Modelling approaches for meta-analyses with dependent effect sizes in ecology and evolution: A simulation study

Coralie Williams[1,2*], Yefeng Yang[1], David I. Warton[1,2], and Shinichi Nakagawa[1,3]

[1]Ecology and Evolution Research Centre, School of Biological Earth and Environmental Sciences, The University of New South Wales, Sydney, Australia.

[2]School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia.

[3]Department of Biological Sciences, Faculty of Science, The University of Alberta, Edmonton, Canada

[*]coralie.williams@unsw.edu.au

Abstract

1. In ecology and evolution, meta-analysis is an important tool to synthesise findings across separate studies. However, ecological and evolutionary data often exhibit complex dependence structures, such as shared sources of variation within studies, phylogenetic relationships, and hierarchical sampling designs. Recent statistical advancements offer approaches for handling such complexities in dependence, yet these methods remain underutilised or unfamiliar to ecologists and evolutionary biologists.

2. We conducted extensive simulations to evaluate modelling approaches for handling dependence in effect sizes and sampling errors in ecological and evolutionary meta-analyses. We assessed the performance of multilevel models, incorporating an assumed sampling error variance-covariance matrix (which account for within-study correlation), cluster robust variance estimation (CRVE) methods and their combination across different true within-study correlations. Finally, we showcased the applications of these models in two case studies of published meta-analyses.

3. Multilevel models produced unbiased regression coefficient estimates and when a sampling variance-covariance matrix was used it provided accurate random effect variance components estimates within and among studies. However, the latter had no impact on regression coefficient estimates if the model was misspecified. The inclusion of CRVE methods, either alone or combined with multilevel models, did not enhance performance. In simulations involving phylogenetic multilevel meta-analysis, models using CRVE methods generated narrower confidence intervals and lower coverage rates than the nominal expectations. The case study results showed the importance of considering a sampling error variance-covariance matrix to improve the model fit.

4. Our results provide clear modelling recommendations for ecologists and evolutionary biologists conducting meta-analyses. To improve the precision of variance component estimates we recommend constructing a variance–covariance matrix that accounts for dependencies in sampling errors within studies. Although CRVE methods provide robust inference under certain conditions, we caution against their use with crossed random effects, such as phylogenetic multilevel meta-analyses, as CRVE methods currently do not account for multi-way clustering and may inflate Type I error rates. Finally, we recommend using multilevel meta-analytic models to account for heterogeneity at all relevant hierarchical levels and to follow guidance on inference methods to ensure accurate coverage of the overall mean.

Key-words: Cross-classified data, Phylogenetic comparative methods, Meta-regression, Mixed-effects models, Multi-species, Non-independence, Sandwich estimators

1 Introduction

In ecology and evolution, meta-analysis has been used to make broader generalisation from results across global scales, long time spans, and across multiple species, while identifying sources of variability (Arnqvist & Wooster, 1995; Nakagawa & Poulin, 2012; Stewart, 2009). By systematically combining the quantitative results of independent studies, meta-analysis estimates an overall effect size and identifies factors influencing variation among effect sizes. However, data in ecology and evolution often exhibit complex dependence structures which require advanced approaches to ensure appropriate meta-analytical inference (Gurevitch & Hedges, 1999; Koricheva & Gurevitch, 2014; Nakagawa & Santos, 2012).

Meta-analytical data can have multiple sources of dependence in their structure which can be broadly divided into two types. The first and most common is dependence among effect sizes. This occurs when effect sizes come from the same primary study, experiment, treatment, location, or another grouping feature, and are therefore correlated with each other. Further, meta-analyses in ecology and evolution often involve multiple species. In this case effect sizes from the same species are also correlated due to shared evolutionary history (Chamberlain et al., 2012; Gurevitch & Hedges, 1999). The second, often overlooked, type of dependence is among sampling errors. This type of dependence may arise, for example, when multiple measurements are taken from the same subject or group of animals, or when treated subjects are compared with the same controls in the context of comparative treatment-control studies. This dependence leads to the sampling errors to be correlated within studies or subgroups. A survey of meta-analysis in environmental sciences found that only 9% of surveyed meta-analysis used methods to account for dependence in sampling errors (Nakagawa et al., 2023). In the past two decades, new and innovative methods for handling these two forms of dependence have emerged, but are currently underutilised by ecologists and evolutionary biologists conducting meta-analyses today.

Historically, there are three approaches to deal with dependence structures in meta-analysis, as described in Becker (2000): (1) ignore dependence, (2) aggregate (making *ad hoc* changes to the data to avoid dependence), and (3) model dependence (using integrative strategies, *i.e.* methods that do not modify the original dataset). The first approach, which ignores dependence, is not recommended as it underestimates standard errors and increases the risk of false positives (Type I errors). The second approach, aggregating data, yields unbiased estimates but leads to loss of information, as it restricts opportunities for meta-regression and the estimation of the variance components of random effects (Nakagawa et al., 2022; Pustejovsky & Chen, 2024). The most flexible approach to dealing with dependence is the third approach of modelling (Tipton

et al., 2019). Multilevel models can account for hierarchical structures in effect sizes by including random effects (Pastor & Lazowski, 2018; Van Den Noortgate & Onghena, 2003; Van den Noortgate et al., 2013). However, the information about the amount of dependence among sampling errors is often not reported in primary studies (Lajeunesse, 2009, 2011; Noble et al., 2017). To model unknown dependencies among sampling errors from the same study one can incorporate an assumed within-study correlation within the sampling variance-covariance (VCV) matrix. To avoid making any assumptions about correlations among effect sizes and potential model misspecification, Hedges et al. (2010) proposed to use cluster robust variance estimation (CRVE) methods, also known as sandwich estimator methods. CRVE methods offer an effective approach to account for dependencies in sampling errors, though it is important to understand their limitations, as certain CRVE methods can perform poorly with small sample sizes. In a recent study, Pustejovsky and Tipton (2022) proposed a new working model that combines multilevel meta-analytical models, an assumed sampling error variance covariance matrix, and cluster robust variance estimation with simulations demonstrating that this approach enhances the precision of regression estimates. Currently, no simulation study has assessed the above modelling approaches and their combination in the context of ecological and evolutionary meta-analyses, specifically, when meta-analyses have an unbalanced design and include multiple species. As meta-analytic findings can inform evidence-based policy decisions (Haddaway & Pullin, 2014; Maynard, 2024), neglecting to account for such dependence structures may lead to erroneous inferences that could misinform such policies and conservation management decisions.

In this paper, we conduct a simulation study to evaluate the performance of different meta-analysis modelling approaches to account for dependence in effect sizes and sampling errors. We compare two approaches under different working models: one that specifies a within-study error variance-covariance (VCV) matrix assuming constant correlation, and another that incorporates a cluster robust variance estimator (CRVE) in the context of ecological and evolutionary data. For practical applicability, we focus on different strategies for including a within-study VCV, CRVE methods, and their combination, while also assessing how the incorporation of phylogenetic random effects influence model efficiency. This study aims to highlight current strategies for dealing with unknown dependence of effect sizes and sampling errors. Despite the emergence of new tools and modelling approaches, the guidance for applying them to complex ecological and evolutionary dependent dataset structures remain limited. Below we provide clear recommendations based on our simulation results.

2 Methods

We registered our study's protocol in May 2024, detailing the methodological plan following the ADEMP-PreReg template provided in Siepe et al. (2023). We reported our simulation items in accordance with the guidance provided by Morris et al. (2019) and Williams et al. (2024).

2.1 Meta-analytic models and assumptions

Meta-analysis synthesises effect size estimates obtained from multiple primary studies, allowing researchers to evaluate the magnitude and direction of a particular effect or association. In ecology and evolution, commonly used effect size measures include standardised mean differences, response ratios, correlation coefficients, and risk or odds ratios. The sampling variances associated with these effect sizes, reflecting the uncertainty in their estimation, are assumed to be known as they are either provided by the original study or can be calculated from estimated parameters in the original study data. Hence, when sample sizes are sufficiently large, these calculated sampling variances can be treated as approximately known.

2.1.1 Fixed-effect (FE) and random-effects (RE) models

Here, we define y_i to be the effect size estimate of the ith study (if all studies report a single effect size, the terms study and effect size are interchangeable, $N_{\text{studies}} = N_{\text{total}}$) and with corresponding sampling variances v_i.

The simplest meta-analytical model is the FE model, defined as

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

$$i = 1, \ldots, N_{\text{studies}}$$

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

where μ is the overall mean and $\mathbf{e}$ is the sampling error term which we assume to be normally distributed with mean 0 and with a variance-covariance matrix $\mathbf{V}$ where sampling variances v_i are along the diagonal. This fixed-effect model (sometimes called common-effects or equal-effects model in the meta-analytical literature)

assumes that the underlying effect sizes have the same true effect, which is often not the case in ecological and evolutionary meta-analyses due to data with multiple species (Senior et al., 2016).

To account for this variability in true effects, the RE model can incorporate a random effect at the estimate level, and is defined as

$$y_i = \mu + u_i + e_i \tag{2}$$

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

where u_i is a random effect corresponding to the ith effect size estimate (i.e. equivalent to study as there is one effect size per study), assumed to be normally distributed with mean 0 and variance σ_u^2, and $\mathbf{I}_u$ is an identity matrix of size $N_{studies} \times N_{studies}$. This random effects model, assumes all effect sizes across studies are independent and that their sampling variances have no dependence structure. However, as we described earlier, most studies in ecology and evolution involve more than one effect size per study (Senior et al., 2016) and sampling errors are likely related due to study design.

2.1.2 Multilevel models (ML)

To address this first type of dependence, **dependence among effect sizes**, we can include an additional random effect at the study level, creating a multilevel model. We define y_{ij} the jth effect size estimate in the ith study as the multilevel model with two 'levels' as

$$y_{ij} = \mu + u_{ij} + s_i + e_{ij} \tag{3}$$

$$i = 1, \ldots, N_{\text{studies}}$$

$$j = 1, \ldots, N_{\text{total}}$$

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

where u_{ij} denotes the random effect of the jth effect estimate in the ith study, s_i is the study level random effect, assumed to be normally distributed with mean 0 and variance σ_s^2 ($\mathbf{I}_s$ denotes the identity matrix).

This multilevel model assumes independence among sampling errors within studies (*i.e.* for any two effect sizes from the ith study the covariance of sampling errors would be zero: $Cov(v_{ij}, v_{ij'}) = 0$; where j and j' are distinct effect sizes). Note that this model can be expanded with more 'levels' (i.e. random effects) to capture other hierarchical dependencies present in the data, for example site, exposure, treatment etc.

The second type of dependence, **dependence among sampling errors**, as described earlier, can occur when estimates are correlated due to effect sizes being calculated from the same cohort, sample, or due to shared controls. Using information from each study's primary data, we can calculate the covariances of effect size pairs. However, this information is often not available, or only available for a few studies, and usually all we have available from study i is the vector of error variances $\mathbf{v}_i$ for each effect size. To address this, an approach is to assume an arbitrary constant correlation, which we define as ρ, between effect size estimates coming from the same study. Then we assume the vector of within-study errors across all studies, $\mathbf{e} = vec(e_{ij})$, is distributed as:

$$\mathbf{e} \sim N(0, \mathbf{V}^*) \tag{4}$$

where the variance-covariance (VCV) matrix $\mathbf{V}^*$ is block diagonal, where the ith block has diagonals equal to the sampling variances $\mathbf{v}_i$ of the respective effect sizes for study i, and its off-diagonals are the covariances between each effect size, assuming common correlation ρ. For example, the covariance of any two effect sizes j and j' from study i is $Cov(v_{ij}, v_{ij'}) = \rho\sqrt{v_{ij}v_{ij'}}$. In ecology and evolution, a constant within-study $\rho = 0.5$ has been recommended (Noble et al., 2017) to assume a conservative correlation among effect sizes. Certain software implementations assume an arbitrary higher constant correlation $\rho = 0.8$ as default (Fisher et al., 2023) which may be more applicable for human studies (e.g. psychology, education) where effect sizes can be more correlated. We further assume there is no correlation between sampling errors from different studies, that is, we assume $Cov(v_{ij}, v_{i'j'}) = 0$ for $i \neq i'$, hence $\mathbf{V}^*$ has a block-diagonal structure.

Below we specify an example of constructing the **V*** block diagonal sampling VCV matrix for a dataset with seven effect sizes from two studies, assuming a constant within-study correlation. To improve readability we have added a comma between the subscripts of studies (i) and effect sizes (j). The first study includes four effect sizes (with associated variances $v_{1,1}$, $v_{1,2}$, $v_{1,3}$, $v_{1,4}$) and the second study includes three effect sizes (with associated variances $v_{2,5}$, $v_{2,6}$, $v_{2,7}$). Variances and covariances are coloured in teal for the first study and in olive for the second to differenciate them.

$$\mathbf{V}^* = \begin{bmatrix} v_{1,1} & \rho\sqrt{v_{1,1}v_{1,2}} & \rho\sqrt{v_{1,1}v_{1,3}} & \rho\sqrt{v_{1,1}v_{1,4}} & 0 & 0 & 0 \\ \rho\sqrt{v_{1,2}v_{1,1}} & v_{1,2} & \rho\sqrt{v_{1,2}v_{1,3}} & \rho\sqrt{v_{1,2}v_{1,4}} & 0 & 0 & 0 \\ \rho\sqrt{v_{1,3}v_{1,1}} & \rho\sqrt{v_{1,3}v_{1,2}} & v_{1,3} & \rho\sqrt{v_{1,3}v_{1,4}} & 0 & 0 & 0 \\ \rho\sqrt{v_{1,4}v_{1,1}} & \rho\sqrt{v_{1,4}v_{1,2}} & \rho\sqrt{v_{1,4}v_{1,3}} & v_{1,4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & v_{2,5} & \rho\sqrt{v_{2,5}v_{2,6}} & \rho\sqrt{v_{2,5}v_{2,7}} \\ 0 & 0 & 0 & 0 & \rho\sqrt{v_{2,6}v_{2,5}} & v_{2,6} & \rho\sqrt{v_{2,6}v_{2,7}} \\ 0 & 0 & 0 & 0 & \rho\sqrt{v_{2,7}v_{2,5}} & \rho\sqrt{v_{2,7}v_{2,6}} & v_{2,7} \end{bmatrix}$$

2.1.3 Phylogenetic multilevel meta-analysis models (PML)

To account for multiple effect sizes across different species we can add random effects at the species level. From recent simulations from Cinar et al. (2022), including both a phylogenetic and non-phylogenetic random effect in meta-analytical models provides improved inference. This then extends the multilevel model in Equation 3 to

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

$$k = 1, \ldots, N_{\text{species}}$$

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

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

where y_{ijk} is the effect size of the jth estimate, of the ith study and of the kth species. The component n_k is a species level random effect, assumed to be normally distributed with mean 0 and variance σ_n^2 and identity matrix $\mathbf{I}_n$, assuming species are independent to each other. To account for the shared evolutionary history between species, a second random effect at the species level p_k is incorporated, which has a variance of σ_p^2 and $\mathbf{A}$ is the phylogenetic correlation matrix of size $N_{species} \times N_{species}$. We note that the species level effects (phylogenetic and non-phylogenetic) are crossed among studies, which means any given species can have effect sizes coming from multiple studies.

The phylogenetic meta-analysis model described in Equation 5 can also incorporate a variance-covariance matrix ($\mathbf{V}^*$) for the sampling errors (see Equation 4) to account for correlated errors within-studies.

2.2 Cluster-robust variance estimators (CRVE)

In the previous section, we described multilevel models with a fixed sampling VCV, in which we needed to assume a known, constant correlation across studies, in order to account for correlated sampling errors (often in the absence of direct measurements of it). To relax this assumption, cluster robust variance-covariance estimators (CRVE) have been introduced in meta-analysis to model dependent effect sizes from the same study when the true dependence structure is unknown (Hedges et al., 2010). CRVE stem from robust variance estimators, also known as sandwich estimators or Huber White estimators, which are designed to handle heteroscedasticity (Sidik & Jonkman, 2005; White, 1980). Even when the working model is misspecified, meta-regression coefficient estimates with CRVE have asymptotically consistent standard errors. Hence, hypothesis tests and confidence intervals are valid when appropriate small-sample adjustments are used and the number of clusters is sufficiently large. We present three of the main CRVE methods implemented in the `clubSandwich` R package (Pustejovsky, 2023), also available in the `metafor` package via the `robust()` function (Viechtbauer, 2023). We note that other methods, such as cluster wild bootstrapping (Joshi et al., 2022), are available but we do not cover them here. The original robust sandwich estimator (as popularised in Liang & Zeger, 1986), which we will refer to as **CR0** as per Cameron and Miller (2015), estimates the standard errors of coefficients empirically and without imposing structural correlation assumptions. However, when cluster numbers are small (less than 50 studies), which is likely in meta-analysis in ecology and evolution, the **CR0** method is downwardly biased for variance components as well as having high Type I error rates of associated hypothesis tests (Tipton & Pustejovsky, 2015; Viechtbauer et al., 2015). To address this issue, a number of CRVE methods have been proposed to enhance inference accuracy when the number of clusters is small. Briefly, the **CR1** method provides an approximate correction for when the number of clusters is small. The **CR2** method provides a "bias-reduced linearisation" adjustment for small (study) sample sizes which was initially proposed by Bell and McCaffry (2002) and further developed in Pustejovsky and Tipton (2018). Using the **CR2** method with the Satterthwaite approximation of effective degrees of freedom controls for Type-I error rates (Tipton & Pustejovsky, 2015). However, currently there is no statistical theory to support multi-way clustered standard errors for models with crossed random effects, hence **CR2** can't be used with phylogenetic meta-analytical models (Equation 5), *i.e.* when species are distributed across multiple studies.

2.3 Simulation study

We conducted two inter-related simulation studies following a similar design as Cinar et al. (2022), to assess performance of the models and CRVE methods we presented earlier under different dependence structures. The first study, Study 1, compared meta-analysis models detailed in Equations 1-3. The second study, Study 2, compared performance of phylogenetic multilevel meta-analysis models (PML) detailed in Equation 5. For ML and PML using a V^* matrix, we assumed a constant within-study correlation ρ for sampling errors, and considered each of $\rho \in (0.2, 0.5, 0.8)$. We summarised the simulation settings per model in Table 1.

For both studies, we used a data-generating process inspired by real meta-analysis data from ecology and evolution (Senior et al., 2016), which also informed the simulation design in Cinar et al. (2022) (see Supporting Information Figure S1). The number of effect sizes per study were simulated as an unbalanced design with random values generated from a beta distribution with parameters $\alpha = 1.5$ and $\beta = 3$ (making a right-skewed distribution), scaled by a factor of 39, rounded to the nearest integer, and incremented by one. For all simulations, we considered an overall mean effect size $\mu = 0.2$. The test statistics and confidence intervals of the overall mean estimate $\hat{\mu}$ were computed assuming a t-distribution and adjusted degrees of freedom (more detail below). We simulated sampling errors assuming dependence of effect sizes within-studies, following a multivariate normal distribution with mean vector 0 and sampling error variance-covariance matrix. We generated the sampling error variance-covariance matrix assuming a true constant within-study effect size correlation, defined as ϕ, and assumed the sampling error variances, v_{ij}, followed a right-skewed beta distribution with parameters $\alpha = 2$ and $\beta = 20$, resulting in a mean sampling variance of 0.091. We considered three values of true correlation within-study, $\phi \in \{0, 0.2, 0.5, 0.8\}$, to reflect different levels of dependence and to match models with assumed sampling error V^* matrix structures. Note that when we fitted models that assumed within-study error correlation, we considered all three values ($\rho \in \{0.2, 0.5, 0.8\}$) irrespective of the actual correlation (ϕ) at which data were simulated, in order to understand robustness of the method to misspecification.

For Study 1, we considered $N_{studies} \in (20, 50)$ studies, and variance components values of $(\sigma^2_u, \sigma^2_s) \in (0.05, 0.3)$. For Study 2, we considered scattershot combinations of the number of studies and the number of species, with two combinations: $(N_{\text{studies}}, N_{\text{species}}) = (20, 40)$ and $(N_{\text{studies}}, N_{\text{species}}) = (50, 100)$. For the variance components in Study 2 we considered $(\sigma^2_u, \sigma^2_s, \sigma^2_p, \sigma^2_n) \in (0.05, 0.3)$. We simulated species indices assuming a beta distribution with parameters $\alpha = 2$ and $\beta = 2$, which were scaled by the number of species minus one, rounded, and increased by one. We randomly generated phylogenetic trees and computed branch

lengths assuming a power parameter α of 1 based on results in Cinar et al. (2022), using the `rtree` function from the `ape` package (Paradis et al., 2023). The phylogenetic correlation matrix (matrix **A** in Equation 5) was computed assuming a Brownian motion model of evolution.

For all models and simulation conditions, we assessed the bias and mean squared error (MSE) of the overall mean estimates, and variance components. Further, we evaluated the precision and consistency of the overall mean estimates by assessing the 95% coverage rates and widths of confidence interval. We performed 5,000 simulation repetitions per condition. The Monte Carlo Standard Error (MCSE) for 5,000 repetitions will be lower than 1% for bias, MSE and coverage measures for each one of the models in the simulation studies (Morris et al., 2019). All our simulations were conducted using open-source software `R` version 4.3.1 (R-Core-Team, 2022). The `metafor` package version 4.6-0 was employed to fit meta-analysis models (Viechtbauer, 2023) assuming a restricted maximum likelihood (REML) estimation, the default setting of the `rma.mv` function. The adjusted degrees of freedom were specified in the model using `dfs="contain"` argument which calculates the degrees of freedom for the overall mean coefficient by checking whether its predictor varies at a specific random effect level, then using the number of unique values of that effect minus one as the degrees of freedom. All simulations were run on the high performance computing (HPC) cluster Katana supported by Research Technology Services at UNSW Sydney (UNSW, 2024).

2.4 Additions and deviations

Meta-analyses often assess whether effect sizes vary based on certain study characteristics. To account for these characteristics (commonly referred as moderators or predictor variables) researchers can employ meta-regression models, which help to explore heterogeneity and control for potential confounders. We extended our protocol to evaluate meta-regression models by simulating phylogenetic multilevel models with moderators *i.e.* predictor variables. This analysis followed the same design as simulation Study 2 but included three moderators: a study-level categorical moderator (e.g., treatment type), a species-level continuous moderator (e.g., species weight), and an observation/effect size level categorical moderator (e.g., sex). Expanding on the phylogenetic meta-analysis from Equation 5, the phylogenetically controlled meta-regression model with the three described moderators is defined as

$$y_{ijk} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2k} + \beta_3 x_{3ij} + u_{ij} + s_i + n_k + p_k + e_{ij} \tag{6}$$

where β_0 is the fixed intercept coefficient, $\beta_1, \beta_2, \beta_3$ are the fixed effect coefficients for the predictor variables x_{1i}, x_{2k}, x_{3ij}. For the simulation study, we assume true values of $\beta_0 = 0.2$, $\beta_1 = 0.6$, $\beta_2 = 0.2$, and $\beta_3 = 0.5$. The tests of individual fixed coefficients in the meta-regression model and the corresponding confidence intervals were based on a t-distribution, and the omnibus test based on a F-distribution. In the meta-regression, each coefficient's adjusted degrees of freedom were computed by subtracting the total number of model coefficients (including the intercept) from the number of unique levels of the random effect over which the corresponding predictor varied using the using `dfs="contain"` argument in `metafor` (Viechtbauer, 2023).

3 Results

3.1 Study 1: Meta-analysis models

Figure 1 displays the performance of the six different working models (FE, RE, ML, ML-VCV-0.2, ML-VCV-0.5, and ML-VCV-0.8) for estimating the overall mean $\hat{\mu}$ across varying true within study correlation ϕ. All models had unbiased overall mean estimates $\hat{\mu}$ (Figure 1A and Table S1). We found that FE (Fixed-Effects) model exhibited higher variability and higher mean squared error (MSE) compared to other models (Figure 1.A, 1.B, and Table S1). Multilevel (ML) models, including ML models with assumed sampling VCV (i.e. $\mathbf{V}^*$), had identical lower and more consistent MSE across all conditions (Figure 1.B). Figure 1.C displays the coverage rates of the 95% confidence intervals, revealing that FE and RE (Random-Effects) generally fail to achieve the nominal 95% coverage, while the four ML models achieves coverage closer to the target across conditions (Table S2). Further, we found the FE model had the narrowest confidence intervals widths (Figure 1.D), whereas they were larger for the multilevel models. We note that ML-VCV-0.8 showed slightly narrower confidence interval widths with higher corresponding MSE. Figure S2 displays the 95% coverage rates of the four ML models across three different inference methods showing the assumed t-distribution with adjusted degrees of freedom is at the nominal coverage rate compared to inferences assuming a z-distribution or t-distribution without any degrees of freedom adjustment.

The coverage rate and width of the 95% confidence interval of the overall mean estimates $\hat{\mu}$ are presented in Figure 2 across six working models and four approaches: no CRVE method, CR0, CR1, and CR2. We found that the multilevel (ML) models with and without assuming a sampling VCV consistently achieved coverage close to the nominal 95% no matter the CRVE method, while FE and RE showed lower coverage

Table 1: Simulation parameters

Simulation	Conditions no.	Model	CRVE method	ρ	N_{studies}	N_{species}	σ_u^2	σ_s^2	σ_n^2	σ_p^2
Study 1	24	FE	none, CR0, CR1, CR2	0	20, 50	1	0	0	0	0
	48	RE	none, CR0, CR1, CR2	0	20, 50	1	0.05, 0.3	0	0	0
	96	ML	none, CR0, CR1, CR2	0	20, 50	1	0.05, 0.3	0.05, 0.3	0	0
	96	ML-VCV-0.2	none, CR0, CR1, CR2	0.2	20, 50	1	0.05, 0.3	0.05, 0.3	0	0
	96	ML-VCV-0.5	none, CR0, CR1, CR2	0.5	20, 50	1	0.05, 0.3	0.05, 0.3	0	0
	96	ML-VCV-0.8	none, CR0, CR1, CR2	0.8	20, 50	1	0.05, 0.3	0.05, 0.3	0	0
Study 2	288	PML	none, CR0, CR1	0	20 or 50	40 or 100	0.05, 0.3	0.05, 0.3	0.05, 0.3	0.05, 0.3
	288	PML-VCV-0.2	none, CR0, CR1	0.2	20 or 50	40 or 100	0.05, 0.3	0.05, 0.3	0.05, 0.3	0.05, 0.3
	288	PML-VCV-0.5	none, CR0, CR1	0.5	20 or 50	40 or 100	0.05, 0.3	0.05, 0.3	0.05, 0.3	0.05, 0.3
	288	PML-VCV-0.8	none, CR0, CR1	0.8	20 or 50	40 or 100	0.05, 0.3	0.05, 0.3	0.05, 0.3	0.05, 0.3

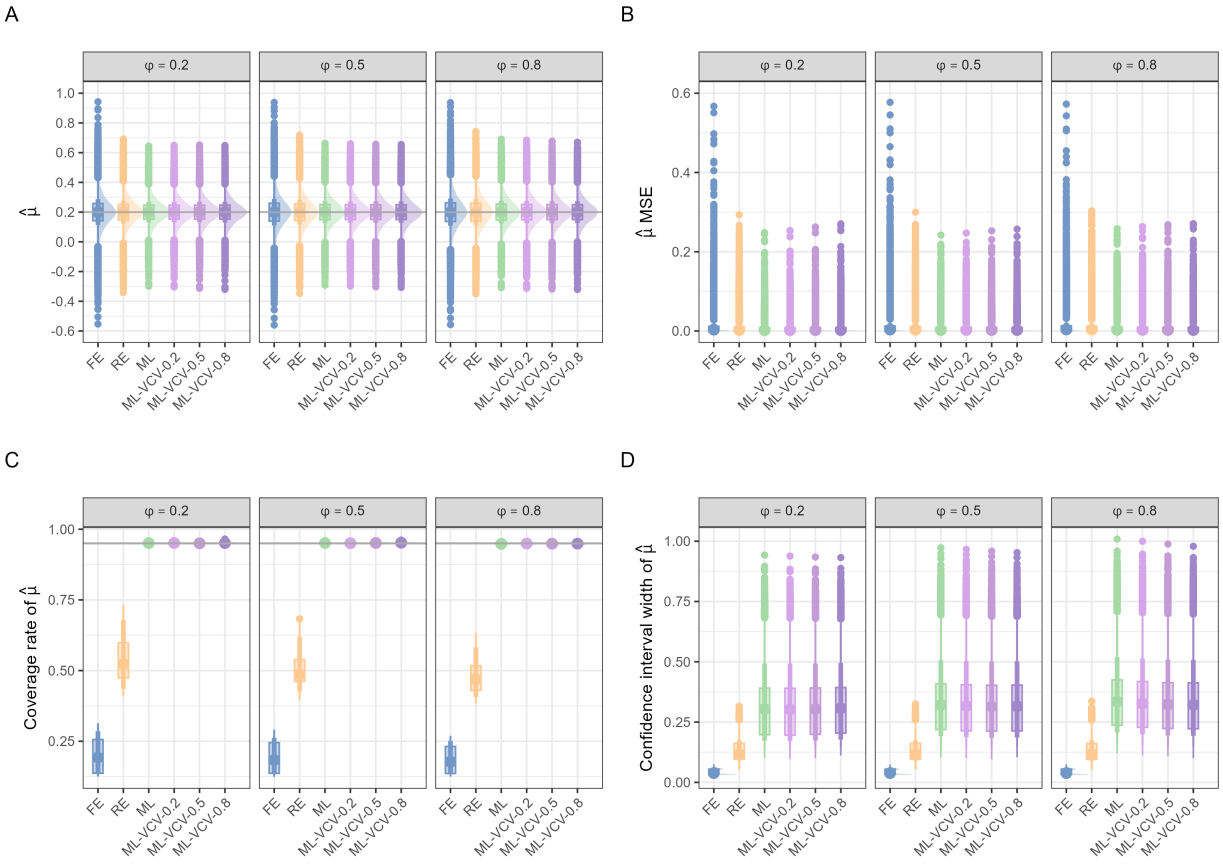

Figure 1: Overall mean estimate $\hat{\mu}$ performance across all working models and conditions assumimg a true within study correlation between effect sizes of $\phi \in (0.2, 0.5, 0.8)$, evaluated over 5,000 simulation iterations. **A.** The bias of the overall mean estimate $\hat{\mu}$, reflecting the deviation from the true mean. Monte Carlo standard errors of the overall mean bias are provided in Table S1. **B.** The mean squared error (MSE) of $\hat{\mu}$, combining both bias and variance to measure accuracy. **C.** The coverage rates of the 95% confidence intervals, indicating the proportion of intervals that include the true mean μ and assessing the reliability and consistency of the interval estimates. Monte Carlo standard errors of the overall mean coverage rate are provided in Table S2. **D.** The widths of the 95% confidence intervals, representing the precision of the estimates across different conditions.

but approximately close to 95% for CR2 method (Figure 2.A). The confidence interval widths of FE and RE models without any CRVE method were narrower while having low coverage of the overall mean estimate (Figure 2.B). The confidence interval widths of ML models were identical and did not change no matter the CRVE method.

Figure 3 displays the distribution of the conditional variance components estimates within study ($\hat{\sigma}_u^2$) and among studies ($\hat{\sigma}_s^2$). The FE models and the RE models are not shown as they did not estimate these variance components. Figure 3.A shows the RE models overestimated the within-study variance components and had high variability, while multilevel (ML) models were closer to the true value when $\sigma_u^2 = 0.3$. For the among-study conditional variance estimates ($\hat{\sigma}_s^2$), Figure 3.B shows the ML without assuming a sampling $\mathbf{V}^*$ matrix overestimated variances for higher correlations within studies ($\phi > 0.2$). Similar patterns were found for other true variance component conditions (see Figure S3, S4, S5, and Table S3). As for the total variance estimates ($\hat{\sigma}_{total}^2 = \hat{\sigma}_u^2 + \hat{\sigma}_s^2$), we found smaller mean squared errors (MSE) in models assuming a sampling $\mathbf{V}^*$ matrix for higher true within-study correlations $\phi > 0.2$. Similar patterns were found for other true variance component conditions displayed in Supporting Figure S6. All models in Study 1 converged and showed no errors in the estimation process, and computed in less than 3 seconds (Supporting Table S8).

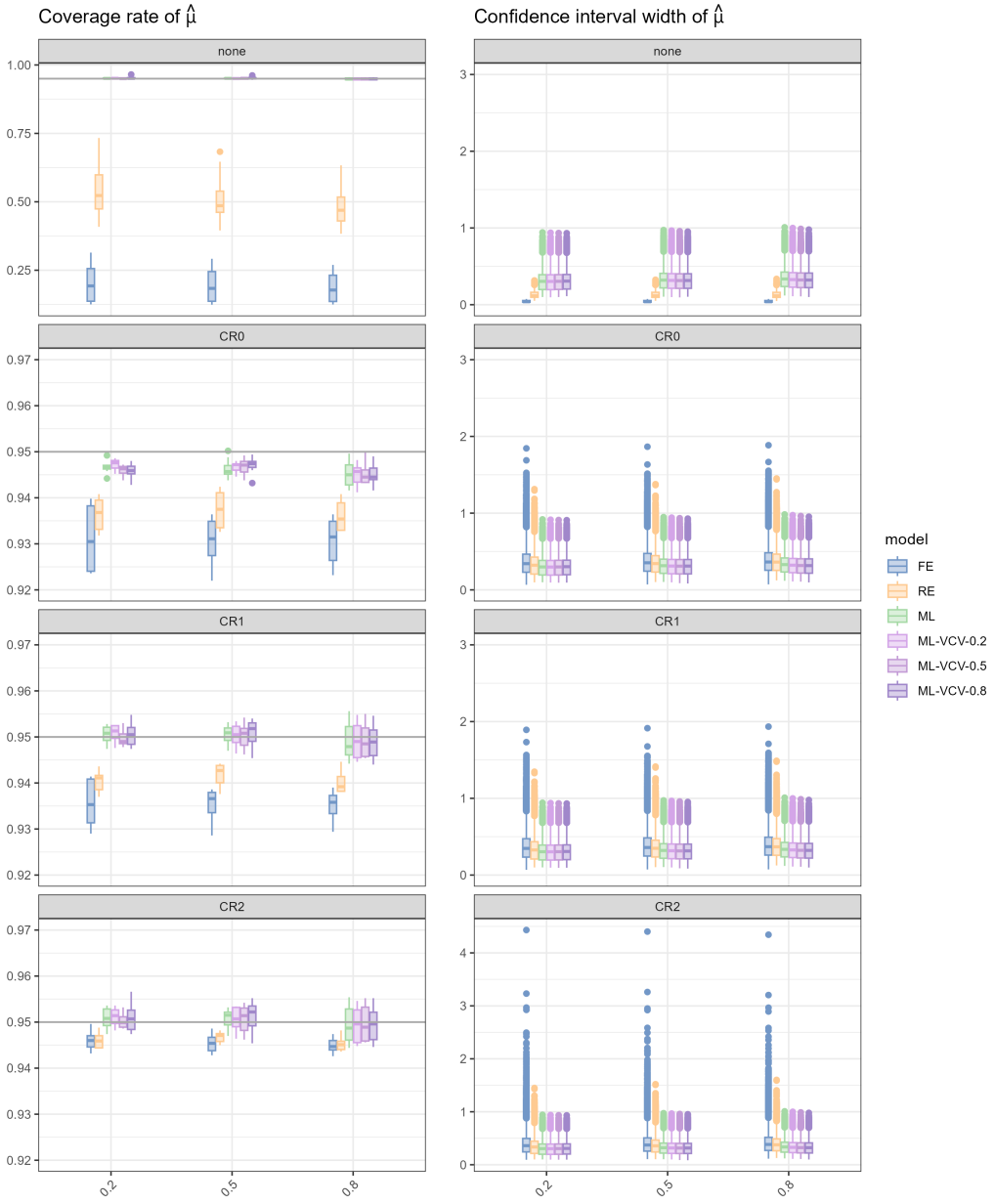

Figure 2: Boxplots of the overall mean estimate $\hat{\mu}$ coverage rate and confidence intervals for each CRVE method under working models across all conditions. **A.** The coverage rates of the 95% confidence intervals, indicating the proportion of intervals that include the true mean μ and assessing the reliability and consistency of the interval estimates **B.** The widths of the confidence intervals. The results were evaluated across 5,000 simulation iterations, eight conditions of variance components (σ_u^2, σ_s^2) and the number of studies (k_{studies}).

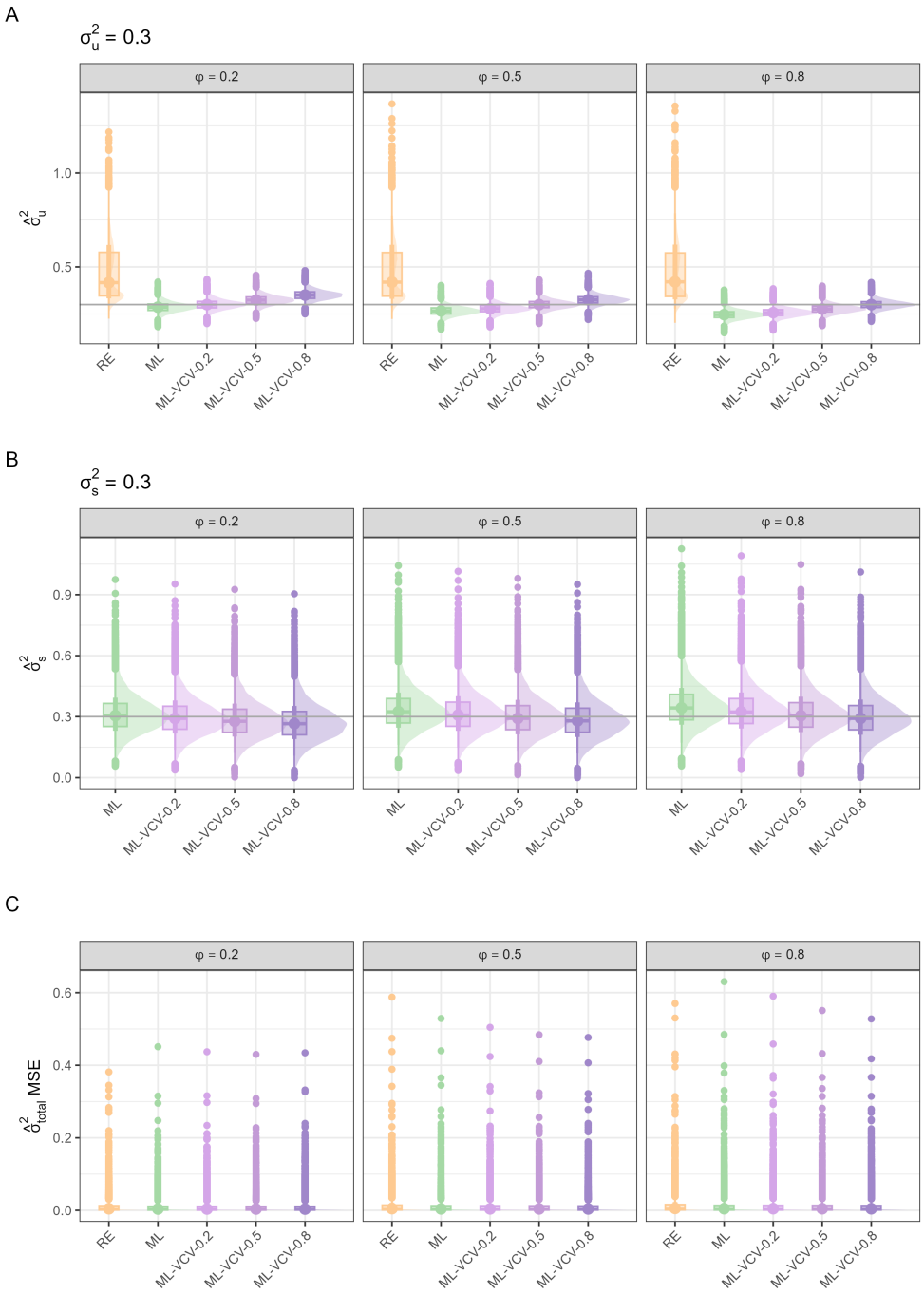

Figure 3: **A.** Boxplots of within-study conditional variance estimates ($\hat{\sigma}_u^2$) under true values of $\sigma_u^2 = 0.3$ and across within study correlation levels $\phi \in 0.2, 0.5, 0.8$. **B.** Boxplots of among study under conditional variance estimates ($\hat{\sigma}_s^2$) under true values of $\sigma_s^2 = 0.3$ and across within study correlation levels $\phi \in 0.2, 0.5, 0.8$. For both panels **A** and **B**, the true variance is shown in the grey bolded line and the boxplot represent the variability of estimates across 5,000 simulations. **C.** Distribution of mean squared error (MSE) of the total conditional variance estimates of models ($\hat{\sigma}_{total}^2 = \hat{\sigma}_u^2 + \hat{\sigma}_s^2$) under true values of $\sigma_u^2 = 0.3$ and $\sigma_s^2 = 0.3$, and within study correlation levels of $\phi \in (0.2, 0.5, 0.8)$. Models that did not estimate among study variation had $\hat{\sigma}_s^2 = 0$.

3.2 Study 2: Phylogenetic meta-analysis and meta-regression models

3.2.1 Phylogenetic multilevel meta-analysis

We found no clear difference in the bias, MSE, coverage rate and width of confidence intervals of the four phylogenetic multilevel working models (PML, PML-VCV-0.2, PML-VCV-0.5, PML-VCV-0.8) across the three true values for within study correlation (see SFigure 2). Figure 4 displays boxplots of coverage rate and confidence interval widths of the overall mean estimates of the four phylogenetic multilevel working models (PML, PML-VCV-0.2, PML-VCV-0.5, PML-VCV-0.8) across three dependence structures for each CRVE method. Coverage rates are closer to 95% nominal when no CRVE method is used, which reached on average 66-68% across all working models (Figure 4A). Confidence intervals were narrower with CRVE, whereas without CRVE, widths were approximately twice as large (Figure 4.B). Figure 5 displays distribution in boxplots of the conditional variances of the four random effects in each working model. As the true correlation within study increases, $\phi \in (0.2, 0.5, 0.8)$, the PML working model, which assumes no correlation among effect sizes from the same study ($\rho = 0$), provided an estimate of the variance component within study ($\hat{\sigma}_u^2$) that was downwardly biased and the estimated variance component among studies ($\hat{\sigma}_s^2$) that was upwardly biased. The majority of models converged (at least 99.99% of models showed no errors in the estimation process) and were computed within 6 seconds (Supporting Table S10).

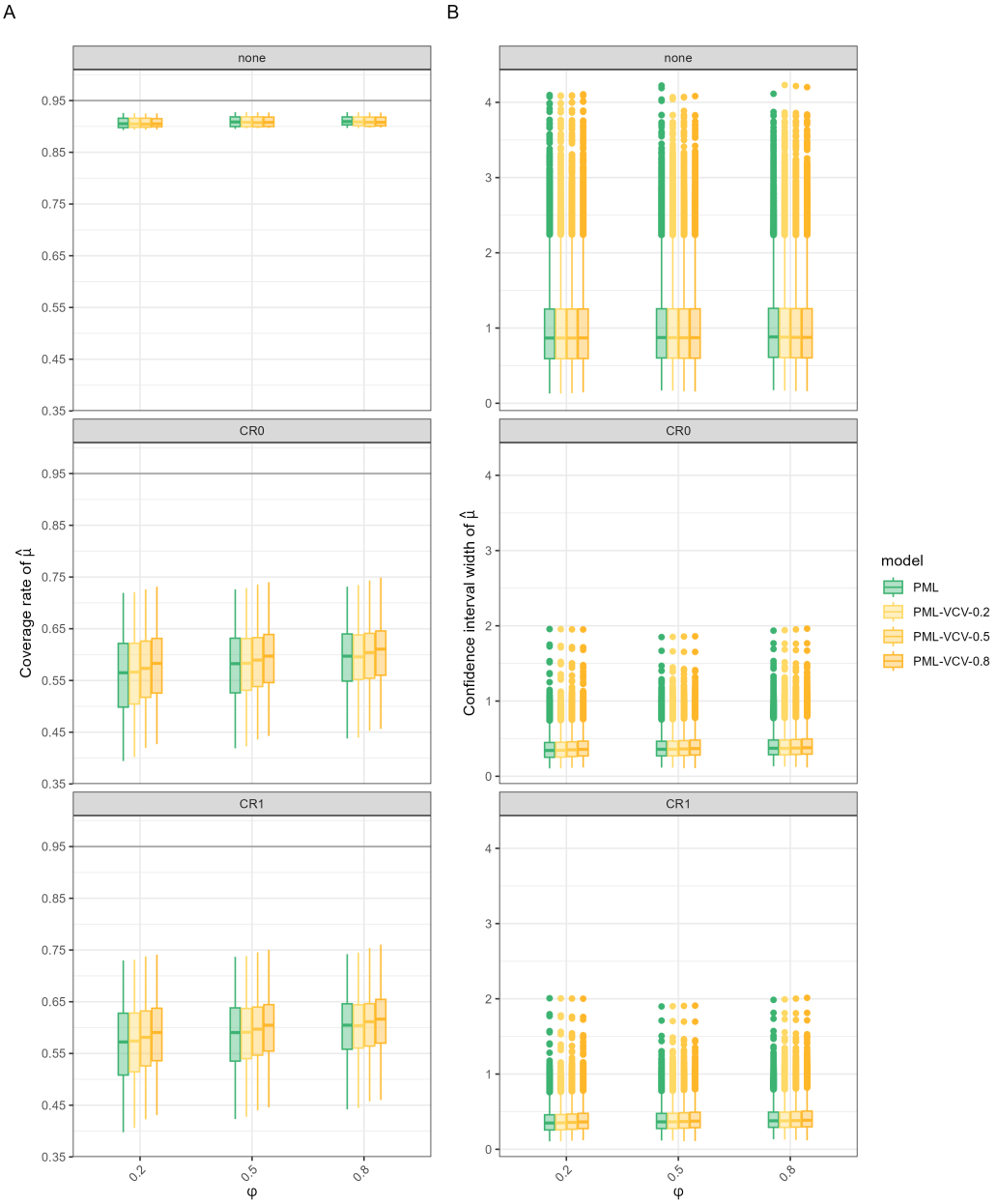

Figure 4: Boxplots of the overall mean estimate $\hat{\mu}$ coverage rate and confidence intervals for each CRVE method under four phylogenetic meta-analysis (PML) working models across all conditions, assessed over 5,000 simulation iterations. **A.** The coverage rates of the 95% confidence intervals, indicating the proportion of intervals that include the true mean μ and assessing the reliability and consistency of the interval estimates **B.** The widths of the confidence intervals. The results were evaluated across 5,000 simulation iterations, eight conditions of variance components (σ_u^2, σ_s^2) and the number of studies (k_{studies}).

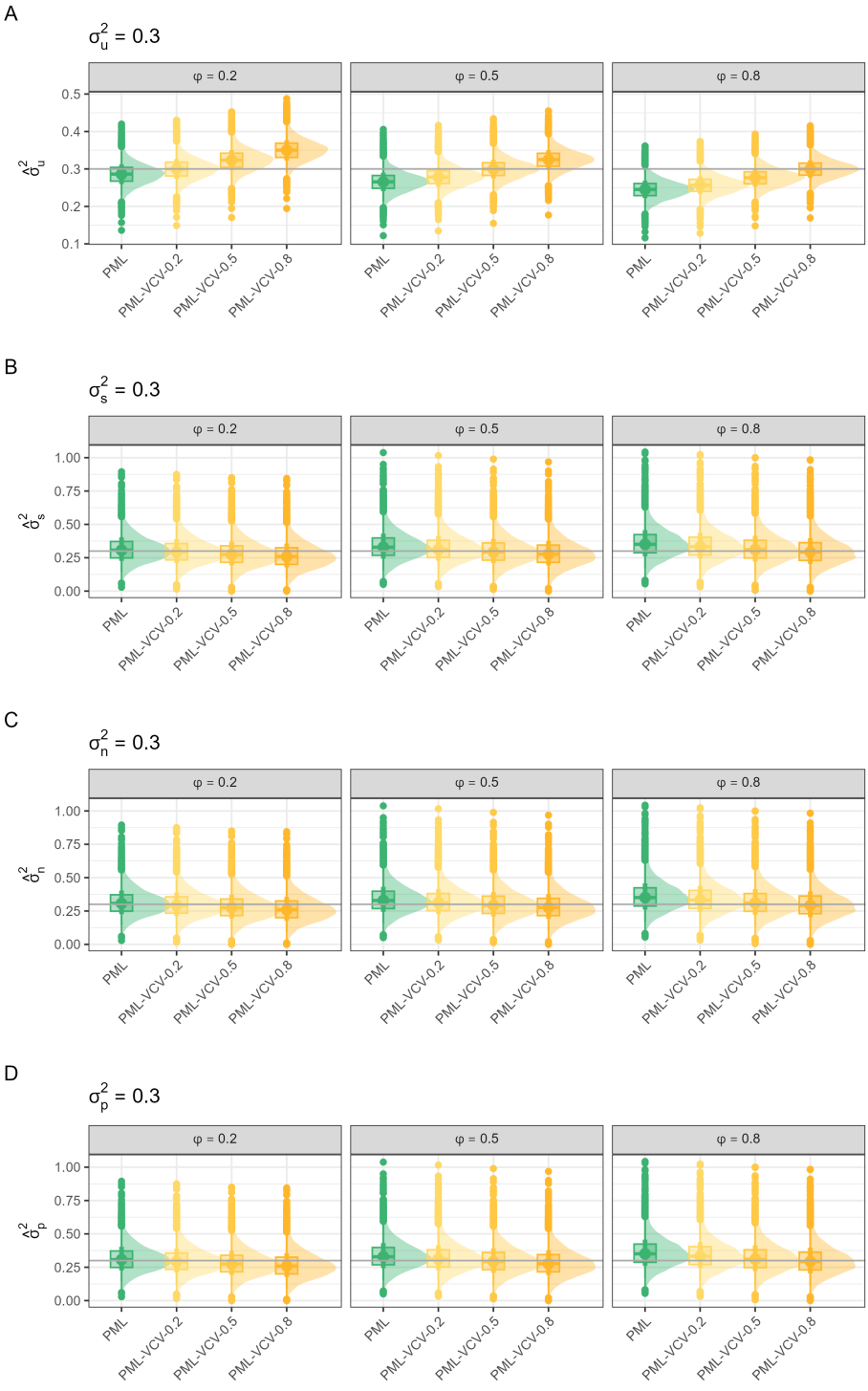

Figure 5: **A.** Boxplots of within-study conditional variance estimates ($\hat{\sigma}^2_u$). **B.** Boxplots of among study conditional variance estimates ($\hat{\sigma}^2_s$). **C.** Boxplots of non-phylogenetic effect conditional variance estimates ($\hat{\sigma}^2_n$). **D.** Boxplots of phylogenetic effect conditional variance estimates ($\hat{\sigma}^2_p$). For all panels, the true variance is shown in the grey bolded line and the boxplot represent the variability of estimates across 5,000 simulations across true within study correlation levels of $\phi \in 0.2, 0.5, 0.8$ and under true values of $\sigma^2_u = \sigma^2_s = \sigma^2_n = \sigma^2_p = 0.3$.

3.2.2 Phylogenetic multilevel meta-regression

For the phylogenetic meta-regression model, the estimates of the four coefficients ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$) were unbiased and did not vary across models with different within-study correlations (see Supporting Figures S12–S15 and Table S9). The 95% confidence interval widths for all coefficients estimates were similarly unaffected even under model misspecification. However, we note slightly narrower widths for the effect size level coefficient $\hat{\beta}_3$ when the model is specified under the true data-generating mechanism of the within-study correlation (Supporting Figure S15). We found coverage rates of the estimates of the moderator coefficients at study level β_1, species level β_2 and effect size level β_3 were approximately at the nominal 95% (see Figures S13, S14, S15). The estimates of the intercept of the meta-regression β_0 showed slightly lower coverage rates around 93% (Figure S12). The fixed effect coefficients in the phylogenetic meta-regression model had worse coverage rates when using CR0 and CR1 cluster robust variance estimation methods (see Figures S16–S19). The majority of models converged (at least 99.99% of models) and were computed within 6 seconds (Supporting Table S3).

3.3 Case studies

We reanalyse two published meta-analyses to illustrate the application of these working models. The models have been simplified from the original studies, so the results are for illustration purpose only and should not be used to draw substantive conclusions. The first case study covers multilevel meta-analysis models which we dealt with in simulation Study 1, while the second focuses on the phylogenetic multilevel meta-analysis models that we conducted for simulation Study 2. Code to run the case studies is provided here.

3.3.1 Case study 1: Multilevel meta-analysis

Crawford et al., 2019 used a large meta-analysis dataset of pairwise plant-soil feedback measures to investigate whether these feedbacks contribute to plant species coexistence. We reanalysed their dataset, focusing on the mycorrhizal having different status consisting of 59 effect sizes across 13 studies. We applied the multilevel meta-analytical models (Equations 3, 4) to account for dependence among effect sizes. For dependence among sampling errors, we assumed a $\mathbf{V}^*$ matrix with a constant within-study correlation, ρ, considering values from 0.1 to 0.9 as well as the case of no correlation (i.e. $\rho = 0$). We also calculated the cluster robust CR2 standard error and P-values for each model. Assuming a higher within-study correlation ($\rho = 0.9$) resulted in a slightly higher log likelihood. The overall mean estimate was near zero and varied little, compared to its standard error, as ρ was changed (although it did change sign at $\rho < 0.5$). The standard errors and P-values did not show any substantial differences as ρ changed or as we moved across to the robust CR2 method. However, we found that the heterogeneity estimates ($\hat{\sigma}^2_u$ and $\hat{\sigma}^2_s$) varied with different assumed correlations.

Table 2: Results of the multilevel meta-analysis working models on the case study 1 dataset. The first column shows the assumed constant correlation among effect sizes from the same study (ρ). The subsequent columns report the estimated overall mean ($\hat{\mu}$), its standard error ($SE[\hat{\mu}]$), the robust CR2 standard error (SE_{CR2}), the P-value (P) (under a t-distribution) and the robust CR2 P-value (P_{CR2}) for testing whether the overall mean is zero, followed by the variance component estimates ($\hat{\sigma}^2_s$ and $\hat{\sigma}^2_u$) and the model's log-likelihood.

ρ	$\hat{\mu}$	$SE[\hat{\mu}]$	$SE[\hat{\mu}]_{\text{CR2}}$	P	P_{CR2}	$\hat{\sigma}^2_u$	$\hat{\sigma}^2_s$	LogLik
0.0	-0.04	0.155	0.154	0.7857	0.7852	0.190	0.229	-56.290
0.1	-0.03	0.152	0.152	0.8413	0.8409	0.193	0.211	-55.794
0.2	-0.02	0.151	0.150	0.8895	0.8891	0.198	0.198	-55.384
0.3	-0.01	0.150	0.149	0.9299	0.9296	0.204	0.187	-55.043
0.4	-0.01	0.150	0.149	0.9628	0.9626	0.211	0.178	-54.757
0.5	0.00	0.150	0.149	0.9887	0.9886	0.219	0.170	-54.517
0.6	0.00	0.150	0.148	0.9918	0.9917	0.229	0.163	-54.316
0.7	0.00	0.150	0.148	0.9783	0.9781	0.239	0.156	-54.151
0.8	0.01	0.150	0.149	0.9705	0.9703	0.251	0.150	-54.020
0.9	0.01	0.150	0.149	0.9682	0.9679	0.264	0.143	-53.923

3.3.2 Case study 2: Phylogenetic multilevel meta-analysis

Horváth et al., 2023 investigated whether behavioural type (mean behaviour) and behavioural predictability (within-individual variation) evolve independently or under system-specific constraints across multiple species. We reanalysed the dataset using phylogenetic multilevel meta-analysis (Equation 5), applying different within-study correlations for effect sizes from the same studies and obtaining CR1 robust standard errors and significance tests. The working model had slightly higher log-likelihoods when no within-study correlation was assumed ($\rho = 0$), but only by a decimal point. The overall mean, standard error, P-value, and variance components ($\hat{\sigma}^2_u$, $\hat{\sigma}^2_s$, $\hat{\sigma}^2_p$, and $\hat{\sigma}^2_n$) remained largely unchanged within two to three decimal places. We note that the CR1 robust standard errors and P-values were substantially smaller than without applying CR1 (the **CR2** method was not applied for the PML as it can't handle cross random effects).

Table 3: Results of the phylogenetic multilevel meta-analysis on the case study 2 dataset. The first column shows the assumed correlation among effect sizes from the same study (ρ). The subsequent columns report the estimated overall mean ($\hat{\mu}$), its standard error ($SE[\hat{\mu}]$), the robust CR1 standard error (SE_{CR1}), the P-value (P) (under a t-distribution) and the robust CR1 P-value (P_{CR1}) for testing whether the overall mean is zero, followed by the variance component estimates ($\hat{\sigma}^2_s$, $\hat{\sigma}^2_u$, $\hat{\sigma}^2_p$ and , $\hat{\sigma}^2_n$) and the model's log-likelihood.

ρ	$\hat{\mu}$	$SE[\hat{\mu}]$	$SE[\hat{\mu}]_{CR1}$	P	P_{CR1}	$\hat{\sigma}^2_u$	$\hat{\sigma}^2_s$	$\hat{\sigma}^2_p$	$\hat{\sigma}^2_n$	LogLik
0.0	-0.05	0.207	0.083	0.7953	0.5199	0.133	0.381	0.115	<0.001	-102.696
0.1	-0.05	0.207	0.083	0.7946	0.5183	0.132	0.382	0.115	<0.001	-102.703
0.2	-0.05	0.207	0.083	0.7940	0.5168	0.130	0.384	0.115	<0.001	-102.710
0.3	-0.05	0.207	0.083	0.7933	0.5153	0.129	0.385	0.116	<0.001	-102.718
0.4	-0.05	0.207	0.083	0.7927	0.5138	0.128	0.386	0.116	<0.001	-102.725
0.5	-0.06	0.208	0.084	0.7920	0.5122	0.127	0.388	0.116	<0.001	-102.733
0.6	-0.06	0.208	0.084	0.7914	0.5107	0.126	0.389	0.117	<0.001	-102.741
0.7	-0.06	0.208	0.084	0.7908	0.5093	0.125	0.390	0.117	<0.001	-102.749
0.8	-0.06	0.208	0.084	0.7901	0.5078	0.124	0.391	0.117	<0.001	-102.757
0.9	-0.06	0.208	0.084	0.7895	0.5063	0.123	0.393	0.117	<0.001	-102.765

4 Discussion

Here, using two extensive simulation studies, we evaluated modelling approaches, including combined methods proposed by Pustejovsky and Tipton (2022), to account for dependence in ecological and evolutionary meta-analytic data. Our simulations are the first to evaluate these combined approaches in an unbalanced design (varying number of effect size per study) and in the context of phylogenetic multispecies meta-analytical data. Our results suggest that multilevel models performed best, given our simulation settings. Additionally, constructing a sampling error variance-covariance matrix ($\mathbf{V}^*$) to account for correlated sampling errors within-studies improved the accuracy of heterogeneity (variance component) estimates. However, neither combining multilevel models with cluster robust variance estimation (CRVE) nor incorporating within-study correlation in sampling error ($\mathbf{V}^*$) improved regression coefficient estimates. We discuss these findings in detail below.

4.1 Regression coefficient estimates

Our simulation results showed that multilevel models provided unbiased and efficient estimates of the overall mean regardless of the specified sampling error dependence structure (Figure 1 and Figure S7). Similar results were also found in the simulations by Moeyaert et al. (2017). Importantly, the inference method and the choice of degrees of freedom in the test statistics and confidence intervals noticeably influenced the coverage rate of the overall mean estimate (to control for Type I error rates), as shown in Figure S2, which was also found in Nakagawa et al. (2022). Further, as expected we found simplistic models (fixed-effects, FE, or random-effects, RE) led to lower coverage rates (increase Type I errors) when there was dependence among effect sizes and sampling errors. However, our results showed that combining simplistic models with CRVE methods improved statistical inference. We highlight below in more details in what context such simplistic models may be of interest even under complex dependence structures.

Regarding multilevel phylogenetic meta-analyses (Study 2), our simulation results found the overall mean estimates were unbiased across all models with and without a specified sampling ($\mathbf{V}^*$) matrix. However, the overall mean had a low coverage rate around 90% for all models, which was also found in the simulations by Cinar et al., 2022. For the phylogenetic multilevel meta-regression models, we found that the estimates of three moderator coefficients were unbiased and precise. Specifically, the effect size level coefficient estimate was slightly more precise under the true model specifications of the sampling error similar to simulation results

by Pustejovsky and Tipton (2022), although the improvement was too small to affect the inference. Further, our results showed the coverage rates of the three moderator coefficients were close to the nominal 95%. However, the estimate of the intercept coefficient showed lower coverage, around 93%. The lower coverage rates for the overall mean and meta-regression intercept estimates could potentially be recovered by using adjusted degrees of freedom (e.g. Satterthwaite method) although such adjustments are not implemented currently in `metafor` under version 4.6-0 (via `clubsandwich`) for models with crossed random effects.

4.2 Variance component estimates

When we assumed a sampling error matrix $\mathbf{V}^*$ that matched the true underlying data-generating mechanisms the multilevel meta-analysis models in both simulation studies provided unbiased estimates of the within and among study variance components. Our findings align with other simulation studies (Fernández-Castilla et al., 2019; Pustejovsky & Tipton, 2022). Further, we found that assuming a higher ρ than the true within-study correlation inflates the within-study variance component, while assuming a lower ρ underestimates it. Although model misspecification does not affect the total variance estimate of the model, it redistributes the variance components, leading to bias variance components. Similar variance redistribution under misspecification has been reported in mixed-effects models (Schielzeth et al., 2020). Modelling accurate variance components is an important part of meta-analysis as it helps distinguish within and among studies variances (Senior et al., 2016). For example, it allows researchers to assess whether an overall mean effect applies across diverse study contexts and to quantify either there is higher variability within or among studies (Yang et al., 2023, 2025). We note that the CRVE methods did not impact the estimation of variance components. The results from Case Study 1 showed that assuming a sampling error $\mathbf{V^*}$ matrix with a higher within-study constant correlation provided better model fit (Table 2). However, in practice, the analyst may not know the true correlations among effect sizes, as described earlier in Section 2. To select the most appropriate correlation structure, researchers can use model fit criteria (e.g., log-likelihood or information criteria) as recommended in Barnett et al., 2010 and as demonstrated in our two case studies. A further issue remains when it is unknown whether correlations among effect sizes are constant or non-constant within and across studies. In such cases, researchers either have to make arbitrary assumptions about these correlations or, if information about another hierarchical level (e.g., different cohorts or samples within studies) is available from primary studies, incorporate this as an additional random effect to avoid assuming a specific $\mathbf{V}^*$ matrix. Yet, such an additional random effect is often unlikely to be distinguishable from the between study effect (or it could lead to non-singularity, for example, if there is only 32 cohorts from 30 studies).

4.3 CRVE methods

We found, interestingly, no substantial benefit in using CRVE methods combined with multilevel modelling even when the model was misspecified. CRVE methods inflate standard errors when samples are small or assumptions are violated, leading to greater uncertainty compared to large samples without violations. However, as discussed above, if the model specifies multi-way clusters (i.e. cross-random effects), the CRVE methods do not work (at least currently). Notably, when CRVE methods are applied to phylogenetic multilevel meta-analysis models it yielded lower coverage rates (increase risk in Type I errors). Further, in our case study 2 we found substantially smaller standard errors and P-values when the cluster robust method was applied, which could lead to incorrect inference (i.e. inflated type I error). Therefore, the current implementation of CRVE methods should not be used for models with crossed random effects, which are common in ecology and evolution (e.g., species, geographical location, experimental method). This is because the current CRVE methods cannot account for cross-classified dependence. When using study-level clustering, CRVE methods assume that estimates from different studies are independent. However, in a model that includes for example species-level random effects (e.g. phylogenetic and non-phylogenetic), there is dependence across studies and ignoring it can lead to underestimated standard errors. Current statistical implementations are limited to support robust variance estimation for multi-way clustered data. There have been methods developed by Cameron et al., 2011 to deal with multi-way clustered standard errors, but these only apply to ordinary least squares models. Currently, the `clubSandwich` does not compute robust estimates when cross-random effects or known correlation matrix for the random effects (*i.e.* the matrix for phylogenetic relationships) are present, which will result in an error. Whereas, `metafor` will compute an estimate for CR0 and CR1 methods when there are crossed-random effects under the current version 4.6-0, which leaves the analyst to interpret whether the results are valid.

4.4 Recommendations

Based on our findings, we recommend the use of multilevel models with adjusted degrees of freedom, and when necessary a constructed sampling error variance-covariance $\mathbf{V}^*$ matrix as the standard approach for ecological and evolutionary meta-analyses. This approach ensures accurate coverage rates and accounts for sampling error dependencies, leading to reliable variance component estimates. We note two important considerations that should guide any meta-analytical model specification. First, carefully select the variables that adequately capture heterogeneity at each hierarchical level, define the hierarchical structure, and decide

whether certain factors should be treated as random or fixed effects (Gelman, 2005). Importantly, always include a random effect at the level of individual effect sizes (*i.e.* modelling the within-study effect), as it accounts for within-study variability and avoids assuming a common true effect. We recommend following a systematic model selection process as described in the decision tree in Pustejovsky and Tipton, 2022. Further consider preregistering this process of model selection, which does not need to include model detail but rather the model selection process, to enhance transparency and reproducibility (Head et al., 2015). Second, use all the information from primary studies. Ideally, the sampling error $\mathbf{V}^*$ matrix should be constructed using this information. However, if there are insufficient data to calculate covariances or to model an additional hierarchical level, using model selection criteria, as in our case studies, can help guide its specification.

4.5 Limitations of study

It is important to note that our findings are limited by the assumptions of the data-generating model and the choice of parameter values in our simulation studies. Although we considered a range of values reflecting ecological and evolutionary meta-analytical data, we did not capture other possible conditions encountered in meta-analysis. This is because these other conditions are less relevant to our main aims. For example, we did not account for varying within-study correlations among effect sizes (*i.e.* non-constant correlations). The consequences of varying within-study correlations and the combination of using known values and arbitrary assumptions has not been investigated in our simulations. Also, we did not evaluate the impact of publication bias (selective reporting of positive findings), a well-documented issue in meta-analysis (Marks-Anglin et al., 2020). Publication bias can distort meta-analytical datasets, leading to biased parameter estimates and inference. Multilevel models, in particular, may overestimate the overall mean effect, as they weigh studies more equally. In contrast, simpler models, such as fixed-effect models (FE), are less sensitive to publication bias but tend to underestimate standard errors, increasing Type I error rates. Approaches to address this suggest combining simpler models that have a sampling error matrix ($\mathbf{V}^*$) with cluster-robust variance estimation (CRVE), which, as our simulation results demonstrate, yields precise and unbiased estimates of the overall mean (Yang et al., 2024). However, further simulation research is needed to confirm their effectiveness as well as applications to real datasets.

5 Conclusions

Dependence among effect sizes and sampling errors in meta-analytical datasets can lead to inaccurate inferences, significantly impacting the conclusions of meta-analyses. Although modern statistical methods that account for this dependence have emerged recently, they remain underutilised in ecology and evolution. Here we recommended specific modelling strategies for ecological and evolutionary meta-analyses to ensure accurate estimation of variance components and reliable coverage of overall mean estimates. Specifically, we advocate the use of multilevel models to explicitly account for heterogeneity at every relevant hierarchical level, use advised inference methods, and incorporate a sampling error variance-covariance matrix using any known values of correlations amongst effect sizes from primary studies to obtain accurate variance component estimates.

References

Arnqvist, G., & Wooster, D. (1995). Meta-analysis: Synthesizing research findings in ecology and evolution. *Trends in Ecology & Evolution*, *10*(6), 236–240. https://doi.org/10.1016/S0169-5347(00)89073-4

Barnett, A. G., Koper, N., Dobson, A. J., Schmiegelow, F., & Manseau, M. (2010). Using information criteria to select the correct variance–covariance structure for longitudinal data in ecology. *Methods in Ecology and Evolution*, *1*(1), 15–24. https://doi.org/10.1111/j.2041-210X.2009.00009.x

Becker, B. J. (2000). 17 - Multivariate Meta-analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 499–525). Academic Press. Retrieved April 12, 2024, from 10.1016/B978-012691360-6/50018-5

Bell, R. M., & McCaffry, D. F. (2002). Bias reduction in standard errors for linear and generalized linear models with multi-stage samples. *Survey Methodology*, *28*(2), 169–181.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249. Retrieved March 8, 2024, from https://www.jstor.org/stable/25800796

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Chamberlain, S. A., Hovick, S. M., Dibble, C. J., Rasmussen, N. L., Van Allen, B. G., Maitner, B. S., Ahern, J. R., Bell-Dereske, L. P., Roy, C. L., Meza-Lopez, M., Carrillo, J., Siemann, E., Lajeunesse, M. J., & Whitney, K. D. (2012). Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. *Ecology Letters*, *15*(6), 627–636. https://doi.org/10.1111/j.1461-0248.2012.01776.x

Cinar, O., Nakagawa, S., & Viechtbauer, W. (2022). Phylogenetic multilevel meta-analysis: A simulation study on the importance of modelling the phylogeny. *Methods in Ecology and Evolution*, *13*(2), 383–395. https://doi.org/10.1111/2041-210X.13760

Crawford, K. M., Bauer, J. T., Comita, L. S., Eppinga, M. B., Johnson, D. J., Mangan, S. A., Queenborough, S. A., Strand, A. E., Suding, K. N., Umbanhowar, J., & Bever, J. D. (2019). When and where plant-soil feedback may promote plant coexistence: A meta-analysis. *Ecology Letters*, *22*(8), 1274–1284. https://doi.org/10.1111/ele.13278

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

Fisher, Z., Tipton, E., & Zhipeng, H. (2023). Robumeta: Robust Variance Meta-Regression. https://cran.r-project.org/web/packages/robumeta/index.html

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53. https://doi.org/10.1214/009053604000001048

Gurevitch, J., & Hedges, L. V. (1999). Statistical Issues in Ecological Meta-Analyses. *Ecology*, *80*(4), 1142–1149. https://doi.org/10.1890/0012-9658(1999)080[1142:SIIEMA]2.0.CO;2

Haddaway, N. R., & Pullin, A. S. (2014). The Policy Role of Systematic Reviews: Past, Present and Future. *Springer Science Reviews*, *2*(1), 179–183. https://doi.org/10.1007/s40362-014-0023-1

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Horváth, G., Garamszegi, L. Z., & Herczeg, G. (2023). Phylogenetic meta-analysis reveals system-specific behavioural type–behavioural predictability correlations. *Royal Society Open Science*, *10*(9), 230303. https://doi.org/10.1098/rsos.230303

Joshi, M., Pustejovsky, J. E., & Beretvas, S. N. (2022). Cluster wild bootstrapping to handle dependent effect sizes in meta-analysis with a small number of studies. *Research Synthesis Methods*, *13*(4), 457–477. https://doi.org/10.1002/jrsm.1554

Koricheva, J., & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology*, *102*(4), 828–844. https://doi.org/10.1111/1365-2745.12224

Lajeunesse, M. J. (2009). Meta-Analysis and the Comparative Phylogenetic Method. *The American Naturalist*, *174*(3), 369–381. https://doi.org/10.1086/603628

Lajeunesse, M. J. (2011). On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology*, *92*(11), 2049–2055. https://doi.org/10.1890/11-0423.1

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Marks-Anglin, A., Duan, R., Chen, Y., Panagiotou, O., & Schmid, C. H. (2020). Publication and Outcome Reporting Bias. In *Handbook of Meta-Analysis*. Chapman; Hall/CRC.

Maynard, R. (2024). Improving the Usefulness and Use of Meta-Analysis to Inform Policy and Practice. *Evaluation Review*, *48*(3), 515–543. https://doi.org/10.1177/0193841X241229885

Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, *20*(6), 559–572. https://doi.org/10.1080/13645579.2016.1252189

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Nakagawa, S., & Poulin, R. (2012). Meta-analytic insights into evolutionary ecology: An introduction and synthesis. *Evolutionary Ecology*, *26*(5), 1085–1099. https://doi.org/10.1007/s10682-012-9593-z

Nakagawa, S., & Santos, E. S. A. (2012). Methodological issues and advances in biological meta-analysis. *Evol. Ecol.*, *26*(5), 1253–1274. https://doi.org/10.1007/s10682-012-9555-5

Nakagawa, S., Senior, A. M., Viechtbauer, W., & Noble, D. W. A. (2022). An assessment of statistical methods for nonindependent data in ecological meta-analyses: Comment. *Ecology*, *103*(1), e03490. https://doi.org/10.1002/ecy.3490

Nakagawa, S., Yang, Y., Macartney, E. L., Spake, R., & Lagisz, M. (2023). Quantitative evidence synthesis: A practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environmental Evidence*, *12*(1), 8. https://doi.org/10.1186/s13750-023-00301-6

Noble, D. W. A., Lagisz, M., O'dea, R. E., & Nakagawa, S. (2017). Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses. *Mol. Ecol.*, *26*(9), 2410–2425. https://doi.org/10.1111/mec.14031

Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claramunt, S., Claude, J., Cuong, H. S., Desper, R., Didier, G., Durand, B., Dutheil, J., Ewing, R. J., Gascuel, O., Guillerme, T., Heibl, C., Ives, A., Jones, B., Krah, F., Lawson, D., … De Vienne, D. (2023). Ape: Analyses of Phylogenetics and Evolution. https://cran.r-project.org/web/packages/ape/index.html

Pastor, D. A., & Lazowski, R. A. (2018). On the Multilevel Nature of Meta-Analysis: A Tutorial, Comparison of Software Programs, and Discussion of Analytic Choices. *Multivariate Behavioral Research*, *53*(1), 74–89. https://doi.org/10.1080/00273171.2017.1365684

Pustejovsky, J. E. (2023). clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. https://cran.r-project.org/web/packages/clubSandwich/index.html

Pustejovsky, J. E., & Chen, M. (2024). Equivalencies Between Ad Hoc Strategies and Multivariate Models for Meta-Analysis of Dependent Effect Sizes. *Journal of Educational and Behavioral Statistics*, 10769986241232524. https://doi.org/10.3102/10769986241232524

Pustejovsky, J. E., & Tipton, E. (2018). Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business & Economic Statistics*, *36*(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prevention Science*, *23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

R-Core-Team. (2022). R: A language and environment for statistical computing. https://www.R-project.org/

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

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

Sidik, K., & Jonkman, J. N. (2005). A Note on Variance Estimation in Random Effects Meta-Regression. *Journal of Biopharmaceutical Statistics*, *15*(5), 823–838. https://doi.org/10.1081/BIP-200067915

Siepe, B. S., Bartoš, F., Morris, T., Boulesteix, A.-L., Heck, D. W., & Pawel, S. (2023, October). Simulation Studies for Methodological Research in Psychology: A Standardized Template for Planning, Preregistration, and Reporting. https://doi.org/10.31234/osf.io/ufgy6

Stewart, G. (2009). Meta-analysis in applied ecology. *Biology Letters*, *6*(1), 78–81. https://doi.org/10.1098/rsbl.2009.0546

Tipton, E., & Pustejovsky, J. E. (2015). Small-Sample Adjustments for Tests of Moderators and Model Fit Using Robust Variance Estimation in Meta-Regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*(2), 180–194. https://doi.org/10.1002/jrsm.1339

UNSW. (2024). Katana. https://doi.org/10.26190/669x-a286

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Van Den Noortgate, W., & Onghena, P. (2003). Multilevel Meta-Analysis: A Comparison with Traditional Meta-Analytical Procedures. *Educational and Psychological Measurement*, *63*(5), 765–790. https://doi.org/10.1177/0013164403251027

Viechtbauer, W. (2023). Metafor: Meta-Analysis Package for R. https://cran.r-project.org/web/packages/metafor/index.html

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, *20*(3), 360–374. https://doi.org/10.1037/met0000023

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Williams, C., Yang, Y., Lagisz, M., Morrison, K., Ricolfi, L., Warton, D. I., & Nakagawa, S. (2024). Transparent reporting items for simulation studies evaluating statistical methods: Foundations for reproducibility and reliability. *Methods in Ecology and Evolution*, *15*(11), 1926–1939. https://doi.org/10.1111/2041-210X.14415

Yang, Y., Lagisz, M., Williams, C., Noble, D. W. A., Pan, J., & Nakagawa, S. (2024). Robust point and variance estimation for meta-analyses with selective reporting and dependent effect sizes. *Methods in Ecology and Evolution*, *15*(9), 1593–1610. https://doi.org/10.1111/2041-210X.14377

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

Yang, Y., Noble, D. W., Senior, A. M., Lagisz, M., & Nakagawa, S. (2025). Interpreting prediction intervals and distributions for decoding biological generality in meta-analyses. *eLife*, *14*. https://doi.org/10.7554/eLife.103339.1