (i.e., the probability that a constructed [e.g., credible] interval contains the true value). Briefly, for my purposes here this involves: (1) fitting a model to the empirical data; (2) using the same model with priors informed by the fit to the empirical data to simulate many new (fake) datasets with the same size, shape and structure as the empirical data (i.e., the same number of observations, and the same number of groups for each level in the hierarchical structure); (3) refitting the model to each simulated data set; and, (4) calculating and plotting SBC diagnostics of model fits to simulated data to check for reasonable coverage (Talts et al. 2020, Modrák et al. 2023).

Model 2.1

To start case-study two I refit the model fit by Chase et al. (2020) to effort-standardized for species richness (S_{ij}) in fragment j from study i:

$$S_{ij} \sim lognormal(\mu_{ij}, \sigma)$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i})X_{ij},$$

$$[\beta_{0i}, \beta_{1i}]' \sim MVN(0, \boldsymbol{SRS}),$$

$$\boldsymbol{S} = \begin{bmatrix} \sigma_{0i} & 0 \\ 0 & \sigma_{1i} \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}\sigma_{1i}} \\ \rho_{\sigma_{0i}\sigma_{1i}} & 1 \end{bmatrix},$$

where X_{ij} is the fragment size on a (natural) log-scale, which was centered by subtracting the overall mean from each observation before modelling; β_{0i} and β_{1i} are study-level departures from the overall intercept and slope, respectively, drawn from a multivariate normal (*MVN*)

distribution with standard deviations, σ_{0i} and σ_{1i}, that allowed for correlations (***R*** matrix) between the varying intercepts and slopes. The model was fit to empirical data using weakly regularizing priors:

$$\beta_0 \sim N(2.5, 1),$$

$$\beta_1 \sim N(0, 1),$$

$$[\sigma_{0i}, \sigma_{1i}, \sigma] \sim N(0, 1),$$

$$\mathbf{R} \sim \text{LKJ}(1). \qquad \text{(Model S2.1)}$$

This model had good convergence (all Rhat <= 1.01), and showed a reasonable fit to the empirical data (Appendix S3: Fig. S1).

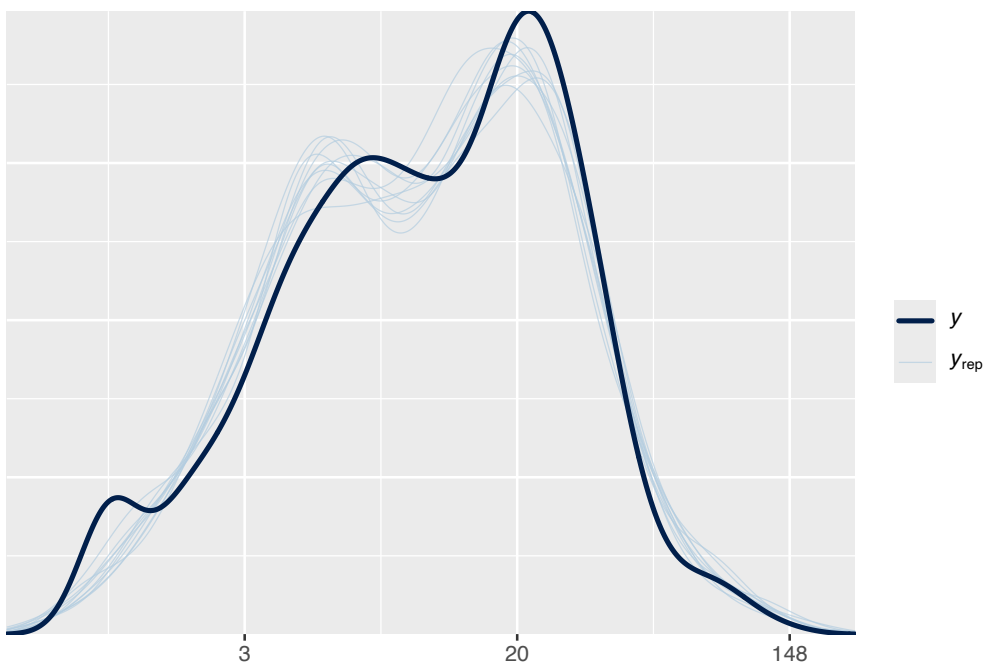


Figure S1: Posterior predictions from model 2.1 showed the model had a good ability to predict the observed data.

The parameter estimates from the fit of model S2.1 to empirical data (β_0: 2.46 [95% credible interval: 2.28 – 2.65]; β_1: 0.06 [95% credible interval: 0.04 – 0.07]; σ_{0i}: 1 [95% credible interval:

922 0.88 – 1.14]; σ_{1i}: 0.06 [95% credible interval: 0.04 – 0.08], and σ: 0.36 [95% credible interval:
923 0.34 – 0.37]) were used to inform the following priors:

924 $$\beta_0 \sim N(2.4, 0.1),$$

925 $$\beta_1 \sim N(0.06, 0.01),$$

926 $$\sigma_{0i} \sim N(1, 0.1),$$

927 $$\sigma_{1i} \sim N(0, 0.05),$$

928 $$\sigma \sim N(0.3, 0.05),$$

929 which were combined with model S2.1 to simulate (fake) data sets. To ensure as much realism as
930 possible in the simulated data, each simulated data set retained the characteristics of the
931 empirical data: 1509 observations distributed across 123 studies (with the same balance, i.e.,
932 fragment size distribution and fragments per study). I refit model S2.1 to each simulated data set,
933 and identified models fit to fake data with poor diagnostics (e.g., divergent transitions and Rhats
934 > 1.05), likely due to the simulated data have extreme values for the simulated data; I report the
935 number of simulated data sets used to calculate simulation diagnostics in figure captions.

936 Simulation-based calibration for model 2.1 showed that the rank statistics were
937 approximately uniformly distributed (Appendix S3: Fig. S2), and that the coverage of parameters
938 was reasonable (Appendix S3: Fig. S3); known parameters were approximately recovered
939 (Appendix S3: Fig. S4). Note, (here and in subsequent calibrations for model extensions) I do not
940 show parameters associated with the intercept (i.e., β_0, σ_{0i}) as these parameters are not readily
941 interpretable for this analysis. They correspond to the average richness (across all studies), and
942 study-level variation around the overall average. Richness (and other diversity measures) are not
943 readily comparable across this heterogeneous data compilation: richness in any given data set
944 will depend on the sample grain and size of the species pool (among other things).

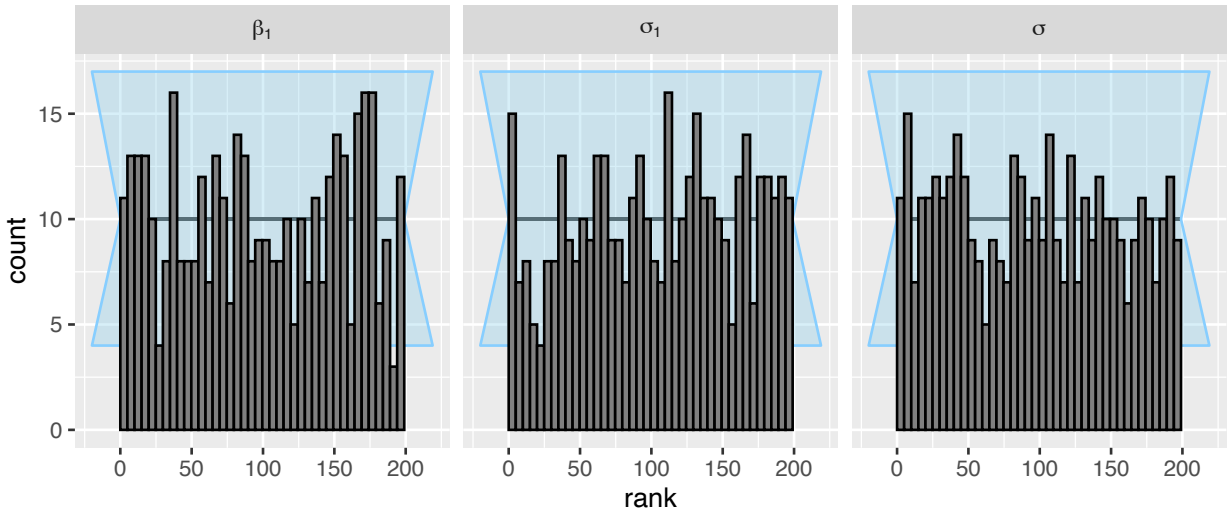


Figure S2: The posterior ranks of the prior draws were approximately normally distributed for the parameters of interest in model S2.1. Results are shown for n = 401 simulated data sets.

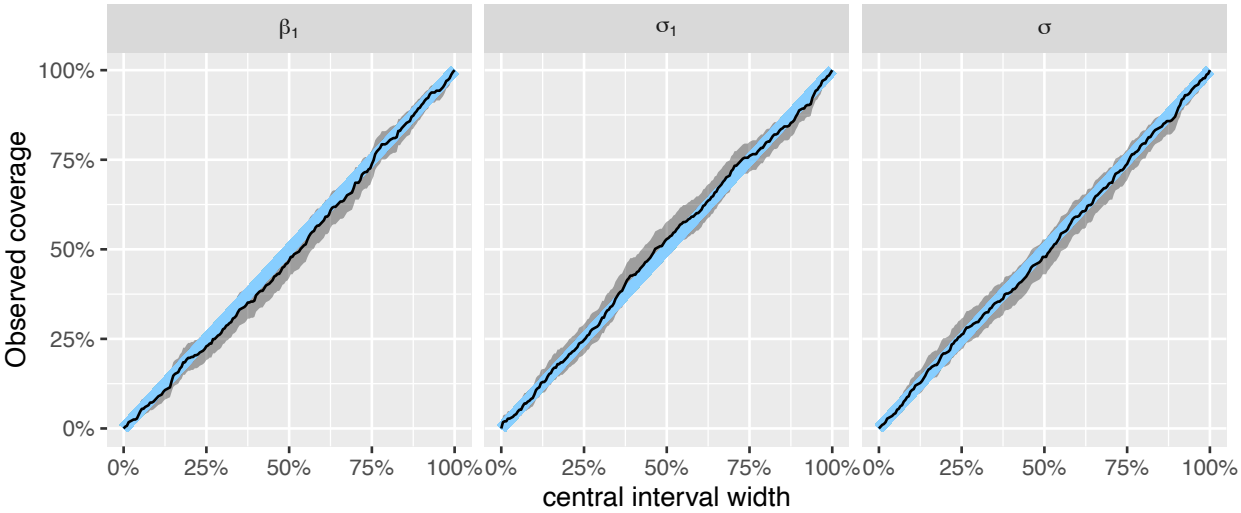


Figure S3: Model S2.1 had good coverage for the parameters of interest. Results are shown for n = 401 simulated data sets.

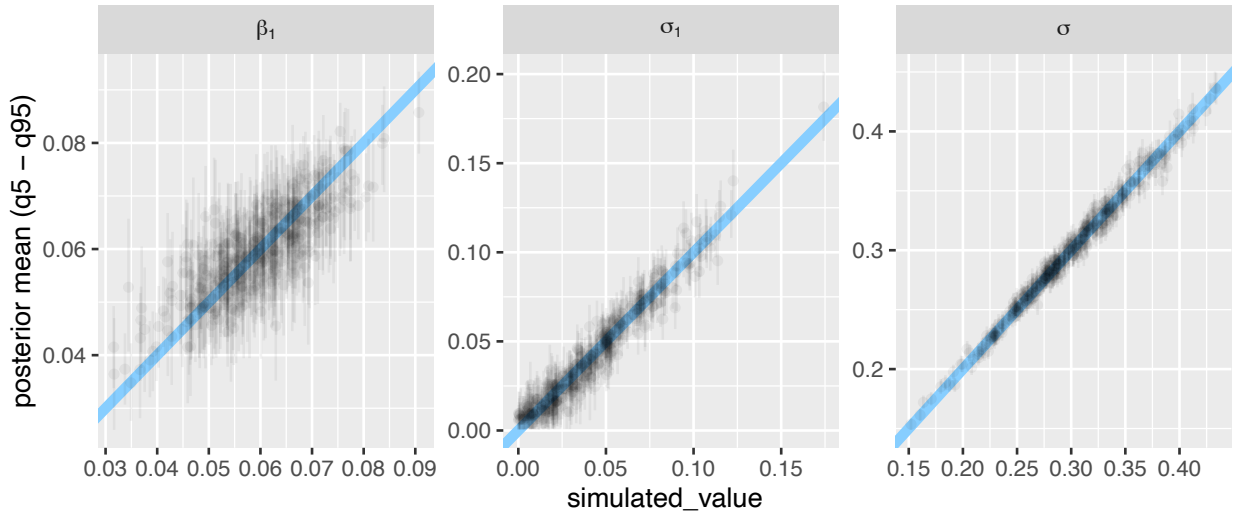


Figure S4: Model S2.1 was able to recover known parameter values, though with considerable uncertainty (with the exception of σ_1 and σ). Results are shown for n = 401 simulated data sets.

Model 2.2

The first heteroscedastic model is motivated similarly to the varying intercepts and slopes per study, and allowed varying study-level residuals around an overall average residual standard deviation. I estimated study-level residual variation independently of other varying study-level parameters for the mean:

$$S_{ij} \sim lognormal(\mu_{ij}, \sigma_i)$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i})X_{ij},$$

$$[\beta_{0i}, \beta_{1i}]' \sim MVN(0, \boldsymbol{SRS}),$$

$$\boldsymbol{S} = \begin{bmatrix} \sigma_{0i} & 0 \\ 0 & \sigma_{1i} \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}\sigma_{1i}} \\ \rho_{\sigma_{0i}\sigma_{1i}} & 1 \end{bmatrix},$$

$$log(\sigma_i) = \beta_0^{\sigma} + \beta_{0i}^{\sigma},$$

$$\beta_{0i}^{\sigma} \sim N(0, \sigma_{0i}^{\sigma}),$$

where β_0^σ is the overall average of residual variation (on a log-scale), and β_{0i}^σ are the varying study-level departures (for the residual variation) drawn from a normal distribution with zero mean and standard deviation σ_{0i}^σ. I fit the model with weakly regularizing priors:

$$\beta_0 \sim N(2.5, 1),$$

$$\beta_1, \beta_0^\sigma \sim N(0, 1),$$

$$[\sigma_{0i}, \sigma_{1i}, \sigma_{0i}^\sigma] \sim N(0, 1),$$

$$\mathbf{R} \sim LKJ(1). \qquad \text{(Model S2.2)}$$

This model had good convergence (all Rhat <= 1.01), and showed a reasonable fit to the empirical data (Appendix S3: Fig. S5).

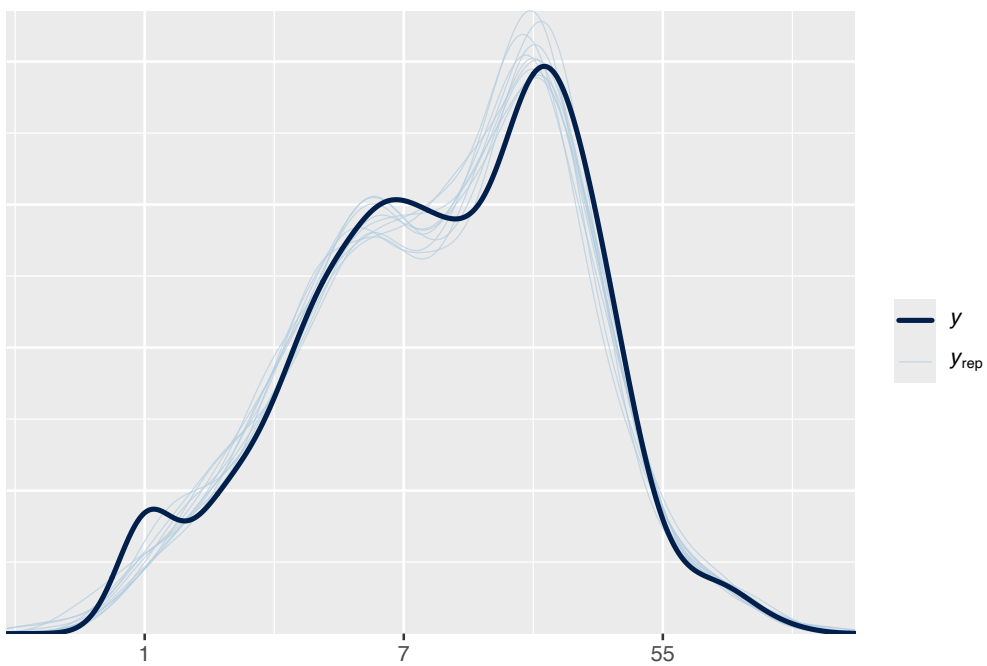


Figure S5: Posterior predictions from model 2.2 showed the model had a good ability to predict the observed data.

The parameter estimates from the fit of model S2.2 to empirical data (β_0: 2.48 [95% credible interval: 2.31 – 2.66]; β_1: 0.05 [95% credible interval: 0.04 – 0.07]; σ_{0i}: 1 [95% credible interval:

0.88 – 1.15]; σ_{1i}: 0.06 [95% credible interval: 0.04 – 0.07], β_0^{σ}: -1.23 [95% credible interval: -1.32 – -1.13], and σ_{0i}^{σ}: 0.44 [95% credible interval: 0.36 – 0.53]) were used to inform the following priors:

$$\beta_0 \sim N(2.5, 0.1),$$

$$\beta_1 \sim N(0.05, 0.01),$$

$$\sigma_{0i} \sim N(1, 0.1),$$

$$\sigma_{1i} \sim N(0.05, 0.05),$$

$$\beta_0^{\sigma} \sim N(-1.2, 0.1),$$

$$\sigma_{0i}^{\sigma} \sim N(0.4, 0.1).$$

Simulation-based calibration for model 2.2 showed that the rank statistics were approximately uniformly distributed (Appendix S3: Fig. S6), and that the coverage of parameters was reasonable (Appendix S3: Fig. S7); known parameters were approximately recovered (Appendix S3: Fig. S8).

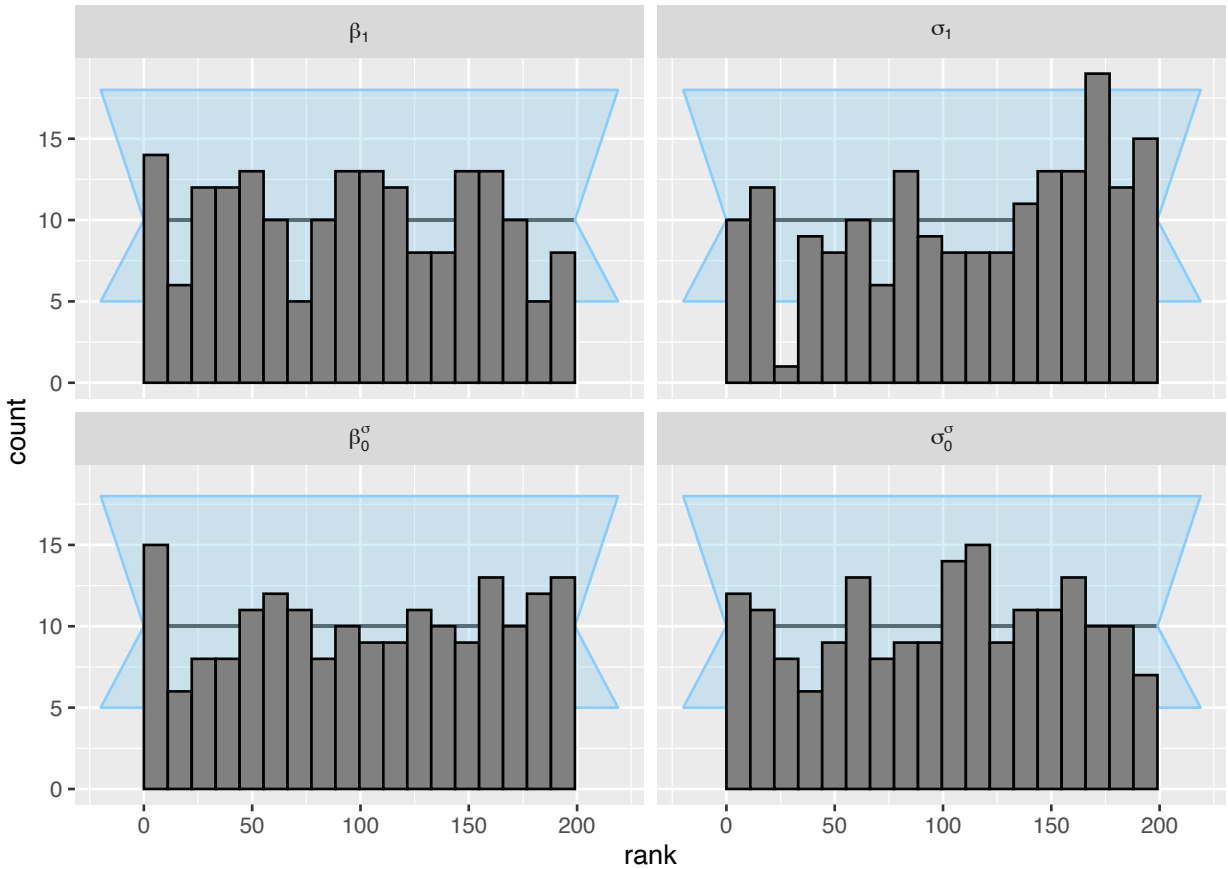


Figure S6: The posterior ranks of the prior draws were approximately normally distributed for the parameters of interest in model S2.2. Results are shown for n = 185 simulated data sets.

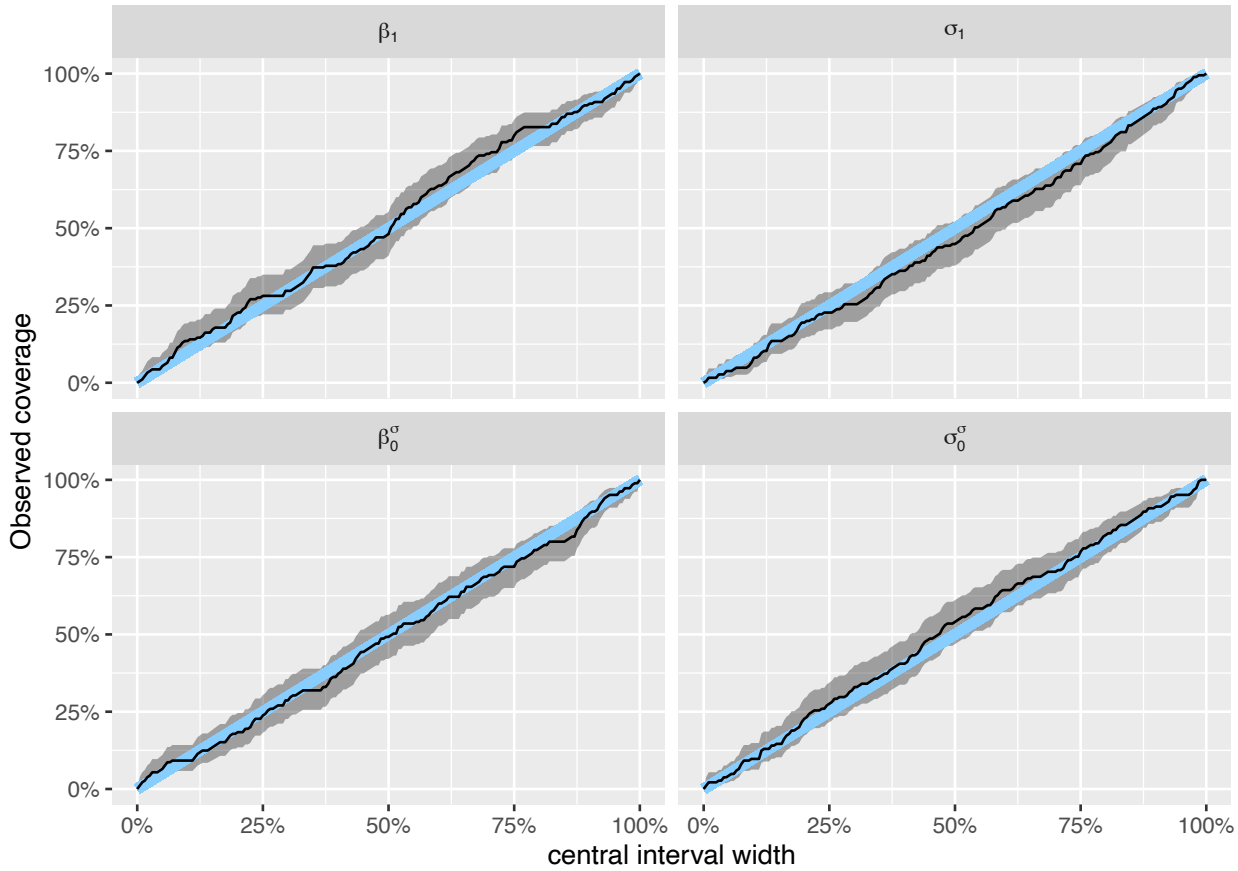


Figure S7: Model S2.2 had good coverage for the parameters of interest. Results are shown for n = 185 simulated data sets.

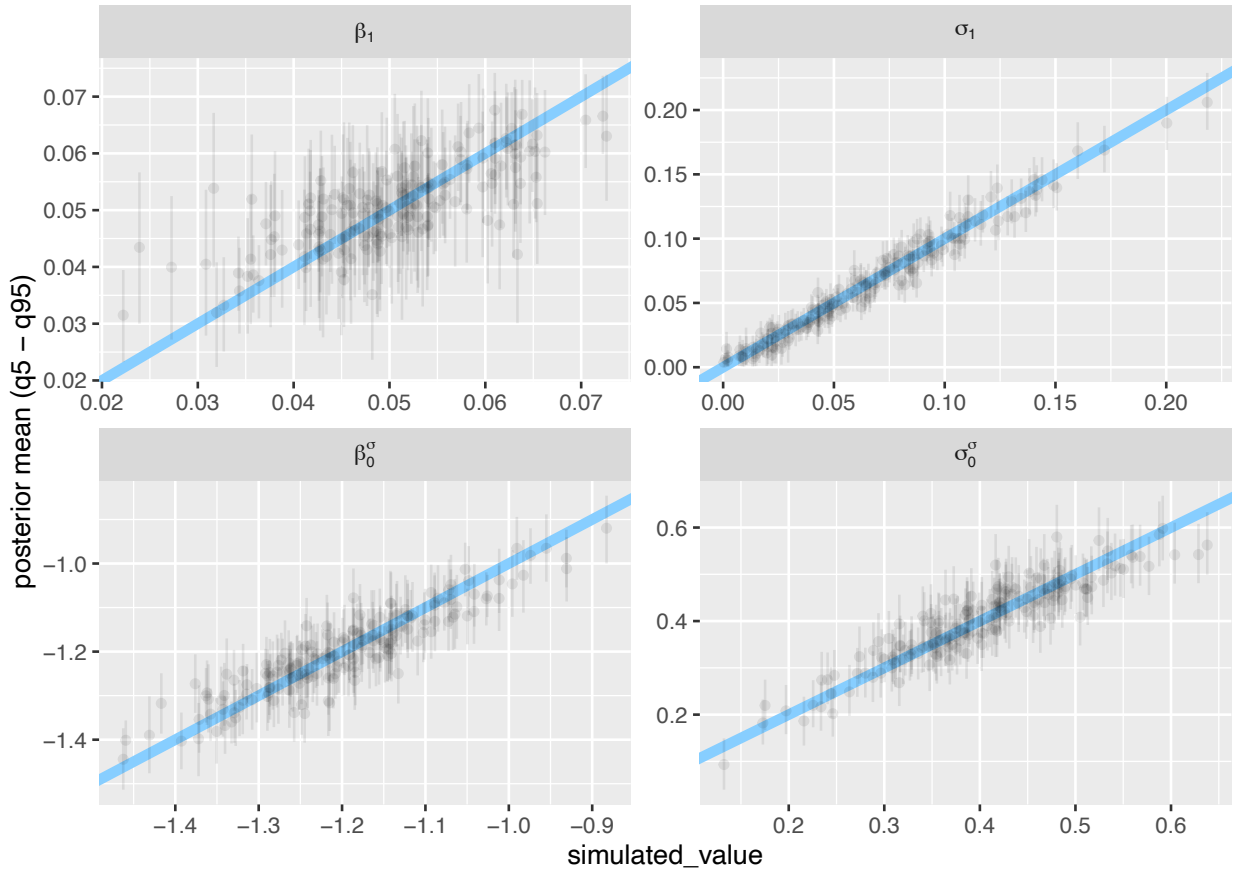


Figure S8: Model S2.2 was able to recover known parameter values, though with considerable uncertainty (with the exception of σ_1). Results are shown for n = 185 simulated data sets.

Model 2.3

Next, I extend this model to include (log) fragment size as a predictor of residual variation, and again allow it to varying between studies independently of varying study-level parameters for the mean:

$$S_{ij} \sim lognormal(\mu_{ij}, \sigma_i)$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i})X_{ij},$$

$$[\beta_{0i}, \beta_{1i}]' \sim MVN(0, \boldsymbol{SRS}),$$

$$\boldsymbol{S} = \begin{bmatrix} \sigma_{0i} & 0 \\ 0 & \sigma_{1i} \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}\sigma_{1i}} \\ \rho_{\sigma_{0i}\sigma_{1i}} & 1 \end{bmatrix},$$

$$log(\sigma_i) = \beta_0^\sigma + \beta_{0i}^\sigma + (\beta_1^\sigma + \beta_{1i}^\sigma)X_i,$$

$$[\beta_{0i}^\sigma, \beta_{1i}^\sigma]' \sim MVN(0, \boldsymbol{S^\sigma R^\sigma S^\sigma}),$$

$$\boldsymbol{S^\sigma} = \begin{bmatrix} \sigma_{0i}^\sigma & 0 \\ 0 & \sigma_{1i}^\sigma \end{bmatrix},$$

$$\boldsymbol{R^\sigma} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}^\sigma\sigma_{1i}^\sigma} \\ \rho_{\sigma_{0i}^\sigma\sigma_{1i}^\sigma} & 1 \end{bmatrix},$$

where β_0^σ is the overall average residual variation, and β_1^σ is the overall average slope of residual variation with fragment size; β_{0i}^σ and β_{1i}^σ are the varying study-level departures from the intercept and slope, respectively, and were drawn from a multivariate normal distribution with zero mean and standard deviation σ_{0i}^σ and σ_{1i}^σ, with correlations estimated in matrix $\boldsymbol{R^\sigma}$. I fit the model with weakly regularizing priors:

$$\beta_0 \sim N(2.5, 1),$$

$$\beta_1, \beta_0^\sigma, \beta_1^\sigma \sim N(0, 1),$$

$$[\sigma_{0i}, \sigma_{1i}] \sim N(0, 1),$$

$$\mathbf{R} \sim LKJ(1),$$

$$[\sigma_{0i}^\sigma, \sigma_{1i}^\sigma] \sim N(0,1),$$

$$\boldsymbol{R^\sigma} \sim \text{LKJ}(1). \qquad \text{(Model S2.3)}$$

This model had good convergence (all Rhat <= 1.02), and showed a reasonable fit to the empirical data (Appendix S3: Fig. S9).

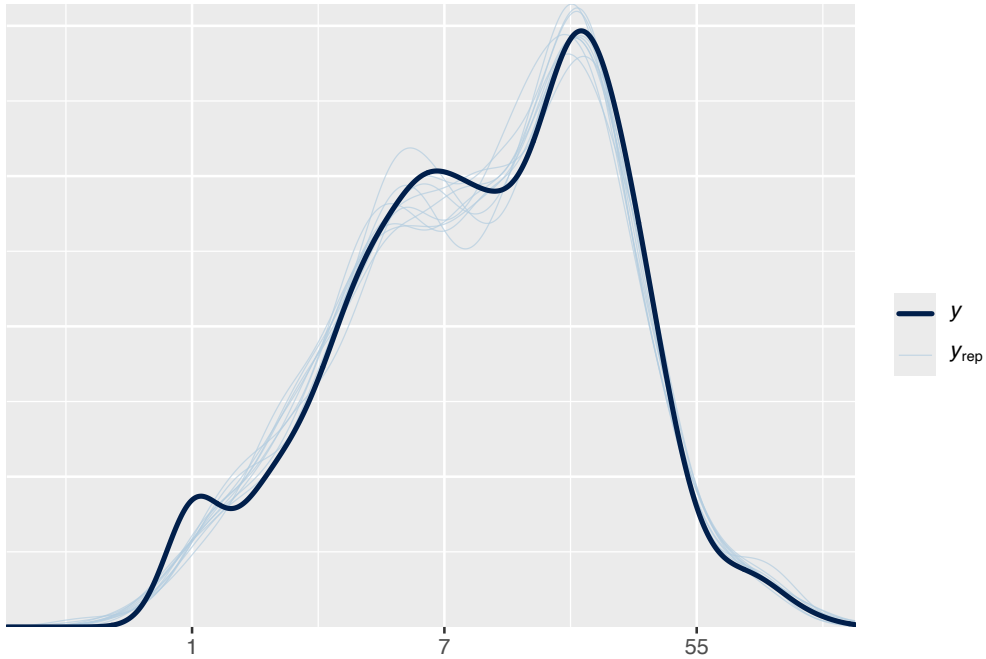


Figure S9: Posterior predictions from model 2.3 showed good fidelity to the observed data.

The parameter estimates from the fit of model S2.3 to empirical data (β_0: 2.48 [95% credible interval: 2.3 – 2.66]; β_1: 0.05 [95% credible interval: 0.03 – 0.06]; σ_{0i}: 1.01 [95% credible interval: 0.89 – 1.14]; σ_{1i}: 0.06 [95% credible interval: 0.04 – 0.07]; β_0^{σ}: -1.25 [95% credible interval: -1.35 – -1.15]; β_1^{σ}: -0.06 [95% credible interval: -0.09 – -0.03]; σ_{0i}^{σ}: 0.45 [95% credible interval: 0.37 – 0.55], and σ_{1i}^{σ}: 0.06 [95% credible interval: 0.01 – 0.1]) were used to inform the following priors:

$$\beta_0 \sim N(2.5, 0.1),$$

$$\beta_1 \sim N(0.05, 0.01),$$

$$\sigma_{0i} \sim N(1, 0.1),$$

$$\sigma_{1i} \sim N(0.05, 0.05),$$

$$\beta_0^{\sigma} \sim N(-1.2, 0.1),$$

$$\sigma_{0i}^{\sigma} \sim N(0.4, 0.1).$$

Simulation-based calibration for model 2.3 showed that the rank statistics were approximately uniformly distributed (Appendix S3: Fig. S10), and that the coverage of parameters was reasonable (Appendix S3: Fig. S11); known parameters were approximately recovered (Appendix S3: Fig. S12).

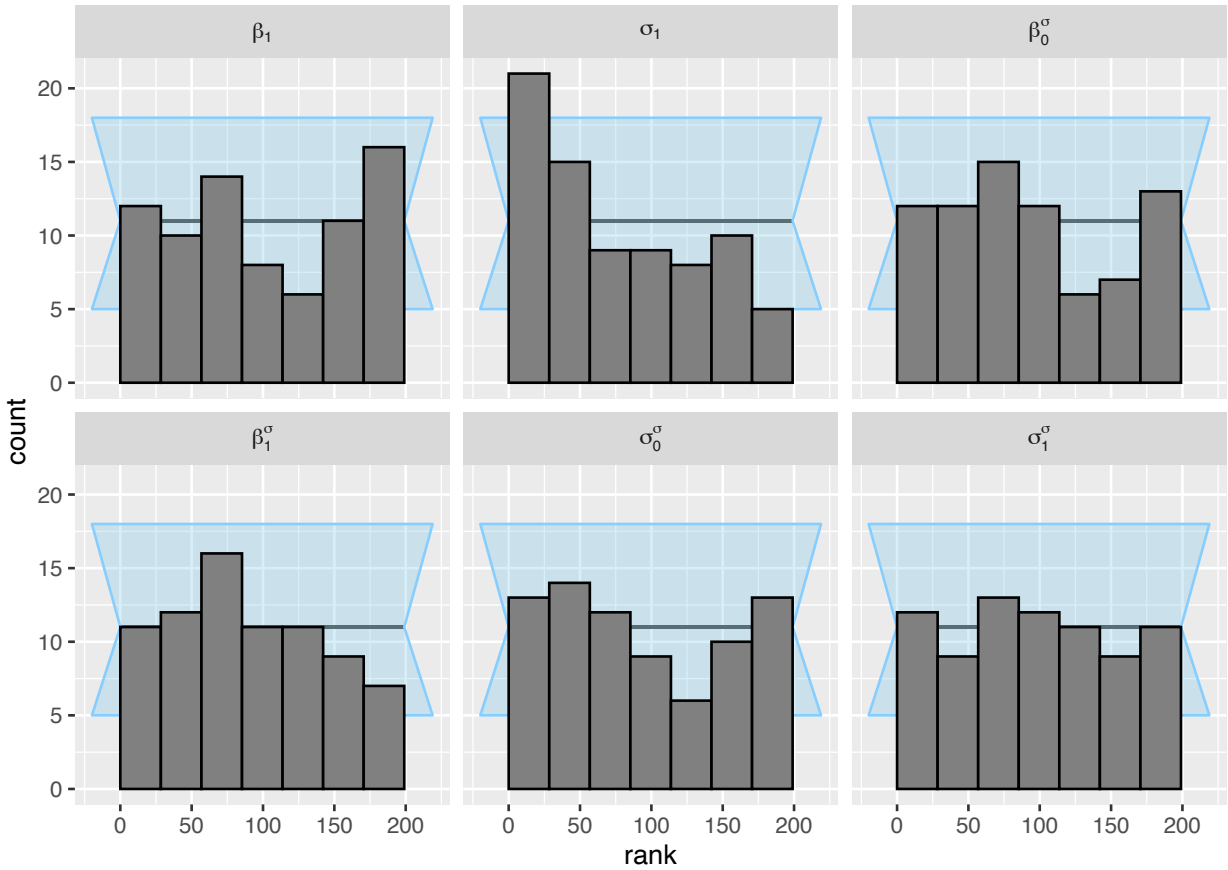


Figure S10: The posterior ranks of the prior draws were approximately normally distributed for the parameters of interest in model S2.3. Results are shown for n = 77 simulated data sets.

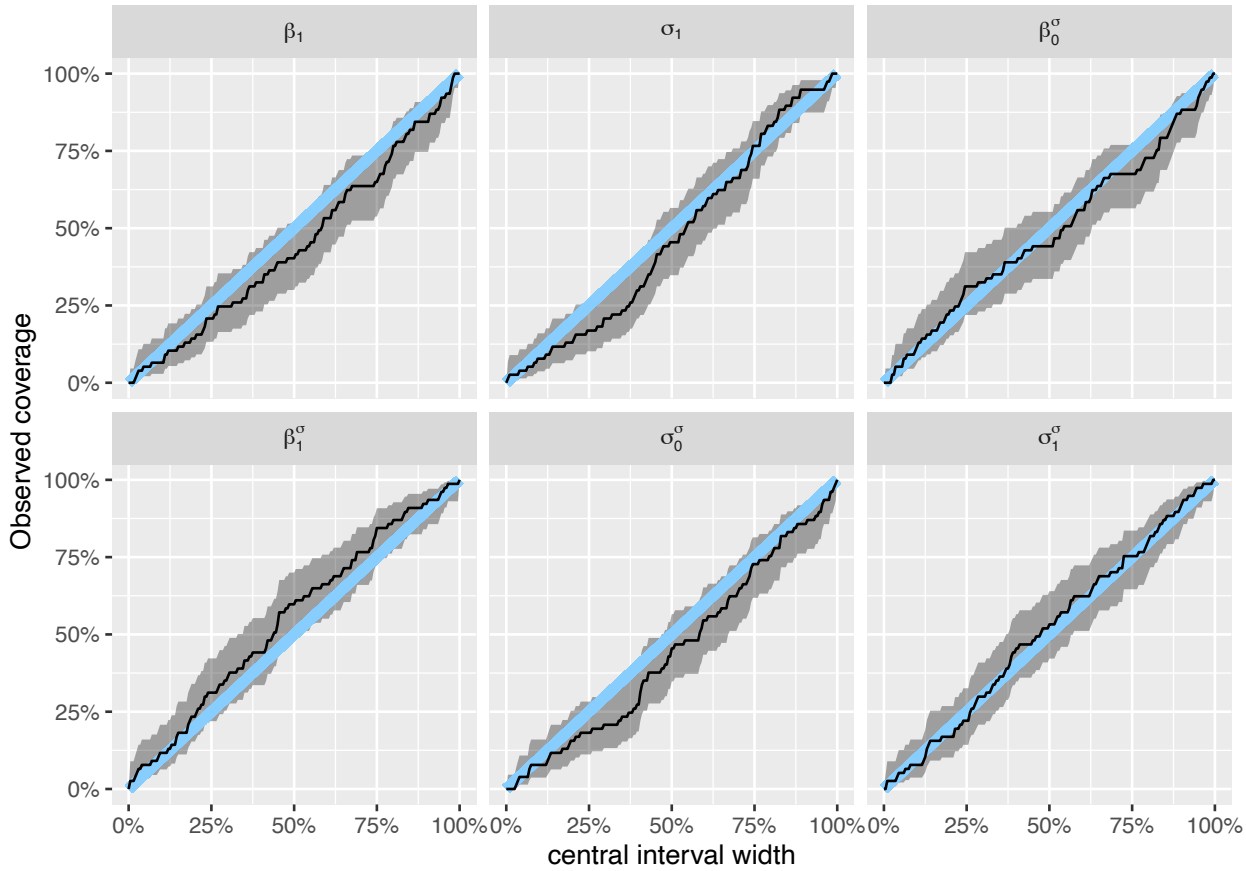


Figure S11: Model S2.3 had good coverage for all parameters of interest. Results are shown for n = 77 simulated data sets.

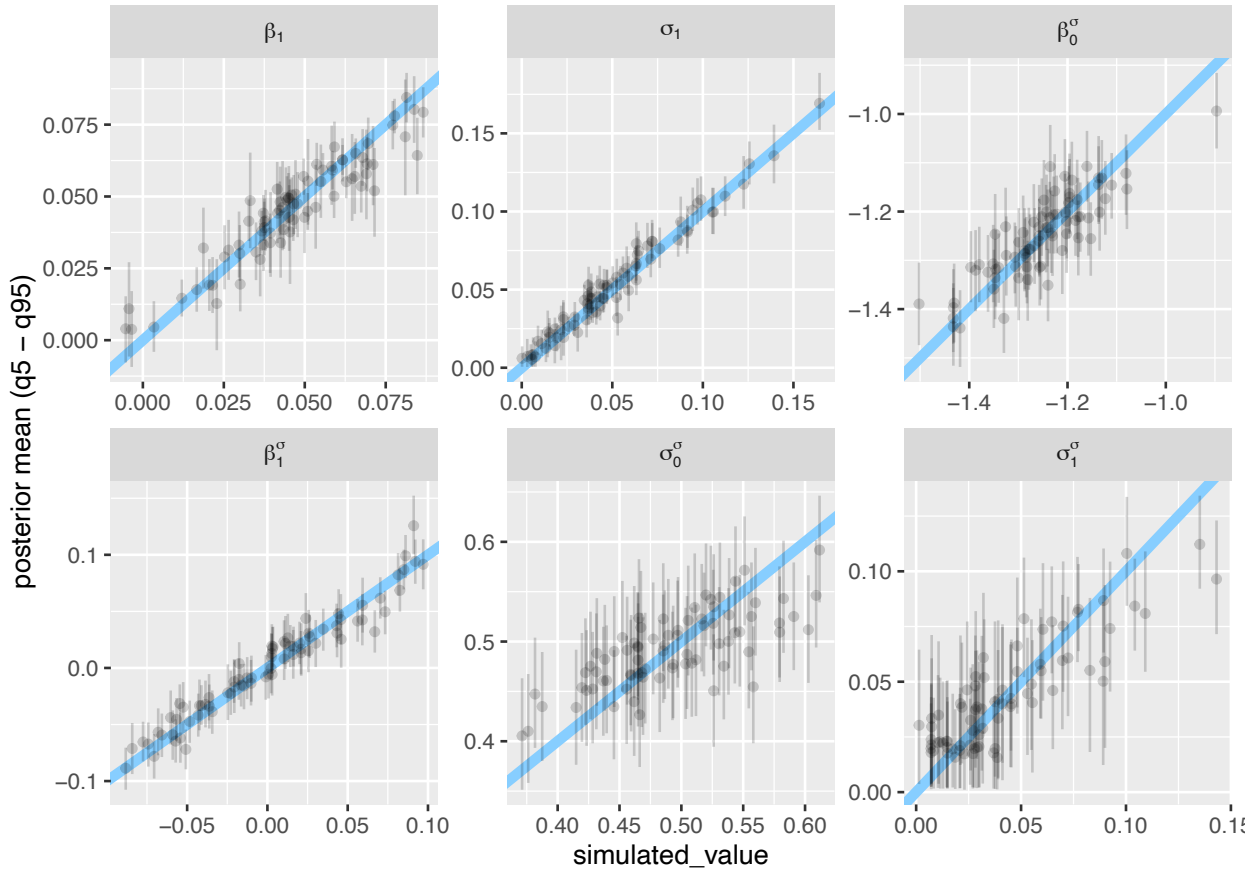


Figure S12: Model S2.3 was able to recover most parameters with reasonable accuracy, albeit with considerable uncertainty. Results are shown for n = 77 simulated data sets.

Model 2.4

The next model is the same as model 2.2 (i.e., varying study-level residuals), but allows the study-level residuals to covary with the other varying parameters for the mean:

$$S_{ij} \sim lognormal(\mu_{ij}, \sigma_i)$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i})X_{ij},$$

$$log(\sigma_i) = \beta_0^\sigma + \beta_{0i}^\sigma,$$

$$[\beta_{0i}, \beta_{1i}, \beta_{0i}^\sigma]' \sim MVN(0, \boldsymbol{SRS}),$$

$$\boldsymbol{S} = \begin{bmatrix} \sigma_{0i} & 0 & 0 \\ 0 & \sigma_{1i} & 0 \\ 0 & 0 & \sigma_{0i}^\sigma \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}\sigma_{1i}} & \rho_{\sigma_{0i}\sigma_{0i}^{\sigma}} \\ \rho_{\sigma_{0i}\sigma_{1i}} & 1 & \rho_{\sigma_{1i}\sigma_{0i}^{\sigma}} \\ \rho_{\sigma_{0i}\sigma_{0i}^{\sigma}} & \rho_{\sigma_{1i}\sigma_{0i}^{\sigma}} & 1 \end{bmatrix}.$$

The model was fit to empirical data with weakly regularizing priors:

$$\beta_0 \sim N(2.5, 1),$$

$$\beta_1, \beta_0^{\sigma} \sim N(0, 1),$$

$$[\sigma_{0i}, \sigma_{1i}, \sigma_{0i}^{\sigma}] \sim N(0, 1),$$

$$\mathbf{R} \sim LKJ(1). \qquad \text{(Model S2.4)}$$

This model had good convergence (all Rhat <= 1.02), and showed a reasonable fit to the empirical data (Appendix S3: Fig. Sx).

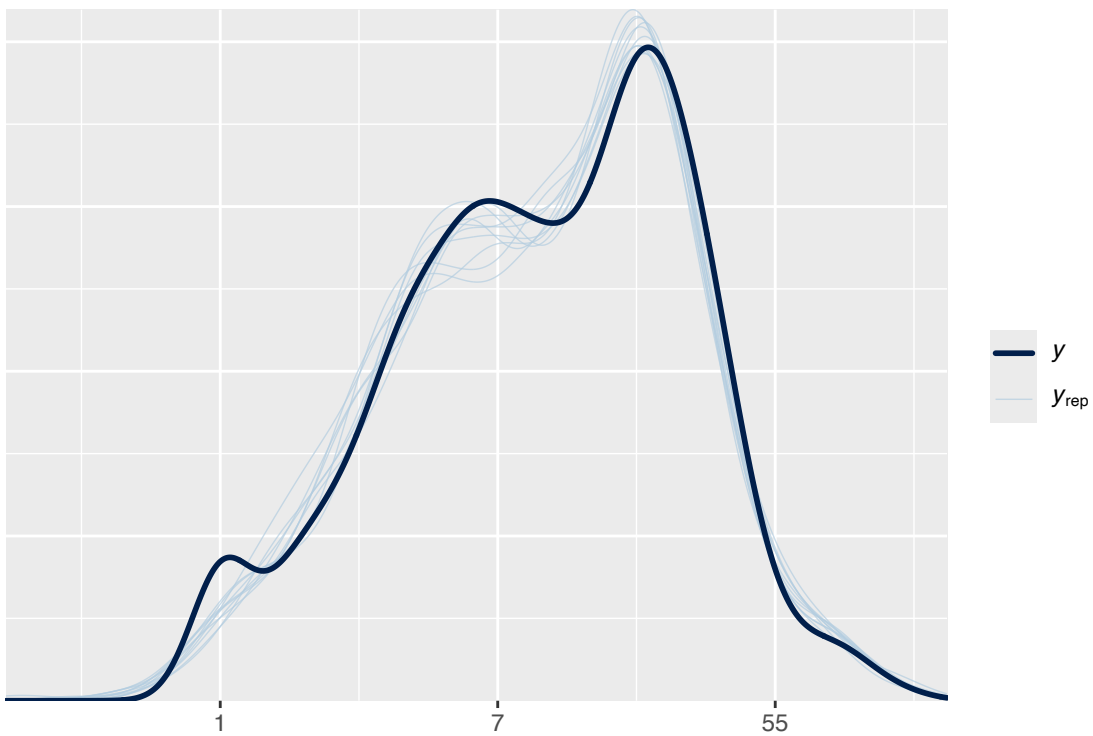


Figure S13: Posterior predictions showed that model 2.4 was able to make predict the observed data well.

The parameter estimates from the fit of model S2.4 to empirical data (β_0: 2.46 [95% credible interval: 2.29 – 2.66]; β_1: 0.05 [95% credible interval: 0.04 – 0.07]; σ_{0i}: 1 [95% credible

interval: 0.89 – 1.13]; σ_{1i}: 0.06 [95% credible interval: 0.04 – 0.07], β_0^{σ}: -1.24 [95% credible interval: -1.35 – -1.15], and σ_{0i}^{σ}: 0.44 [95% credible interval: 0.37 – 0.53]) were used to inform the following priors:

$$\beta_0 \sim N(2.5, 0.1),$$

$$\beta_1 \sim N(0.05, 0.01),$$

$$\sigma_{0i} \sim N(1, 0.1),$$

$$\sigma_{1i} \sim N(0.05, 0.05),$$

$$\beta_0^{\sigma} \sim N(-1.2, 0.1),$$

$$\sigma_{0i}^{\sigma} \sim N(0.4, 0.1).$$

Simulation-based calibration for model 2.4 showed that the rank statistics were approximately uniformly distributed (Appendix S3: Fig. S14), and that the coverage of parameters was reasonable (Appendix S3: Fig. S15); known parameters were approximately recovered (Appendix S3: Fig. S16).

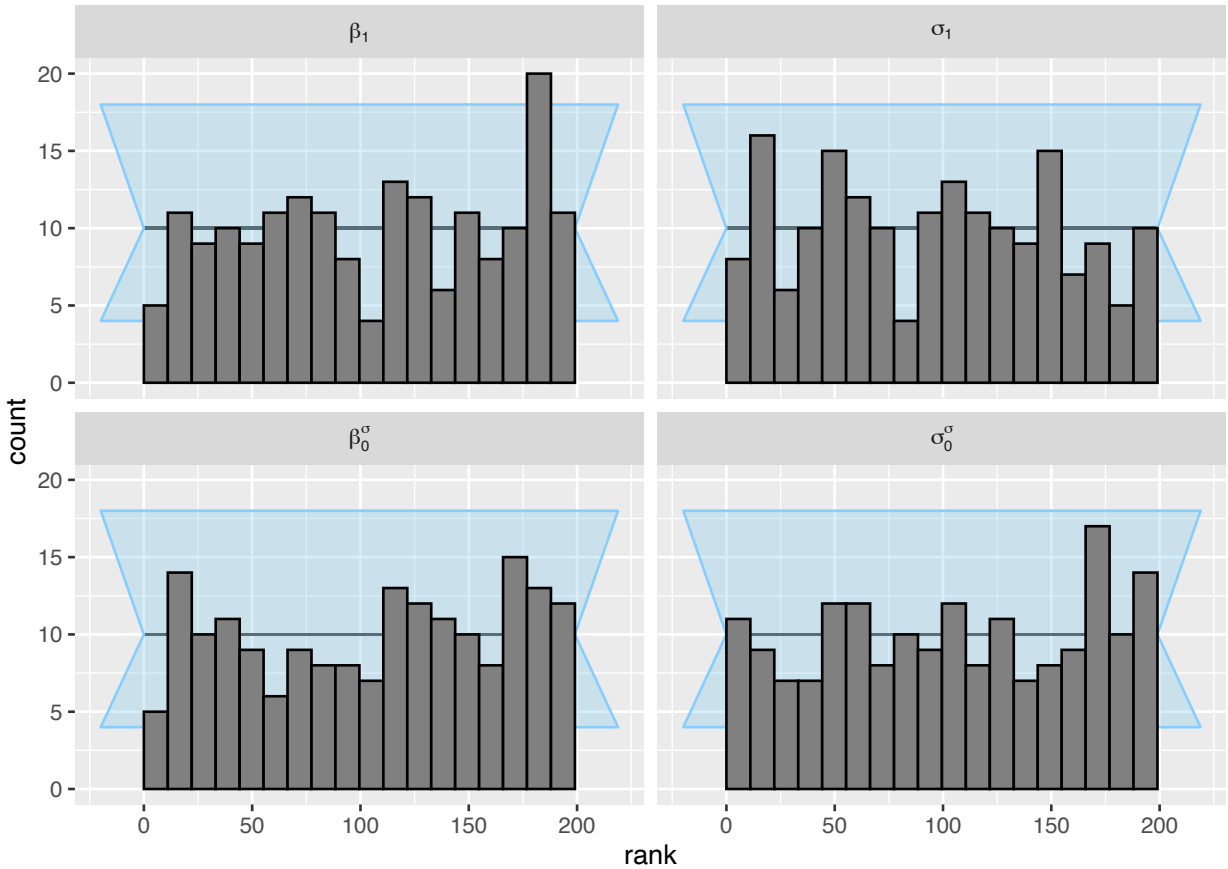


Figure S14: The posterior ranks of the prior draws were approximately normally distributed for the parameters of interest in model S2.4. Results are shown for n = 181 simulated data sets.

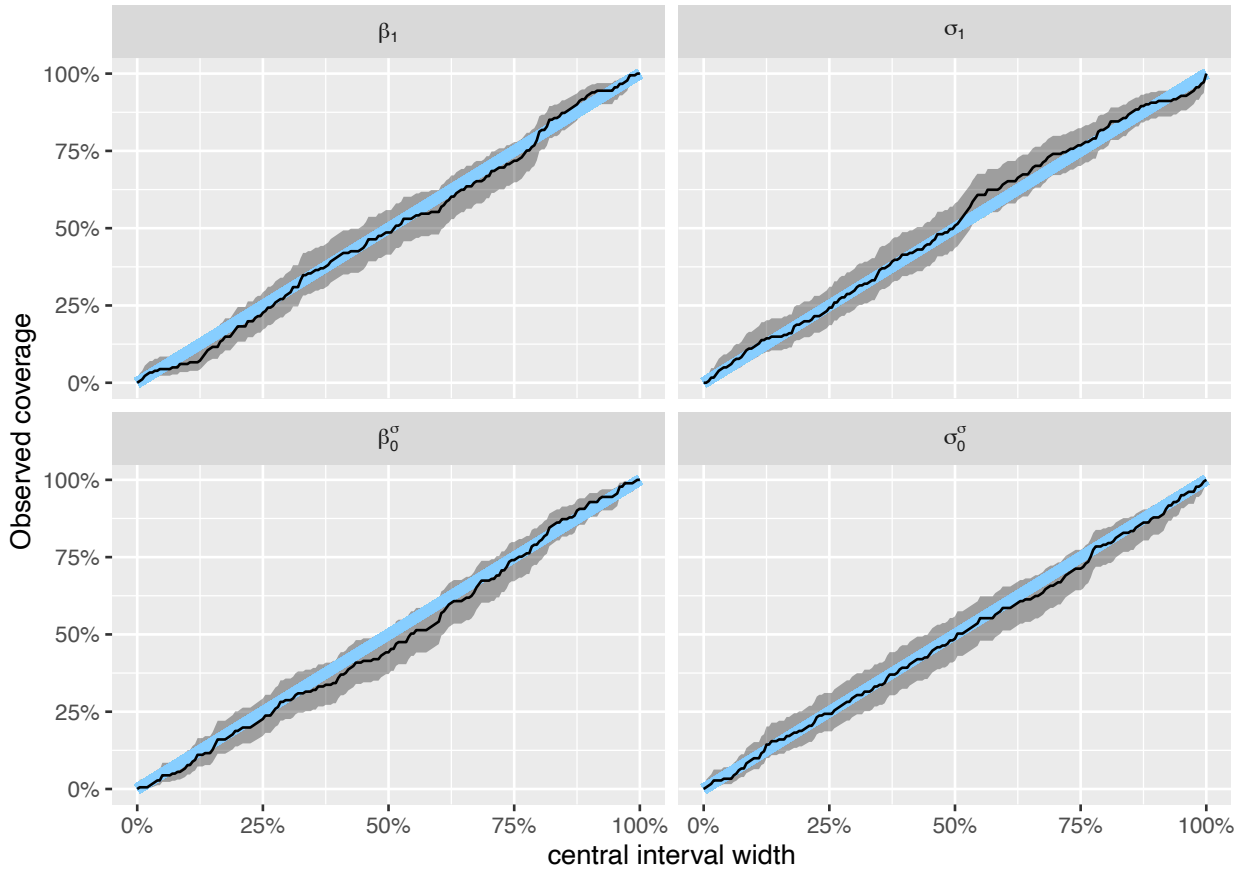


Figure S15: Model S2.4 had good coverage for all parameters of interest. Results are shown for n = 181 simulated data sets.

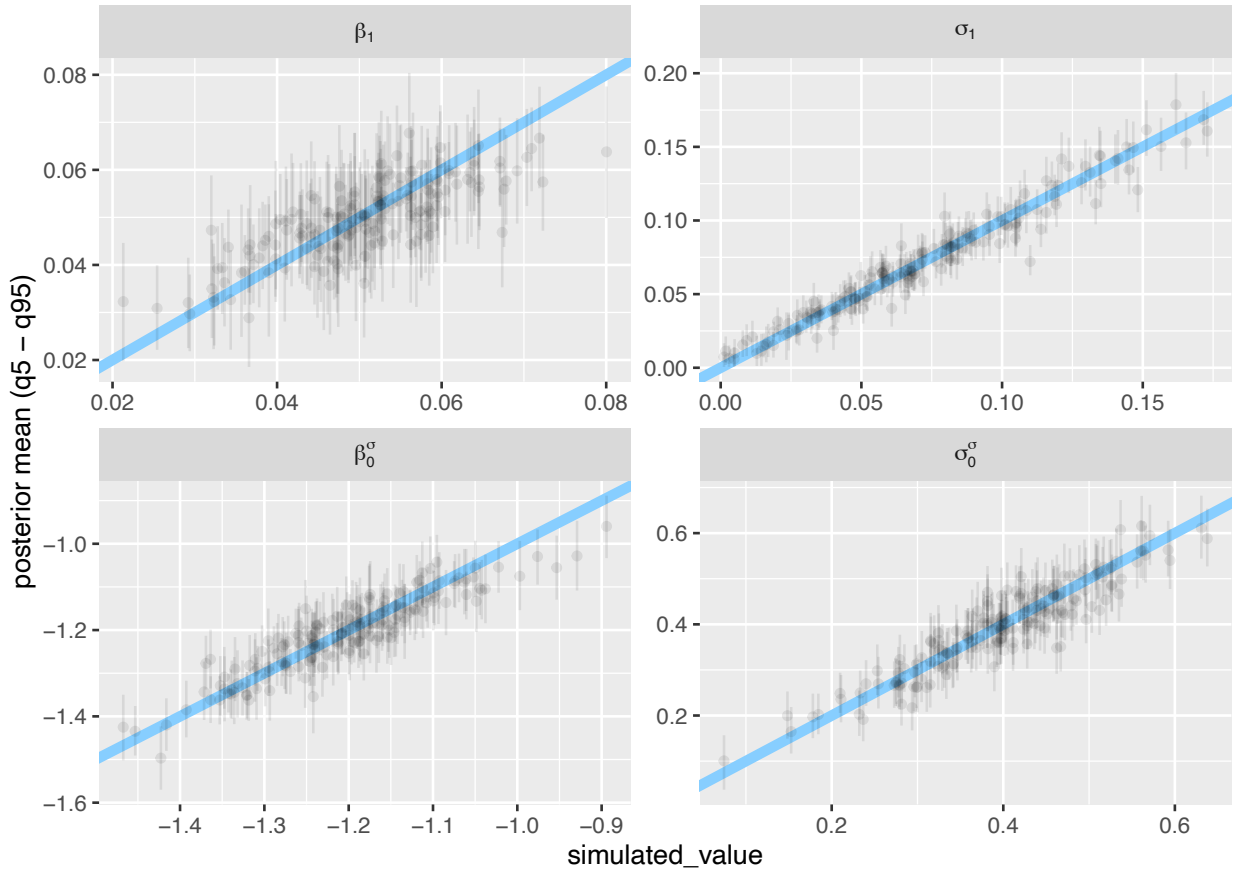


Figure S16: Model S2.4 was able to recover known parameter values with reasonable accuracy. Results are shown for n = 181 simulated data sets.

Model 2.5

The final model extended model 2.3 to allow for correlations between all of the varying study-level parameters (i.e., for the location and the scale):

$$S_{ij} \sim lognormal(\mu_{ij}, \sigma_i)$$

$$\mu_{ij} = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i})X_{ij},$$

$$log(\sigma_i) = \beta_0^\sigma + \beta_{0i}^\sigma + (\beta_1^\sigma + \beta_{1i}^\sigma)X_i,$$

$$[\beta_{0i}, \beta_{1i}, \beta_{0i}^\sigma, \beta_{1i}^\sigma]' \sim MVN(0, \boldsymbol{SRS}),$$

$$\boldsymbol{S} = \begin{bmatrix} \sigma_{0i} & 0 & 0 & 0 \\ 0 & \sigma_{1i} & 0 & 0 \\ 0 & 0 & \sigma_{0i}^\sigma & 0 \\ 0 & 0 & 0 & \sigma_{1i}^\sigma \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\sigma_{0i}\sigma_{1i}} & \rho_{\sigma_{0i}\sigma_{0i}^{\sigma}} & \rho_{\sigma_{0i}\sigma_{1i}^{\sigma}} \\ \rho_{\sigma_{0i}\sigma_{1i}} & 1 & \rho_{\sigma_{1i}\sigma_{0i}^{\sigma}} & \rho_{\sigma_{1i}\sigma_{1i}^{\sigma}} \\ \rho_{\sigma_{0i}\sigma_{0i}^{\sigma}} & \rho_{\sigma_{1i}\sigma_{0i}^{\sigma}} & 1 & \rho_{\sigma_{0i}^{\sigma}\sigma_{1i}^{\sigma}} \\ \rho_{\sigma_{0i}\sigma_{1i}^{\sigma}} & \rho_{\sigma_{1i}\sigma_{1i}^{\sigma}} & \rho_{\sigma_{0i}^{\sigma}\sigma_{1i}^{\sigma}} & 1 \end{bmatrix}.$$

I fit the model with weakly regularizing priors:

$$\beta_0 \sim N(2.5, 1),$$

$$\beta_1, \beta_0^{\sigma}, \beta_1^{\sigma} \sim N(0, 1),$$

$$[\sigma_{0i}, \sigma_{1i}, \sigma_{0i}^{\sigma}, \sigma_{1i}^{\sigma}] \sim N(0, 1),$$

$$\mathbf{R} \sim LKJ(1). \qquad \text{(Model S2.5)}$$

This model had good convergence (all Rhat < 1.01), and showed a reasonable fit to the empirical data (Appendix S3: Fig. S17).

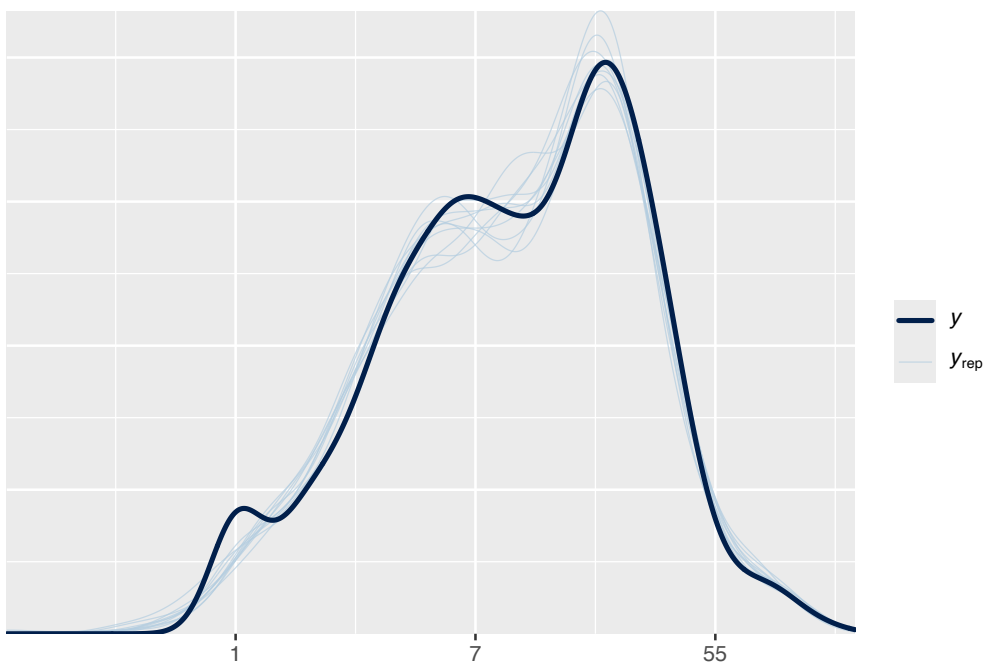


Figure S17: Model 2.5 made predictions that resembled the observed data closely.

The parameter estimates from the fit of model S2.5 to empirical data (β_0: 2.47 [95% credible interval: 2.3 – 2.66]; β_1: 0.05 [95% credible interval: 0.03 – 0.06]; σ_{0i}: 1.01 [95% credible interval: 0.89 – 1.15]; σ_{1i}: 0.06 [95% credible interval: 0.04 – 0.07]; β_0^{σ}: -1.28 [95%

credible interval: -1.39 – -1.18]; β_1^{σ}: -0.06 [95% credible interval: -0.09 – -0.04]; σ_{0i}^{σ}: 0.46 [95% credible interval: 0.39 – 0.56], and σ_{1i}^{σ}: 0.07 [95% credible interval: 0.03 – 0.11]) were used to inform the following priors:

$$\beta_0 \sim N(2.5, 0.1),$$

$$\beta_1 \sim N(0.05, 0.01),$$

$$\sigma_{0i} \sim N(1, 0.1),$$

$$\sigma_{1i} \sim N(0.05, 0.05),$$

$$\beta_0^{\sigma} \sim N(-1.2, 0.1),$$

$$\sigma_{0i}^{\sigma} \sim N(0.4, 0.1).$$

Simulation-based calibration for model 2.5 showed that the rank statistics were approximately uniformly distributed (Appendix S3: Fig. S18), and that the coverage of parameters was reasonable (Appendix S3: Fig. S19); known parameters were approximately recovered (Appendix S3: Fig. S20).

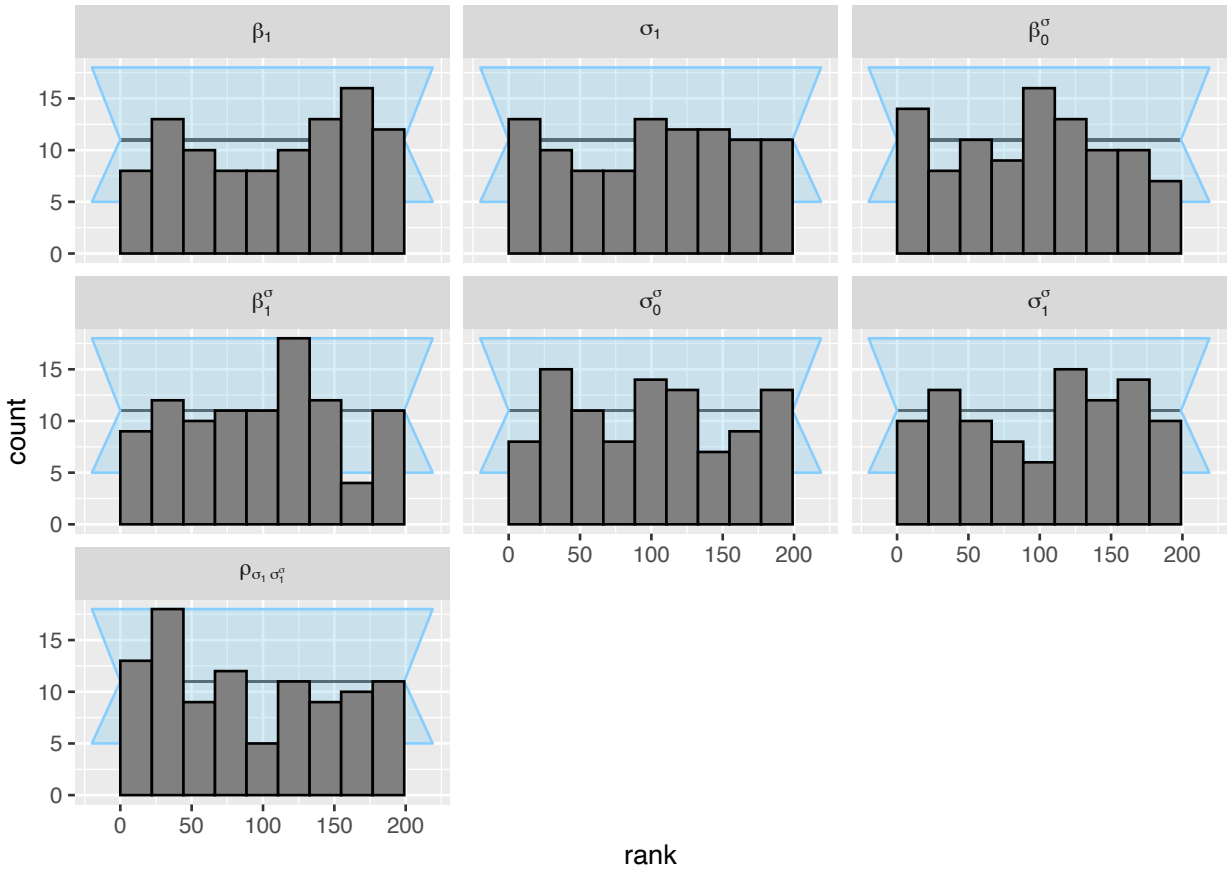


Figure S18: The posterior ranks of the prior draws were approximately normally distributed for the parameters of interest in model S2.5. Results are shown for n = 98 simulated data sets.

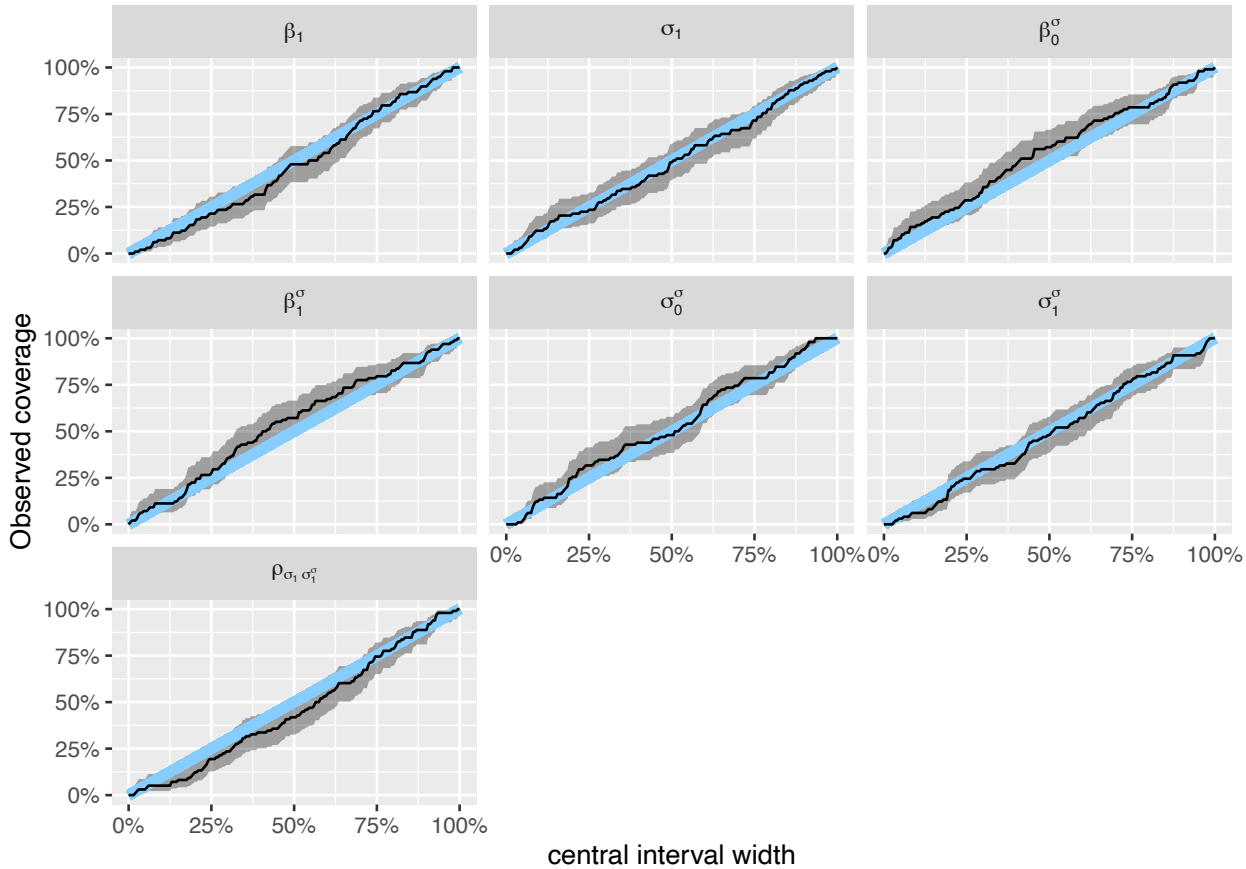


Figure S19: Model S2.5 had good coverage for all parameters of interest. Results are shown for n = 98 simulated data sets.

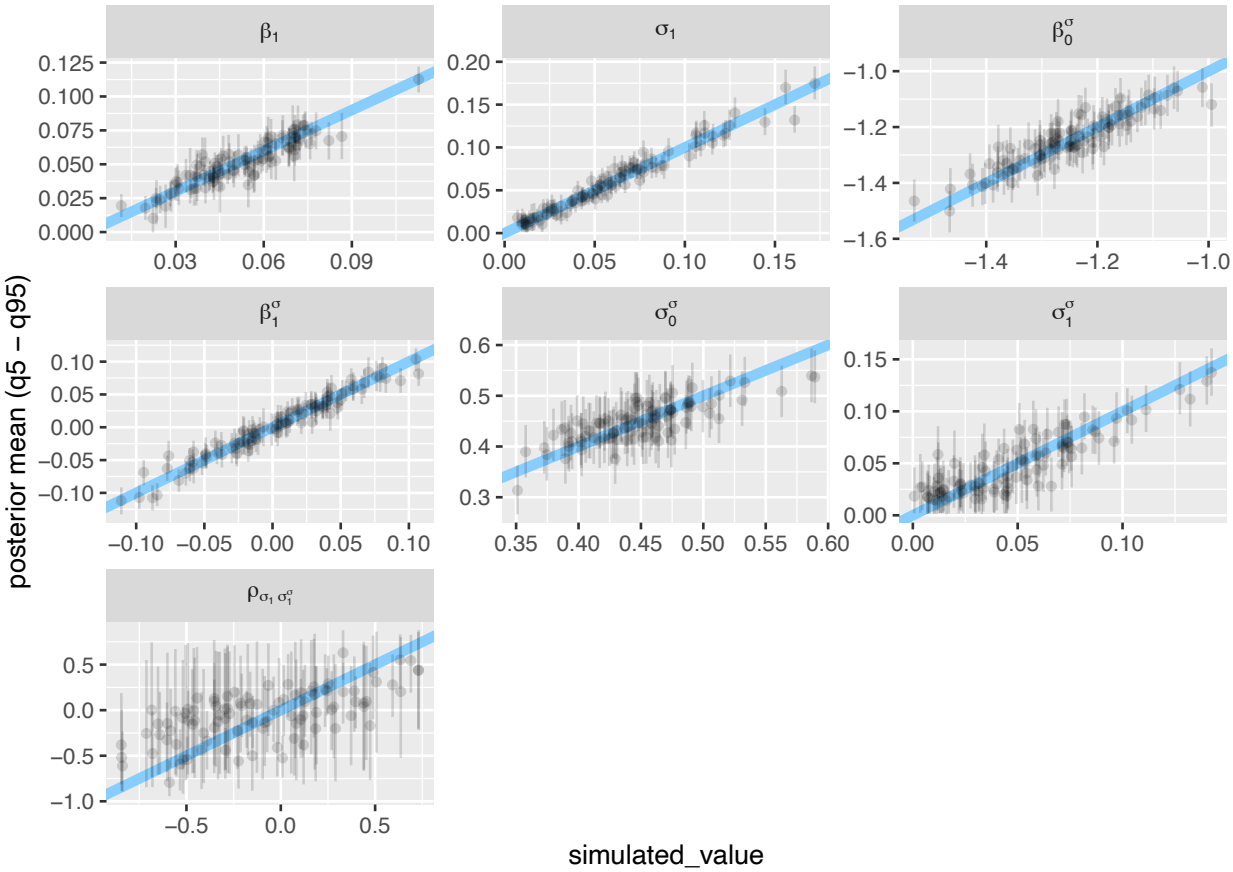


Figure S20: Model S2.5 was able to recover known parameter values with reasonable accuracy, with the exception of the correlation ($\rho_{\sigma_1\sigma_1^\sigma}$), which was estimated with high uncertainty. Results are shown for n = 98 simulated data sets.

1162

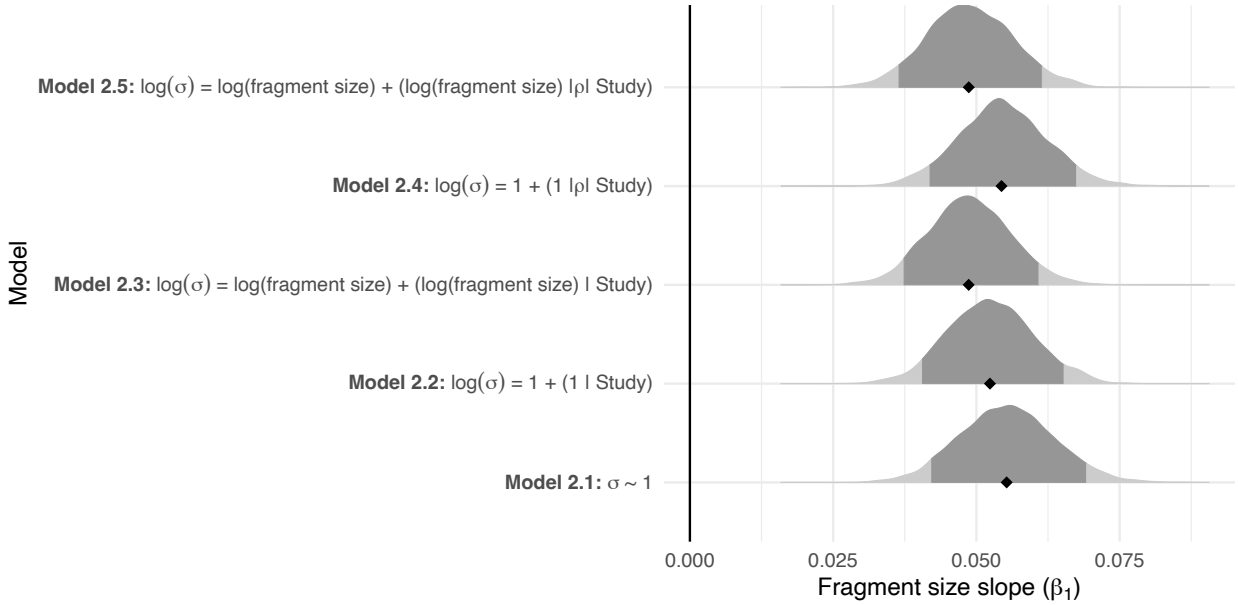


Figure S21: All models fit in case study two supported the ecosystem decay hypothesis.

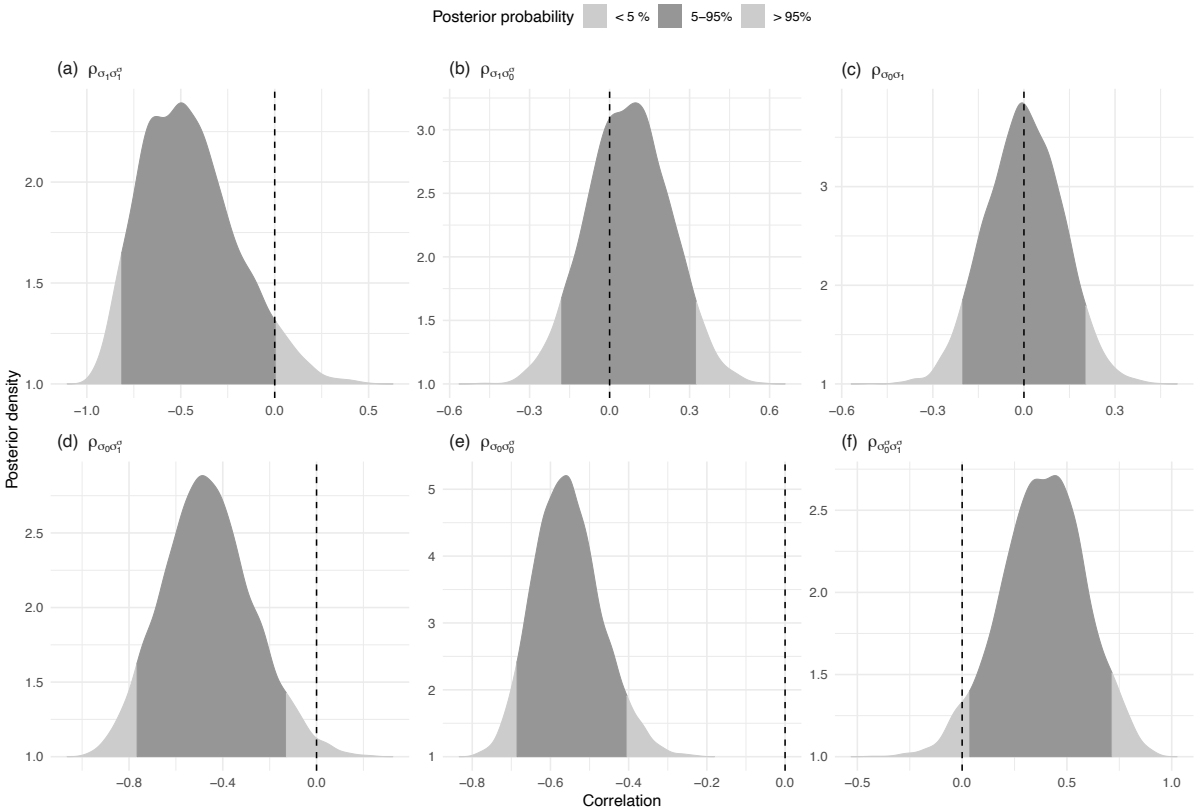


Figure S21: Posterior distributions (density plots for 1000 posterior draws) showing correlations among the varying study-level parameters.

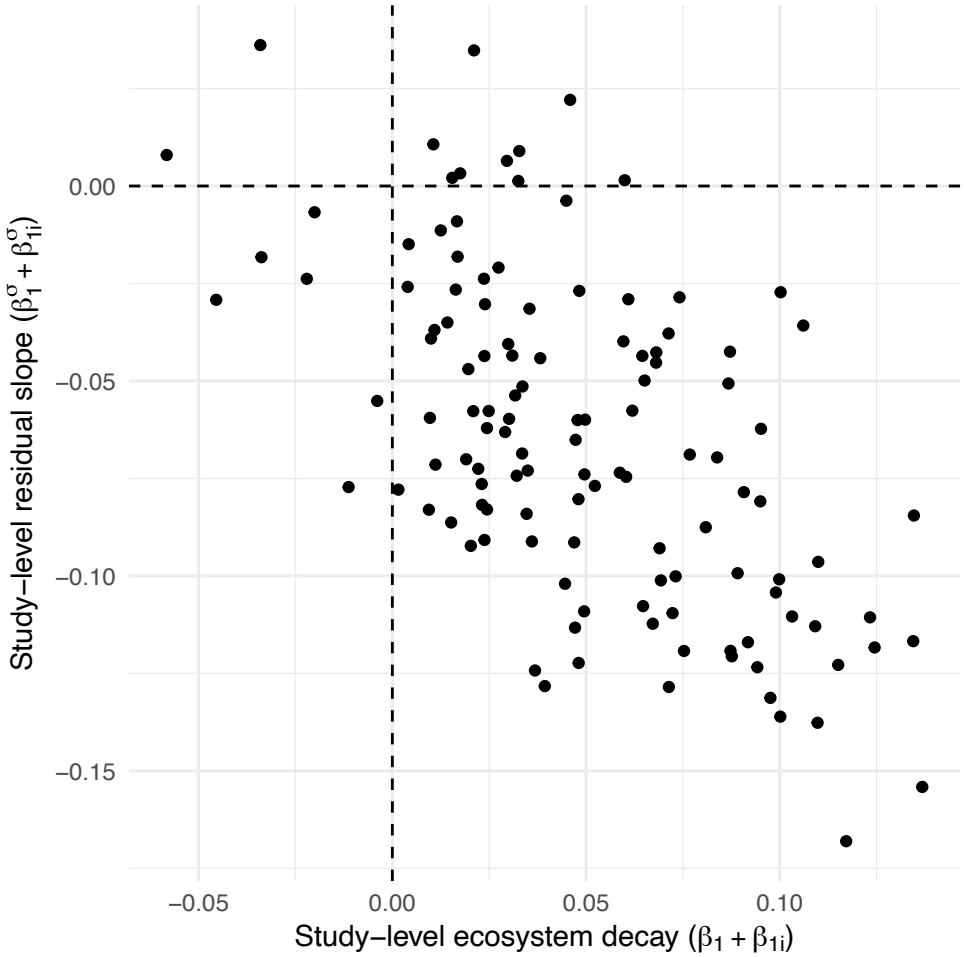


Figure S22: The decline of residual variation (scale) with increasing patch size was greatest for studies with the strongest ecosystem decay. Each point shows the estimate of a study-level slope, with slopes for the location on the x-axis, and the scale on the y-axis.

References

Chase, J. M., S. A. Blowes, T. M. Knight, K. Gerstner, and F. May. 2020. Ecosystem decay exacerbates biodiversity loss with habitat loss. Nature 584:238–243.

Modrák, M., A. H. Moon, S. Kim, P. Bürkner, N. Huurre, K. Faltejsková, A. Gelman, and A. Vehtari. 2023. Simulation-based calibration checking for Bayesian computation: The choice of test quantities shapes sensitivity. Bayesian Analysis 1:1–28.

Talts, S., M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. 2020, October 21. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv.