Phylogenetic multilevel meta-analysis: A simulation study on the importance of modeling the phylogeny

Ozan Cinar, Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Faculty of Health, Medicine, and Life Sciences, Maastricht University, Vijverdalseweg 1, 6226 NB Maastricht, the Netherlands, ozan.cinar@maastrichtuniversity.nl

Shinichi Nakagawa, Evolution & Ecology Centre, and School of Biological, Earth and Environmental Sciences, BEES, University of New South Wales, Randwick NSW 2052, Sydney, Australia, s.nakagawa@unsw.edu.au

Wolfgang Viechtbauer, Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Faculty of Health, Medicine, and Life Sciences, Maastricht University, Vijverdalseweg 1, 6226 NB Maastricht, the Netherlands, wolfgang.viechtbauer@maastrichtuniversity.nl

Corresponding author: Ozan Cinar, Address: Vijverdalseweg 1, 6226 NB Maastricht, the Netherlands, Email: ozan.cinar@maastrichtuniversity.nl

Running Headline: Phylogenetic multilevel meta-analysis

Abstract

1. Meta-analyses in ecology and evolution require special attention due to certain study characteristics in these fields. First, the primary articles in these fields usually report results that are observed from studies conducted with different species, and the phylogeny among the species violates the independence assumption. Second, articles frequently allow the computation of multiple effect sizes which cannot be accounted for by conventional meta-analytic models. While both issues can be dealt with by utilizing a multilevel model that accounts for phylogeny, the performance of such a model has not been examined extensively. In this article, we investigate the performance of this model in comparison with some simpler models.

2. We conducted an extensive simulation study where data with different hierarchical structures (in terms of study and species levels) were generated and then various models were fitted to examine their performance. The models we used include the conventional random-effects and multilevel random-effects models along with more complex multilevel models that account for species-level variance with different variance components. Furthermore, we present an illustrative application of these models based on the data from a meta-analysis on size-assortative mating and comment on the results in light of the findings from the simulation study.

3. Our simulation results show that, when the phylogenetic relationships among the species are at least moderately strong, only the most complex model that decomposes the species-level variance into non-phylogenetic and phylogenetic components provides approximately unbiased estimates of the overall mean and variance components and yields confidence intervals with an approximately nominal coverage rate. Contrarily, removing the phylogenetic or non-phylogenetic component leads to biased variance component estimates and an increased risk for incorrect inferences about the overall mean. These findings are supported by the results derived from the illustrative application.

4. Based on our results, we suggest that meta-analyses in ecology and evolution should use the model that accounts for both the non-phylogenetic and phylogenetic species-level variance in addition to the multilevel structure of the data. Any attempts to simplify this model, such as using only the phylogenetic variance component, may lead to erroneous inferences from the data.

Keywords: comparative analysis, mixed-effects models, model efficiency, multilevel models, phylogenetic meta-analysis, random-effects variance estimation.

1 Introduction

Meta-analysis encompasses an array of methods for synthesizing information from studies examining some phenomenon of interest and evaluating the consistency of their results (Glass, 1976; Hedges and Olkin, 1985; Cooper et al., 2009; Senior et al., 2016). Although these methods have been mostly developed in the medical and social sciences (Egger et al., 2001; Sutton and Higgins, 2008; Cooper et al., 2009), ecologists and evolutionary biologists have successfully adopted these techniques for conducting research syntheses in their respective fields (Jessica Gurevitch et al., 2001; Koricheva et al., 2013; J. Gurevitch et al., 2018). However, meta-analyses in ecology and evolution typically have several features that require special attention so that trustworthy evidence can be obtained.

To start, meta-analyses in these fields often incorporate data from multiple species which share an evolutionary history, described by a phylogeny (Arnqvist and Wooster, 1995; J. Gurevitch and Hedges, 1999; Chamberlain et al., 2012). As a result, the samples (and the effect sizes obtained from these samples) are not independent which violates the independence assumption underlying conventional meta-analytic models. For example, the standard fixed- and random-effects models (Hedges and Olkin, 1985; Hedges and Vevea, 1998), often used for ecological meta-analyses (Nakagawa and Santos, 2012), assume independence among the effect sizes and therefore do not account for phylogeny (Chamberlain et al., 2012; Noble et al., 2017). This issue was first addressed by Adams (2008) and Lajeunesse (2009) who incorporated phylogenies into the fixed- and random-effects models, respectively.

Chamberlain et al. (2012) empirically investigated how the inclusion of phylogeny affects the estimate of the overall mean based on data from 30 meta-analyses in ecology and evolution. While the estimate of the overall mean did not change considerably in most cases (especially when using a random-effects model), a substantial portion of the meta-analyses, which reported significant results before, produced non-significant results when the phylogeny was incorporated into the model. Therefore, including phylogeny might be an important factor to reduce Type I error rates and to obtain an accurate reflection of the uncertainty of meta-analytic estimates.

Although Chamberlain et al. (2012) is the most extensive study to date examining the effects of phylogeny in meta-analysis, their work was based on available meta-analyses. To investigate the issue of phylogeny more broadly, we require a simulation study to explore a wider parameter space and under controlled conditions. Moreover, Chamberlain et al. (2012) did not address the fact that ecological and evolutionary studies usually report multiple effect sizes per study, which leads to dependence among the effect sizes belonging to the same study (Nakagawa and Santos, 2012; Noble et al., 2017). Although past and current meta-analyses have sometimes avoided this issue by selecting a single effect size from each study or by collapsing multiple effect sizes into one, these procedures can lead to a severe loss of information (Nakagawa and Santos, 2012; Nakagawa et al., 2021).

As an alternative, Hadfield and Nakagawa (2010) proposed a mixed-effects model that accounts for the multilevel structure via a study-level random effect (i.e., multiple effect sizes per study are nested within this random effect). In the same model, they include two additional random effects to estimate the non-phylogenetic and the phylogenetic species-level variance. This way, among-species variance is decomposed into two components, the one resulting from species similarities due to evolutionary history and the other from species similarities due to shared ecology and other factors (Lynch, 1991). Although the model by Hadfield and Nakagawa (2010) addresses two major statistical issues in ecological and evolutionary meta-analyses, the complexity of the model poses certain challenges.

Partitioning the species variance into its two components is a challenging endeavor, because both components are modeled using random effects at the species level, with the only difference being that the phylogenetic component assumes that the random effects are correlated according to a phylogenetic correlation matrix – which is derived from a phylogenetic tree constructed based on the similarities and differences of species in terms of their (usually) genetic (but sometimes also physical) characteristics (Felsenstein, 2004). This raises concerns about the identifiability of the variance components and potential bias in their estimates, issues that have also been raised outside the meta-analytic context when analyzing the data of primary studies including multiple species (Paradis, 2012).

Moreover, the complexity of the model poses a threat to the convergence of optimization algorithms (Bates et al., 2015). Accordingly, Nakagawa and Santos (2012) suggested that model fitting may only be feasible with larger datasets, which would limit the applicability of the model in practice. To avoid these problems, some ecological and evolutionary meta-analyses have been carried out using a simplified model without the non-phylogenetic random effect and that therefore accounts for species variance only via the phylogenetic component (e.g., Garamszegi et al., 2012; Moore et al., 2016). However, the consequences of doing so, and the performance of the more complex model, has yet to be evaluated in a simulation study.

We therefore investigated the performance of models for conducting a phylogenetic multilevel meta-analysis in a comprehensive simulation study. We simulate studies that report multiple effect sizes and use several models that vary in their complexity, starting from a simple model (including only a random effect at the effect sizes level) to the most complex model which incorporates a study-level and two among-species random effects. Further, we generate specific conditions to examine the performance of the most complex model when phylogenetic relationships are weak and the consequences of removing the non-phylogenetic component. Finally, we present an illustrative application of these models based on the data from a meta-analysis on size-assortative mating and comment on the results in light of the findings from the simulation study.

2 Materials and Methods

2.1 Meta-Analytic Models

To conduct a meta-analysis, the phenomenon of interest (e.g., the size of a treatment effect or the strength of the association between two variables) needs to be quantified in terms of an effect size estimate for each study to be included in the analysis. We use the term ‘study’ broadly here (and essentially in the sense of ‘paper’ or ‘publication’), as a single study may contribute multiple estimates (i.e., multiple effect sizes, for instance, for multiple species, subgroups, treatments), but for the moment we assume that each study contributes a single estimate to the meta-analysis. Depending on the purpose of a meta-analysis and the information reported in the individual studies, one might use raw or standardized mean differences, response ratios, odds/risk ratios, or correlation coefficients to quantify the relevant results (see Borenstein et al., 2011, for a review). In addition, we need to compute the sampling variances of the estimates, that is, the variability in each estimate that would be expected under repeated sampling of new study units under identical circumstances (Nakagawa and Cuthill, 2007; Cooper et al., 2009; Borenstein et al., 2011).

Regardless of the specific measure used in a meta-analysis, let y_i denote the effect size estimate for the ith study (with $i = 1, \ldots, N_{studies}$) and v_i the corresponding sampling variance (note that the terms ‘study’ and ‘effect size’ are interchangeable when each study reports a single effect size). The most basic model that can be considered for synthesizing the estimates is the fixed-effects model, which is given by

$$y_i = \mu + e_i, \tag{1}$$

$$\mathbf{e} \sim N(\mathbf{0}, \mathbf{V}), \tag{2}$$

where μ is the overall mean, e_i is the sampling error for the ith study, $\mathbf{e}$ is a $1 \times N_{studies}$

column vector with the e_i values (which are assumed to be normally distributed with mean 0 and variance v_i), $\mathbf{0}$ is a column vector of zeros, and $\mathbf{V}$ is an $N_{studies} \times N_{studies}$ matrix with the v_i values along the diagonal.

The fixed-effects model assumes that the included studies share a single common true effect. This assumption, however, is rarely met in multi-population and multi-species meta-analyses of ecology and evolution studies (Senior et al., 2016). The random-effects model addresses this potential 'heterogeneity' among the true effects by adding a random effect corresponding to each estimate and is given by

$$y_i = \mu + u_i + e_i \tag{3}$$

$$\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_u), \tag{4}$$

where u_i is the random effect corresponding to the ith estimate, $\mathbf{u}$ is a $1 \times N_{studies}$ column vector with the u_i values (which are assumed to be normally distributed with mean 0 and variance σ_u^2), and $\mathbf{I}_u$ is an $N_{studies} \times N_{studies}$ identity matrix.

Although the models above are suitable for conducting a meta-analysis in many circumstances, they do not account for the multilevel structure that arises when at least some studies provide multiple effect size estimates (e.g., when the same experiment was conducted under varying circumstances within the same study) and they do not account for phylogenetic dependence (when studies are conducted with multiple species that differ in similarity due to differences in their shared evolutionary history).

To address the first issue, we can use a multilevel meta-analytic model (Konstantopoulos, 2011; Nakagawa and Santos, 2012) which includes a random effect at the effect size level (as in model 3 – for brevity, we use the equation numbers to refer to the various models throughout this article), but which now captures variability in the true effects within studies, and a random effect at the study level, which captures between-study variability. Let y_{ij}

denote the jth effect in the ith study (with $j = 1, \ldots, N_i$, where N_i is the number of effect sizes reported in the ith study), v_{ij} the corresponding sampling variance, and let $N_{total} = \sum_{i=1}^{N_{studies}} N_i$ denote the total number of effects. The model is then given by

$$y_{ij} = \mu + u_{ij} + s_i + e_{ij} \tag{5}$$

$$\mathbf{s} \sim N(\mathbf{0}, \sigma_s^2 \mathbf{I}_s), \tag{6}$$

where u_{ij} is a random effect corresponding to the jth effect size in the ith study, s_i is a random effect at the study level, $\mathbf{u}$ is now a $1 \times N_{total}$ column vector with the u_{ij} values, $\mathbf{s}$ is a $1 \times N_{studies}$ column vector with the s_i values (which are assumed to be normally distributed with mean 0 and variance σ_s^2), and $\mathbf{I}_u$ and $\mathbf{I}_s$ are $N_{total} \times N_{total}$ and $N_{studies} \times N_{studies}$ identity matrices, respectively. Finally, $\mathbf{e}$ is now a $1 \times N_{total}$ column vector with the e_{ij} values and $\mathbf{V}$ is the corresponding (diagonal) variance-covariance matrix with dimensions $N_{total} \times N_{total}$, and the remaining terms are defined as described earlier.

When the effect size estimates are computed based on a set of $N_{species}$ different species, we will need an additional index. Let y_{ijk} denote the jth effect in the ith study as before, but now let $k = 1, \ldots, N_{species}$ be the index that indicates for which species a particular effect size estimate was computed. Model 5 can then be extended to account for species-level variability as follows:

$$y_{ijk} = \mu + u_{ij} + s_i + n_k + e_{ij}, \tag{7}$$

$$\mathbf{n} \sim N(\mathbf{0}, \sigma_n^2 \mathbf{I}_n), \tag{8}$$

where n_k is a species-specific random effect, $\mathbf{n}$ is a $1 \times N_{species}$ column vector with the n_k values (which are assumed to be normally distributed with mean 0 and between-species variance σ_n^2), and $\mathbf{I}_n$ has dimensions $N_{species} \times N_{species}$, with the remaining terms as defined earlier. Note that n_k is a crossed random effect (e.g., Fernández-Castilla et al., 2019) and not nested within studies and we therefore do not put subscript k on u_{ij}, s_i, or e_{ij}.

Model 7, however, does not account for phylogeny. For this, we further extend the model by including an additional species-level random effect (Hadfield and Nakagawa, 2010), but instead of assuming independence for different species (as for the n_k values), we allow the values of the random effect to be correlated according to a phylogenetic correlation matrix, which in turn is derived from a phylogenetic tree based on some model of evolution (e.g., Brownian motion) prior to the analysis (e.g., Lajeunesse, 2009; Felsenstein, 1985; Felsenstein, 2004; Freckleton et al., 2002). The model is given by

$$y_{ijk} = \mu + u_{ij} + s_i + n_k + p_k + e_{ij}, \tag{9}$$

$$\mathbf{p} \sim N(\mathbf{0}, \sigma_p^2 \mathbf{A}), \tag{10}$$

where p_k denotes the phylogenetic random effect for the kth species, $\mathbf{p}$ is a $1 \times N_{species}$ column vector with the p_k values (which are assumed to follow a multivariate normal distribution with mean 0 and variance-covariance matrix $\sigma_p^2\mathbf{A}$, where σ_p^2 denotes between-species variance due to the phylogeny, and $\mathbf{A}$ is the $N_{species} \times N_{species}$ phylogenetic correlation matrix), with the remaining terms as defined earlier. Hence, the model includes a non-phylogenetic species-level random effect (i.e., the n_k values) to account for heterogeneity in the effects sizes due to differences between species unrelated to phylogeny (e.g., the influence of differences in the environments they live in) and a phylogenetic random effect (i.e., the p_k values) that captures dependencies in the effect sizes according to the similarities between species due to phylogenetic relatedness.

Since model 9 includes the species random effect twice (once assumed to be independent and once assumed to be correlated according to the values in $\mathbf{A}$), concerns about identifiability and potential bias in the estimates of the variance components may be raised. In fact, when phylogenetic relationships are weak (i.e., when the off-diagonal values in $\mathbf{A}$ are close to

0 and hence the phylogenetic tree resembles a star phylogeny), then $\mathbf{A}$ starts to approximate $\mathbf{I}_n$ and hence σ_p^2 and σ_n^2 are confounded and are not uniquely identifiable. This concern, or the complexity of model 9 in general, has led some researchers to adopt a simplified model in their meta-analyses where the non-phylogenetic variance component is removed. This leads to the model

$$y_{ijk} = \mu + u_{ij} + s_i + p_k + e_{ij}, \tag{11}$$

with all terms as explained before. Whether this simplified version is an adequate substitute for model 9 is currently unknown.

The models described above can be fitted with the `metafor` package (Viechtbauer, 2010) for `R` (R Core Team, 2021). Maximum likelihood (ML) or restricted maximum likelihood (REML) estimation can be used for model fitting (the latter usually being the preferred choice; see Patterson and Thompson, 1971), providing estimates of the variance components included in a particular model, the estimate of μ (i.e., $\hat{\mu}$), and its standard error (i.e., SE[$\hat{\mu}$]). Likelihood ratio tests and profile likelihood confidence intervals provide inferences for the variance components. An approximate 95% Wald-type confidence interval for μ can be obtained with $\hat{\mu} \pm t_{.975,df}\text{SE}[\hat{\mu}]$, where $t_{.975,df}$ denotes the 97.5th percentile of a t-distribution with *df* degrees of freedom. Based on Nakagawa et al. (2021), we set $df = N_{studies} - 1$, which we expected would bring the coverage rate of the confidence interval closer to its nominal 95% level (when compared to a confidence interval based on a standard normal distribution).

Although fitting the models and deriving inferences from them is feasible, the consequences of using the various models have not been examined systematically. We therefore conducted an extensive simulation study to investigate the performance of the various model under varying circumstances.

Table 1: Overview of the conditions examined in the simulation study. The first two columns show the number of studies and species, respectively. The next four columns indicate the true values of the variance components. The α column represent the power parameter. All values were crossed within a particular row of the table. The last two columns respectively indicate the number of conditions generated in each row and the model that corresponds to the true data generating mechanism for the conditions in a particular row.

$N_{studies}$	$N_{species}$	σ_u^2	σ_s^2	σ_n^2	σ_p^2	α	Conditions	True model
20	40	0, 0.05, 0.30	0	0	0	1	3	Model 3
20	40	0.05, 0.30	0.05, 0.30	0	0	1	4	Model 5
20	40	0.05, 0.30	0.05, 0.30	0.05, 0.30	0	0.5, 1, 2	24	Model 7
20	40	0.05, 0.30	0.05, 0.30	0.05, 0.30	0.05, 0.30	0.5, 1, 2	48	Model 9
50	100	0, 0.05, 0.30	0	0	0	1	3	Model 3
50	100	0.05, 0.30	0.05, 0.30	0	0	1	4	Model 5
50	100	0.05, 0.30	0.05, 0.30	0.05, 0.30	0	0.5, 1, 2	24	Model 7
50	100	0.05, 0.30	0.05, 0.30	0.05, 0.30	0.05, 0.30	0.5, 1, 2	48	Model 9

2.2 Simulation Setup

In our setup, the primary studies could provide one or multiple effect size estimates for one or multiple species. We set $(N_{studies}, N_{species})$ either to $(20, 40)$ or $(50, 100)$ to examine the difference between a smaller versus larger meta-analysis. Furthermore, we set σ_u^2, σ_s^2, σ_n^2, and σ_p^2 to either 0, 0.05, or 0.3 (plus an additional parameter α to be described below to either 0.5, 1, or 2) to define a particular condition within the simulation study. Table 1 provides an overview of the 158 conditions that were studied in this manner. Note that, instead of a full factorization of all parameters, we introduced the variance components successively (in the order of σ_u^2, σ_s^2, σ_n^2, and σ_p^2) using the non-zero values (i.e., 0.05 and 0.3) to keep the number of conditions manageable and to generate scenarios where one of the models described in equations 3, 5, 7, and 9 corresponds to the true data generating mechanism (see Table 1). Within a particular condition, the following steps were repeated 1000 times.

First, the number of effect sizes provided by the studies (i.e., the N_i values) were simulated from a right-skewed distribution, as typically observed in practice. For this, we generated

$N_{studies}$ random values from a Beta$(1.5, 3)$ distribution, which were then multiplied by 39, rounded to the closest integer, and increased by 1. Therefore, the number of estimates per study could vary between 1 and 40 (with a mean, median, and mode of approximately 14, 13, and 9, respectively).

In the next step, we simulated the species indices (i.e., the k values) by generating N_{total} random values from a Beta$(2, 2)$ distribution, which were multiplied by $N_{species} - 1$, rounded to the closest integer, and then increased by 1. Accordingly, the number of times that the various species were studied followed a symmetric unimodal distribution (with mean equal to $(N_{species} + 1)/2$). In order to guarantee that all species appear at least once in each meta-analysis, a randomly chosen $N_{species}$ random numbers generated this way were replaced with the integers from 1 to $N_{species}$.

Next, we generated a phylogenetic tree for the species using the `rtree()` function from the `R` package `ape` (Paradis and Schliep, 2019), which uses a recursive random splitting algorithm to simulate a phylogeny (Paradis, 2012). The branch lengths were then computed using the `compute.brlen()` function based on the method by Grafen (1989), using the power parameter α to adjust the 'height' of branch lengths at the tips of the phylogenetic tree, leading to phylogenetic relationships that are generally stronger when branches are shorter at the tips or weaker when branches are longer at the tips. Fig. 1 shows an example of such a simulated tree for 40 species modified by different α values. Finally, the correlation matrix that represents the phylogenetic relationships (matrix $\mathbf{A}$ in equation 10) was calculated from the tree by using the `vcv()` function based on a Brownian model of evolution (i.e., $\mathbf{A}_{k,k'} = 1 - b_{k,k'}$, where $b_{k,k'}$ is the branch length for a pair of species to their most recent common ancestor). Hence, as α decreases, the off-diagonal values in $\mathbf{A}$ converge to 0, whereas as α increases, the off-diagonal values in $\mathbf{A}$ increase on average.

We then generated the values for the four random effects, corresponding to the variance components σ^2_u, σ^2_s, σ^2_n, and σ^2_p, either as independent draws from normal distributions for

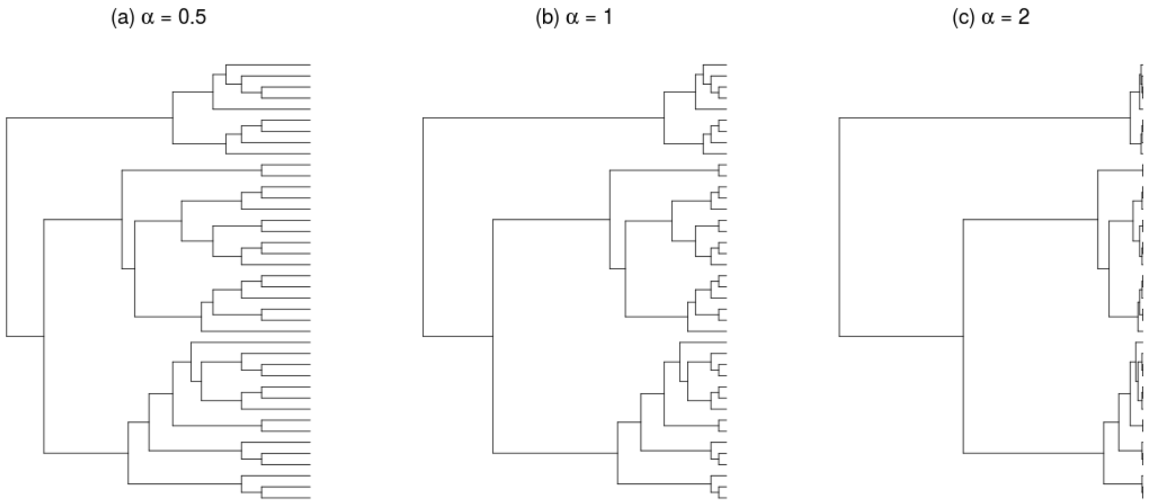

Figure 1: An example of a simulated phylogenetic tree for 40 species modified with different values of the power parameter α (i.e., 0.5, 1, and 2).

the first three components or from a multivariate normal distribution for the last one. In conditions where a variance component is equal to 0, the corresponding random effect values are then just a series of 0s of the appropriate length. To complete the data generating step, the sampling variances (i.e., the v_{ij} values) were simulated from a right-skewed $\text{Beta}(2, 20)$ distribution (and hence had a value of .091 on average) which were then used to generate the N_{total} sampling errors from a normal distribution with mean 0 and variance v_{ij}. We then summed the random effects and sampling errors as shown in equation 9, setting $\mu = 0$ without loss of generality (as scalar changes to μ do not affect any other parts of the models).

After generating the data, we fitted the four models shown in equations 3, 5, 7, and 9, using REML estimation as implemented in the `rma.mv()` function from the `metafor` package. For model 3, we simply treated each estimate as a separate study (one can also think of this as model 5 without the addition of the study-level random effect). For each model, we then saved the estimate of μ, the variance component estimates, the bounds of the 95% Wald-type confidence interval for μ, and the model fitting time to assess how demanding the computations are when fitting these models. In case any one of the four models did not converge within a particular iteration (with the default settings of the `rma.mv()` function), the iteration was discarded and a new iteration was run to guarantee that a 1000 successful model fits were available for all four models (in all conditions, >99% of the analyses converged on solutions).

After the 1000 iterations, we computed the mean of the $\hat{\mu}$ values for each model, the mean of the variance component estimates, the proportion of iterations where 0 was included in the confidence interval (i.e., the empirical coverage rate for μ), the mean confidence interval width, the mean absolute bias in the estimates of μ and the variance components, the convergence rate, and the mean model fitting time. The simulation was run on a workstation with two AMD EPYC 7551 32-Core CPUs utilizing 60 cores in parallel. Completion time for the simulation was approximately 35 hours (roughly 2100 core hours).

We generated two other sets of conditions to investigate specific questions. First, we examined conditions where the phylogenetic relationships could also be weaker than in the main scenarios to test the performance of model 9 under such conditions. These conditions were generated by setting α to $(0.1, 0.2, 0.3, 0.4, 0.5, 1, 2)$ when $(N_{studies}, N_{species}) = (50, 100)$, the estimate- and study-level variance components were both large (0.3), and the levels of the remaining variance components were factorized with values of 0.05 and 0.3 (for a total of 28 different conditions). Second, we compared the performance of model 9 and the simplified model 11 (that leaves out the non-phylogenetic species-level random effect). For this, we set $(N_{studies}, N_{species}) = (50, 100)$, $\sigma^2_u = 0.05$, $\sigma^2_s = 0.05$, and $\alpha = 1$, and then generated different conditions by factorizing different values of only σ^2_n and σ^2_p, where the former was set to values from 0 to 0.3 with increments of 0.05, whereas the latter was set to either 0, 0.05, or 0.3 (for a total of 21 different conditions). The R code to reproduce the simulation and its results are available at the Open Science Framework (https://osf.io/ms8eq/).

3 Results

3.1 Simulation Results

Fig. 2a displays boxplots of the mean $\hat{\mu}$ values (over the 1000 iterations) for each of the four models across the 158 conditions, separated by which model was the true data generating mechanism. Generally, the means were clustered tightly around 0, indicating little to no bias in $\hat{\mu}$, although in a small set of conditions there was some slight positive bias in the estimates of the overall mean. These conditions were characterized by non-zero values for all four variance components (i.e., when model 9 was the true model), $(N_{studies}, N_{species}) = (20, 40)$, a weak phylogenetic relationship ($\alpha = 0.5$), and a large phylogenetic variance ($\sigma_p^2 = 0.3$).

In contrast to the results for the overall mean, the coverage rates of the 95% confidence interval for μ differed markedly across models (Fig. 2b). For conditions where model 3 was the true data generating mechanism, all models achieved coverage rates close to or slightly above the nominal 95% confidence level regardless of the condition. As the other variance components were introduced into the data, however, the coverage rates of models that did not account for these additional sources of variability started to decrease, at times severely so. Only model 9 was able to achieve rates close to the nominal level across the majority of conditions, although the rates also fell somewhat below the nominal level for certain conditions when all variance components were larger than zero.

Given that estimates of μ were relatively unbiased for all models, the closer to nominal coverage rates of model 9 would be expected to be mainly a consequence of wider confidence intervals (that consequently have a better chance of capturing the true value of μ). Fig. 2c confirms this, showing the mean confidence interval widths for the various models across the various conditions. However, what is particularly noteworthy is that the use of model 9 under conditions where actually a simpler model is the true data generating mechanism only leads to a relatively minor increase in the mean interval width.

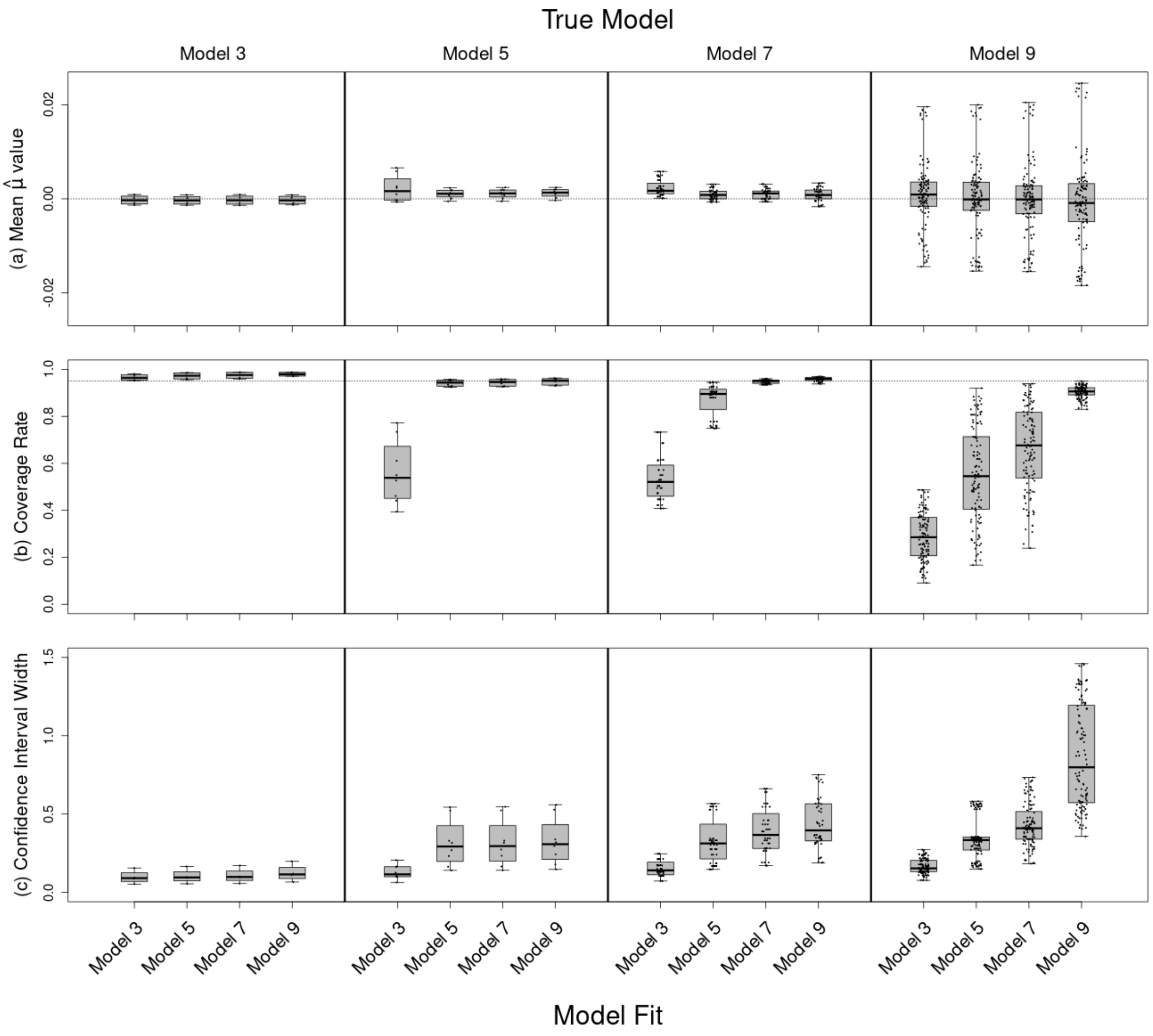

Figure 2: Boxplots (representing the five-number summaries) based on the (a) mean $\hat{\mu}$ values (over the 1000 iterations), (b) coverage rates of the 95% confidence interval for μ, and (c) mean confidence interval widths for each of the four models across the 158 conditions, separated by which model was the true data generating mechanism.

Fig. 3 displays the bias in the variance component estimates of model 9 under the 28 different conditions generated by varying α, σ_n^2, and σ_p^2 (while holding σ_u^2 and σ_s^2 constant at 0.3). The results show no bias in the estimates of σ_u^2 and σ_s^2. Furthermore, the model is able to estimate σ_n^2 and σ_p^2 with little to no bias, except when the strength of the phylogenetic relationships decreased. As expected, under such conditions, the model struggles to provide

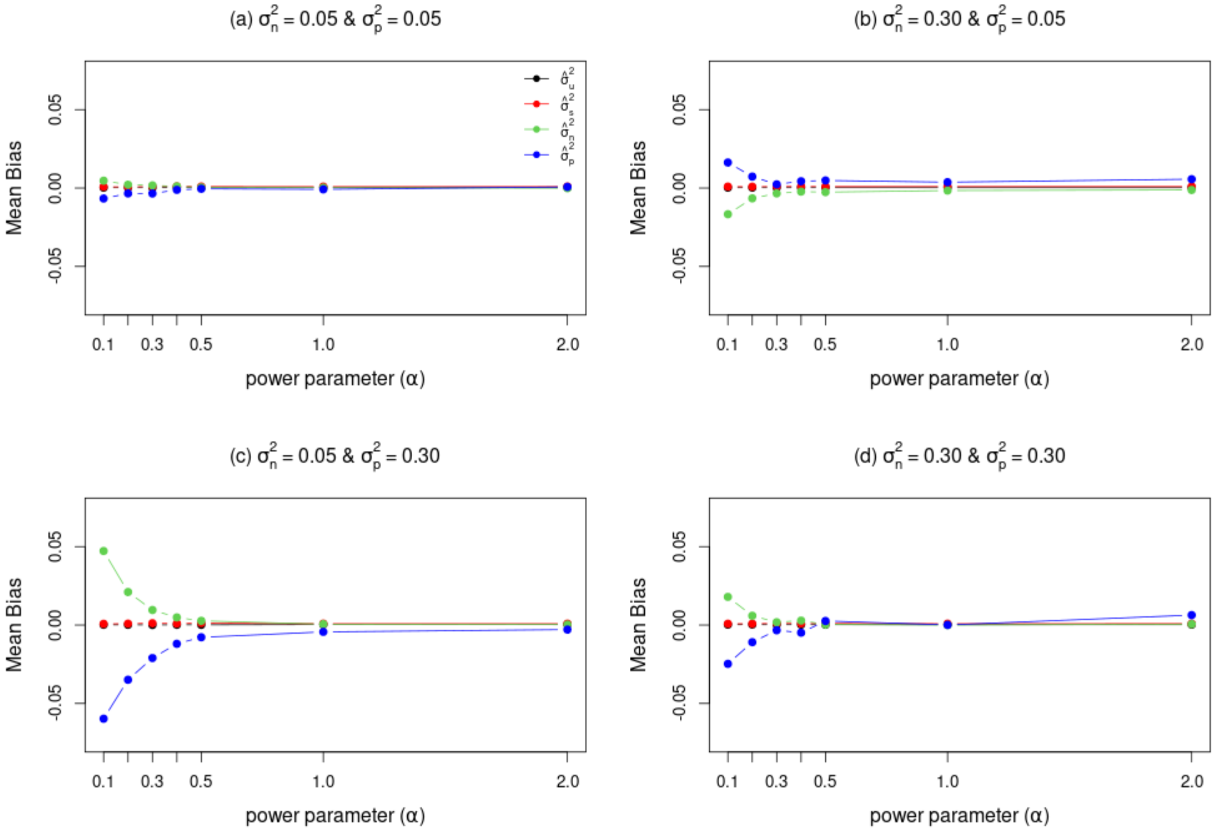

Figure 3: Mean bias of the variance component estimates of model 9 under different combinations of the power parameter (α) and the non-phylogenetic and phylogenetic variance components (σ_n^2 and σ_p^2, respectively). The variance components in model 9, σ_u^2, σ_s^2, σ_n^2, and σ_p^2 are presented as black, red, green, and blue lines.

unbiased estimates of the non-phylogenetic and phylogenetic species-level variance components. Regardless, model 9 still provided overall estimates with mean absolute bias lower than 0.024 across all 28 conditions, although the coverage rate of the CI for μ again tended to fall somewhat below the nominal 95% level (with a mean coverage rate of 92% over the 28 conditions).

Fig. 4a shows the coverage rates of the confidence interval for μ for models 9 and 11 as the size of the non-phylogenetic species-level variance component (i.e., σ_n^2) was increased. While model 9 provided rates close to or somewhat below the nominal level, the rates for model 11 were often equal to 100% and hence the confidence interval tended to be too wide

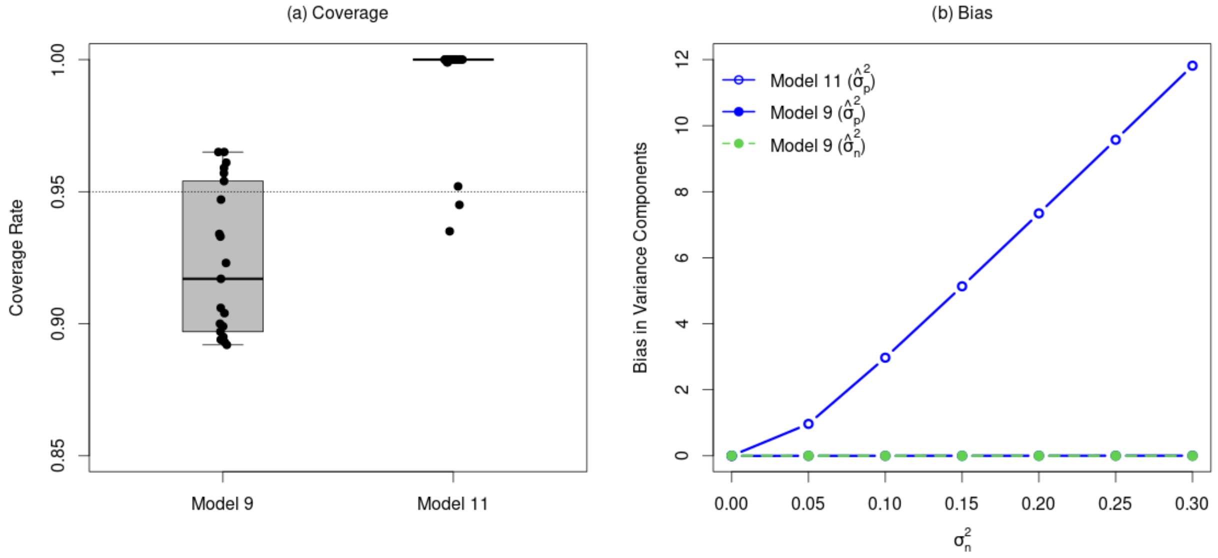

Figure 4: Comparison of models 9 and 11 as the size of the non-phylogenetic species-level variance component (i.e., σ_n^2) was systematically increased. (a) Coverage rates of the 95% confidence intervals for μ, (b) bias in the non-phylogenetic and phylogenetic variance components.

(except for the three conditions where $\sigma_n^2 = 0$ and hence where model 11 was the true model). Furthermore, Fig. 4b demonstrates that the bias in the phylogenetic variance component of model 11 inflated rapidly as the value of σ_n^2 increased (the value of σ_p^2 had no noteworthy influence on the bias and hence we averaged these results over the three possible values of σ_p^2). In contrast, model 9 estimated these two variance components essentially without bias under these scenarios.

Model fitting times differed between the various models (Table 2), with model 9 requiring the most amount of time on average, regardless of the true data generating mechanism. The most challenging conditions for the more complex models were those scenarios where model 3 corresponded to the true data generating mechanism. In this case, a single fit of model 9 took around 33 seconds on average when $(N_{studies}, N_{species}) = (50, 100)$. In these conditions, convergence rates were also the lowest, although even model 9 then converged in more than 99% of the iterations.

Table 2: Average model fitting times in seconds and convergence rates (in parentheses) of all models under the different data generating mechanisms.

(a) $(N_{studies}, N_{Species}) = (20, 40)$

	True Model			
Model Fit	Model 3	Model 5	Model 7	Model 9
Model 3	0.841	0.852	0.830	0.858
	(100.00%)	(100.00%)	(100.00%)	(100.00%)
Model 5	3.052	1.433	1.418	1.475
	(100.00%)	(100.00%)	(100.00%)	(100.00%)
Model 7	2.753	2.227	1.015	1.045
	(99.75%)	(100.00%)	(100.00%)	(100.00%)
Model 9	3.805	3.671	2.781	1.825
	(99.26%)	(99.68%)	(99.99%)	(100.00%)

(b) $(N_{studies}, N_{Species}) = (50, 100)$

	True Model			
Model Fit	Model 3	Model 5	Model 7	Model 9
Model 3	1.625	1.643	1.687	1.551
	(100.00%)	(100.00%)	(100.00%)	(100.00%)
Model 5	4.446	2.506	2.573	2.379
	(100.00%)	(100.00%)	(100.00%)	(100.00%)
Model 7	24.611	19.649	9.862	9.528
	(100.00%)	(100.00%)	(100.00%)	(100.00%)
Model 9	32.897	31.880	25.287	14.405
	(99.31%)	(99.53%)	(100.00%)	(100.00%)

3.2 Illustrative Example

We use the data from the meta-analysis by Rios Moura et al. (2021) on size-assortative mating (SAM) to illustrate an application of the models. Each study included in the meta-analysis provided one or multiple correlation coefficients describing the similarity in some measure of body size in mating couples. For the analysis, the correlation coefficients were transformed with Fisher's r-to-z transformation (i.e., the inverse hyperbolic tangent transformation). We focus here on the estimate of the overall mean (transformed) correlation coefficient, leaving aside the issue of differences between studies where correlations were computed with or without pooling of data across different timepoints or areas (i.e., temporal/spatial pooling). Also, using the method by Grafen (1989), we turned the phylogenetic tree used by Rios Moura et al. (2021) into an ultrametric tree before fitting models 9 and 11, to bring these analyses more in line with how our simulation study was conducted. The dataset includes 1828 effect size estimates (i.e., transformed correlations) collected from 457 studies and 341 species.

Table 3 presents the results obtained from each model. Interestingly, the estimate of the overall mean tended to be somewhat larger in the more complex models, although differences

Table 3: Results derived from fitting the various models to the example dataset. The first five columns show the estimated overall mean, its standard error, the 95% confidence interval, the test statistic, and the p-value for testing H$_0$: $\mu = 0$, respectively. The next four columns show the estimates of the variance components in the respective models. The last column shows the Akaike Information Criteria (AIC) values.

	$\hat{\mu}$	$SE[\hat{\mu}]$	95% CI	Z	p	$\hat{\sigma}_u^2$	$\hat{\sigma}_s^2$	$\hat{\sigma}_n^2$	$\hat{\sigma}_p^2$	AIC
Model 3	0.24	0.007	0.23, 0.25	34.15	<0.0001	0.0641	–	–	–	1082.8
Model 5	0.30	0.015	0.27, 0.33	20.42	<0.0001	0.0149	0.0806	–	–	429.0
Model 7	0.34	0.019	0.30, 0.38	17.37	<0.0001	0.0143	0.0195	0.0815	–	386.3
Model 9	0.37	0.130	0.11, 0.62	2.83	0.0046	0.0145	0.0192	0.0555	0.0512	344.7
Model 11	0.36	0.172	0.02, 0.70	2.07	0.0382	0.0149	0.0557	–	0.0913	367.2

between models 7, 9, and 11 were relatively small. More importantly, we see a substantial increase in the standard error of the estimated overall mean for the more complex models. As a result, the confidence intervals become wider, the values of the test statistics smaller, while the respective p-values increase. Although each model suggests that the overall mean significantly differs from 0 (at the conventional 0.05 level of significance), the p-value for model 11 was approaching the rejection threshold.

The estimates of the variance components also show some interesting patterns. While the simple random-effects model 3 cannot distinguish between different sources of variability and attributes all of the heterogeneity to differences between the individual effect size estimates, model 5 suggests that the variance in the effects is more related to differences between studies than particular estimates within studies. However, once species-level variability is considered in model 7, it becomes apparent that this is actually the dominant source of heterogeneity. Moreover, model 9 shows that this variability is approximately equally attributable to non-phylogenetic and phylogenetic species-level differences. In contrast, when ignoring the non-phylogenetic variance component in the simplified model 11, part of the variance from that component is forced back into the study-level variance component. Furthermore, $\hat{\sigma}_p^2$ in the simplified model is substantially inflated compared to model 9 which may be an example of the inflation in this component when σ_n^2 is excluded (see Fig. 4b). Based on these findings and the Akaike Information Criteria (AIC) values of the various models, we would strongly favor model 9 in this comparison, illustrating that both non-phylogenetic and phylogenetic variance components should be considered in the analysis.

4 Discussion

Meta-analyses in the fields of ecology and evolution typically need to address the fact that multiple effect size estimates can be extracted from at least some of the studies and that the estimates are based on various species that are related to each other due to their shared evolutionary history. In this paper, we investigated the performance of the phylogenetic multilevel meta-analytic model by Hadfield and Nakagawa (2010) and Nakagawa and Santos (2012) that captures these intricacies along with some simpler models. Despite the concerns raised in the introduction, the model can successfully estimate the overall mean and its uncertainty. It also provides approximately unbiased estimates of all variance components, including the non-phylogenetic and phylogenetic species-level variances, as long as there are at least moderately strong phylogenetic relationships among the species. In addition, despite its complexity, the model does not appear to suffer from convergence problems and model fitting does not require excessive computational times.

4.1 Estimating the Overall Mean and its Uncertainty

Not only the phylogenetic multilevel meta-analytic model, but also the simpler models that leave out certain variance components provide essentially unbiased estimates of the overall mean, regardless of the nature of the true model that underlies the data (Fig. 2a). However, the uncertainty in the overall mean will only be estimated accurately when the fitted model includes the variance components that contribute to the heterogeneity and the dependencies among the underlying true effects. Fitting underspecified models typically led to severe undercoverage of the confidence interval for the overall mean and hence anticonservative inferences. In fact, subtracting the coverage rates shown in Fig. 2b from 1 yields the Type I error rates for the test of the overall mean, which could go as high as 91% when using a simple random-effects model that ignores the multilevel structure and the species-level variance components.

These findings are in line with those by Chamberlain et al. (2012), who demonstrated, based on 30 published meta-analyses, that the inclusion of phylogeny into a random-effects model usually only led to minor changes in the pooled effect size, but had a more substantial impact on the statistical significance of the finding (turning significant findings into non-significant ones in the majority of cases where changes occurred).

Our findings can also be used to alleviate concerns with using the phylogenetic multilevel meta-analytic model when it is actually an overspecified model (i.e., when the actual data generating mechanism is simpler). In those cases, the mean confidence interval width of the model was just barely wider than that of the simpler models, indicating little to no loss in efficiency by fitting an overly complex model (Fig. 2c). The superfluous variance components then converge towards 0 (or close to it), which appears to be slightly more challenging for the optimization algorithm, leading to longer model fitting times and occasional convergence problems, but not to any worrisome degree (Table 2). Moreover, in practice, for any particular dataset, convergence problems can typically be resolved by selecting a different optimizer or making changes to the settings for the optimization routine, so the convergence rates as given only apply to the default settings.

At the same time, we should point out that the coverage rate of the model did fall slightly below the nominal 95% level in the majority of conditions when all variance components were in fact non-zero (see Fig. 2b, rightmost panel). A similar issue, but for a simpler model with only between- and within-study variance components (i.e., model 5 in our simulation) was also recently pointed out by Song et al. (2020). Improved methods based on a t-distribution with various approximations for the degrees of freedom have been proposed and studied extensively in the context of the standard random-effects model (e.g., Sanchez-Meca and Marin-Martinez, 2008) and mixed-effects models in general (e.g., Luke, 2017). Following Nakagawa et al. (2021), we actually based the confidence interval on a t-distribution with $N_{studies} - 1$ as the degrees of freedom (as an improvement to using a confidence interval based on a standard normal distribution), although this was apparently not conservative enough, presumably due to the additional dependency among the effect sizes introduced by the phylogeny. Further work will be needed to find an even better approximation to the degrees of freedom in the present context.

4.2 Including and Testing the Phylogenetic Effect

Phylogenies play a central role in the context of phylogenetic comparative studies (Freckleton et al., 2002; Blomberg et al., 2003; Ives et al., 2007). An important step in such studies is testing the significance of the 'phylogenetic signal' in some trait of interest. This test is often performed through a statistic such as λ (Pagel, 1999) or K (Blomberg et al., 2003). Although model 9 does not parameterize the phylogenetic effect in this manner, one can derive information from its output that shows its relationship to the λ statistic. In particular, Pagel's λ is a multiplicative factor that is applied to the off-diagonal values of the correlation matrix that represents the phylogenetic relationships (i.e., the **A** matrix). For example, the variance-covariance matrix for three species would be given by

$$\sigma^2 \begin{bmatrix} 1 & \lambda a_{12} & \lambda a_{13} \\ & 1 & \lambda a_{23} \\ & & 1 \end{bmatrix}$$

while the decomposition of the species-level heterogeneity in model 9 implies the variance-covariance matrix

$$\sigma_n^2 \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} + \sigma_p^2 \begin{bmatrix} 1 & a_{12} & a_{13} \\ & 1 & a_{23} \\ & & 1 \end{bmatrix} = (\sigma_n^2 + \sigma_p^2) \begin{bmatrix} 1 & \left(\frac{\sigma_p^2}{\sigma_n^2+\sigma_p^2}\right) a_{12} & \left(\frac{\sigma_p^2}{\sigma_n^2+\sigma_p^2}\right) a_{13} \\ & 1 & \left(\frac{\sigma_p^2}{\sigma_n^2+\sigma_p^2}\right) a_{23} \\ & & 1 \end{bmatrix}$$

and hence $\sigma^2 = \sigma_n^2 + \sigma_p^2$ and $\lambda = \sigma_p^2/(\sigma_n^2 + \sigma_p^2)$ (see also Lynch, 1991; Freckleton et al., 2002). Hence, $\sigma_p^2/(\sigma_n^2 + \sigma_p^2)$ indicates the degree of the phylogenetic signal in the overall variance sourced from the species. A likelihood ratio test of $H_0\colon \sigma_p^2 = 0$ can be easily performed by comparing $X^2 = -2(ll_7 - ll_9)$ against a chi-squared distribution with one degree of freedom, where ll_7 and ll_9 are the (restricted) log likelihoods of models 7 and 9, respectively. However, we do not advocate making changes to the model based on this test (i.e., by dropping the phylogenetic species random effect from the model if the test is not significant), since making changes to an a priori chosen model based on the data at hand affects the statistical properties of all inferential methods in unknown and unpredictable ways. Finally, we note that the (asymptotic) null distribution of the likelihood ratio test statistic is actually more complex than simply a chi-squared distribution with one degree of freedom, a result of the parameter being on the boundary of the parameter space under the null distribution (Self and Liang, 1987). The appropriate reference distribution for this test in the present context remains to be determined.

4.3 Estimating the Non-Phylogenetic and Phylogenetic Variance

Given the informative nature of these two variance components, it is essential to estimate their true values accurately to properly account for the sources of heterogeneity and dependency in the data. We found that model 9 was usually able to estimate these components with little to no bias, but should note that the model struggles to separate the non-phylogenetic and phylogenetic species effects when phylogenetic relationships are weak. In essence, the two sources of variability then start to collapse into one, with a total variance of $\sigma_n^2 + \sigma_p^2$. The way this total variance is then distributed into the two estimates is in essence arbitrary and can depend on the starting values or other settings of the model fitting algorithm. Therefore, we would caution against the use of model 9 when phylogenetic relationships are weak. As a rough guideline, for $\alpha = 0.5$, the mean correlation in the $\mathbf{A}$ matrix (excluding the diagonal) is around 0.2 and hence a lower mean correlation would call into question the trustworthiness of the estimates of σ_n^2 and σ_p^2.

Some meta-analyses in ecology and evolution have used model 11 to reduce model complexity (e.g., Garamszegi et al., 2012; Moore et al., 2016). Our results indicate that this approach cannot be recommended. As we increased the value of σ_n^2, the bias in the phylogenetic variance component inflated massively in this simplified model (Fig. 4b). As a result, the relevance of the phylogeny could be greatly overestimated. In addition, the confidence interval for the overall mean then becomes extremely conservative with coverage rates at or very close to 100%. This in turn implies a loss of efficiency for estimating the overall mean and a loss of power for testing $H_0\colon \mu = 0$. The illustrative example also shows this phenomenon.

4.4 Caveats and Conclusions

For the simulation study, we used a 'generic' effect size measure, that is, we directly simulated the sampling errors from a normal distribution and treated the sampling variances (i.e., the v_{ij} values) as known. These conditions only apply asymptotically to measures typically used in practice (e.g., standardized mean differences, response ratios, correlation coefficients, risk/odds ratios). The present results therefore reflect the performance of the various models under idealized conditions (i.e., when the sample sizes of the individual studies are sufficiently large, such that the sampling distributions of the estimates are indeed approximately normal and any inaccuracies in the estimated sampling variances are negligible). Although such ideal conditions are rare in practice (Hillebrand and J. Gurevitch, 2014; Pappalardo et al., 2020), the advantage of using a generic measure is that we were able to identify problems that are inherent to certain models and not (potentially) a consequence of violations to the model assumptions (i.e., if a particular model performs poorly for a measure that violates model assumptions, we do not know whether the poor performance is attributable to deficiencies of the model itself or a consequence of model assumptions being violated). On the other hand, it remains to be determined how well the phylogenetic multilevel model performs when the effect sizes are generated based on the exact distributional assumptions underlying specific measures.

Also, an issue we did not tackle in the present simulation study is the influence of the distribution of the different species over the simulated studies. In particular, concerns may arise when many of the primary studies included in a meta-analysis have examined only a single or closely related species. This may make it difficult to accurately estimate and differentiate between the study- and the species-level variance components. We did not generate conditions to specifically simulate such scenarios; thus, this issue still remains to be investigated in future simulation studies.

Therefore, at least for the moment, the present results suggest that model 9 is the most appropriate tool for conducting a multi-species meta-analysis in ecology and evolution (unless the phylogenetic relationships are weak, in which case model 7 may be preferable). For the vast majority of conditions examined, it provides approximately unbiased estimates of the variance components and the overall mean and a confidence interval for the latter with a close to nominal coverage rate. Therefore, we recommend that meta-analysts in ecology and evolution use the phylogenetic multilevel model as the de facto standard when analyzing multi-species datasets.

Conflict of interest statement: The authors declare that they have no competing interests.

Author contributions: SN provided contextual and literature review support, OC and WV wrote the code to run and analyze the results of the simulation, all authors contributed to the manuscript.

Data accessibility statement: No new data were used in this study. The material to reproduce the results are available at: https://osf.io/ms8eq/.

References

Adams, D. C. (2008). “Phylogenetic meta-analysis”. *Evolution* 62(3), 567–572. 10.1111/j.1558-5646.2007.00314.x.

Arnqvist, G. and D. Wooster (1995). “Meta-analysis: Synthesizing research findings in ecology and evolution”. *Trends in Ecology & Evolution* 10(6), 236–240. 10.1016/s0169-5347(00)89073-4.

Bates, D., R. Kliegl, S. Vasishth, and H. Baayen (2015). “Parsimonious mixed models”. *arXiv preprint arXiv:1506.04967.*

Blomberg, S. P., T. Garland Jr, and A. R. Ives (2003). “Testing for phylogenetic signal in comparative data: Behavioral traits are more labile”. *Evolution* 57(4), 717–745. 10.1554/0014-3820(2003)057[0717:TFPSIC]2.0.CO;2.

Borenstein, M., L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein (2011). *Introduction to meta-analysis.* Chichester, UK: Wiley.

Chamberlain, S. A., S. M. Hovick, C. J. Dibble, N. L. Rasmussen, B. G. Van Allen, B. S. Maitner, J. R. Ahern, L. P. Bell-Dereske, C. L. Roy, M. Meza-Lopez, et al. (2012). “Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis". *Ecology Letters* 15(6), 627–636. 10.1111/j.1461-0248.2012.01776.x.

Cooper, H., L. V. Hedges, and J. C. Valentine (2009). *The handbook of research synthesis and meta-analysis*. 2nd. New York: Russell Sage Foundation.

Egger, M., G. Davey-Smith, and D. Altman (2001). *Systematic reviews in health care: Meta-analysis in context*. 2nd. London: Wiley.

Felsenstein, J. (1985). "Phylogenies and the comparative method". *The American Naturalist* 125(1), 1–15. 10.1086/284325.

Felsenstein, J. (2004). *Inferring phylogenies*. 2nd. Sunderland, MA: Sinauer Associates.

Fernández-Castilla, B., M. Maes, L. Declercq, L. Jamshidi, S. N. Beretvas, P. Onghena, and W. Van den Noortgate (2019). "A demonstration and evaluation of the use of cross-classified random-effects models for meta-analysis". *Behavior Research Methods* 51(3), 1286–1304. 10.3758/s13428-018-1063-2.

Freckleton, R. P., P. H. Harvey, and M. Pagel (2002). "Phylogenetic analysis and comparative data: A test and review of evidence". *The American Naturalist* 160(6), 712–726. 10.1086/343873.

Garamszegi, L. Z., G. Markó, and G. Herczeg (2012). "A meta-analysis of correlated behaviours with implications for behavioural syndromes: Mean effect size, publication bias, phylogenetic effects and the role of mediator variables". *Evolutionary Ecology* 26(5), 1213–1235. 10.1007/s10682-012-9589-8.

Glass, G. V. (1976). "Primary, secondary, and meta-analysis of research". *Educational Researcher* 5(10), 3–8. 10.3102/0013189x005010003.

Grafen, A. (1989). "The phylogenetic regression". *Philosophical Transactions of the Royal Society of London, Series B* 326(1233), 119–157.

Gurevitch, J. and L. V. Hedges (1999). "Statistical issues in ecological meta-analyses". *Ecology* 80(4), 1142–1149. 10.1890/0012-9658(1999)080[1142:siiema]2.0.co;2.

Gurevitch, J., J. Koricheva, S. Nakagawa, and G. Stewart (2018). “Meta-analysis and the science of research synthesis”. *Nature* 555(7695), 175. 10.1038/nature25753.

Gurevitch, Jessica, Peter S Curtis, and Michael H Jones (2001). “Meta-analysis in ecology”. *Advances in Ecological Research* 32, 199–247.

Hadfield, J. D. and S. Nakagawa (2010). “General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters”. *Journal of Evolutionary Biology* 23(3), 494–508. 10.1111/j.1420-9101.2009.01915.x.

Hedges, L. V. and I. Olkin (1985). *Statistical models for meta-analysis.* New York: Academic Press.

Hedges, L. V. and J. L. Vevea (1998). “Fixed- and random-effects models in meta-analysis”. *Psychological Methods* 3(4), 486–504. 10.1037/1082-989x.3.4.486.

Hillebrand, H. and J. Gurevitch (2014). “Meta-analysis results are unlikely to be biased by differences in variance and replication between ecological lab and field studies”. *Oikos* 123(7), 794–799. 10.1111/oik.01288.

Ives, A. R., P. E. Midford, and T. Garland Jr (2007). “Within-species variation and measurement error in phylogenetic comparative methods”. *Systematic Biology* 56(2), 252–270. 10.1080/10635150701313830.

Konstantopoulos, S. (2011). “Fixed effects and variance components estimation in three-level meta-analysis”. *Research Synthesis Methods* 2(1), 61–76. 10.1002/jrsm.35.

Koricheva, J., J. Gurevitch, and K. Mengersen, eds. (2013). *Handbook of meta-analysis in ecology and evolution.* Princeton, NJ: Princeton University Press.

Lajeunesse, M. J. (2009). “Meta-analysis and the comparative phylogenetic method”. *The American Naturalist* 174(3), 369–381. 10.2307/40306065.

Luke, S. G. (2017). “Evaluating significance in linear mixed-effects models in R”. *Behavior Research Methods* 49(4), 1494–1502. 10.3758/s13428-016-0809-y.

Lynch, M. (1991). “Methods for the analysis of comparative data in evolutionary biology”. *Evolution* 45(5), 1065–1080. 10.1111/j.1558-5646.1991.tb04375.x.

Moore, F. R., D. M. Shuker, and L. Dougherty (2016). “Stress and sexual signaling: A systematic review and meta-analysis”. *Behavioral Ecology* 27(2), 363–371. 10.1093/beheco/arv195.

Nakagawa, S. and I. C. Cuthill (2007). “Effect size, confidence interval and statistical significance: A practical guide for biologists”. *Biological Reviews* 82(4), 591–605. 10.1111/j.1469-185x.2007.00027.x.

Nakagawa, S. and E. S. A. Santos (2012). “Methodological issues and advances in biological meta-analysis”. *Evolutionary Ecology* 26(5), 1253–1274. 10.1007/s10682-012-9555-5.

Nakagawa, S., A. M. Senior, W. Viechtbauer, and D. W. A. Noble (2021). “An assessment of statistical methods for non-independent data in ecological meta-analyses: Comment”. *Ecology*, 107–122. 10.1002/ecy.3490.

Noble, D. W. A., M. Lagisz, R. E. O’dea, and S. Nakagawa (2017). “Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses”. *Molecular Ecology* 26(9), 2410–2425. 10.1111/mec.14031.

Pagel, M. (1999). “Inferring the historical patterns of biological evolution”. *Nature* 401(6756), 877. 10.1038/44766.

Pappalardo, P., K. Ogle, E. A. Hamman, J. R. Bence, B. A. Hungate, and C. W. Osenberg (2020). “Comparing traditional and Bayesian approaches to ecological meta-analysis”. *Methods in Ecology and Evolution* 11(10), 1286–1295. 10.1111/2041-210X.13445.

Paradis, E. (2012). *Analysis of phylogenetics and evolution with R*. 2nd. New York: Springer.

Paradis, E. and K. Schliep (2019). “ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R”. *Bioinformatics* 35, 526–528. 10.1093/bioinformatics/bty633.

Patterson, H. D. and R. Thompson (1971). “Recovery of inter-block information when block sizes are unequal”. *Biometrika* 58(3), 545–554. 10.1093/biomet/58.3.545.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria.

Rios Moura, R., M. Oliveira Gonzaga, N. Silva Pinto, J. Vasconcellos-Neto, and G. S. Requena (2021). "Assortative mating in space and time: Patterns and biases". *Ecology Letters* 24, 1089–1102. 10.1111/ele.13690.

Sanchez-Meca, J. and F. Marin-Martinez (2008). "Confidence intervals for the overall effect size in random-effects meta-analysis". *Psychological Methods* 13(1), 31–48. 10.1037/1082-989x.13.1.31.

Self, S. G. and K. Y. Liang (1987). "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions". *Journal of the American Statistical Association* 82(398), 605–610.

Senior, A. M., C. E. Grueber, T. Kamiya, M. Lagisz, K. O'Dwyer, E. S. A. Santos, and S. Nakagawa (2016). "Heterogeneity in ecological and evolutionary meta-analyses: Its magnitude and implications". *Ecology* 97(12), 3293–3299. 10.1002/ecy.1591.

Song, C., S. D. Peacor, C. W. Osenberg, and J. R. Bence (2020). "An assessment of statistical methods for nonindependent data in ecological meta-analyses". *Ecology* 101(12), e03184. 10.1002/ecy.3184.

Sutton, A. J. and J. P. T. Higgins (2008). "Recent developments in meta-analysis". *Statistics in Medicine* 27(5), 625–650. 10.1002/sim.2934.

Viechtbauer, W. (2010). "Conducting meta-analyses in R with the metafor package". *Journal of Statistical Software* 36(3), 1–48. 10.18637/jss.v036.i03.