Methods in Ecology and Evolution

DR PAULA PAPPALARDO (Orcid ID : 0000-0003-0853-7681)

DR KIONA OGLE (Orcid ID : 0000-0002-0652-8397)

Article type : Research Article

Editor : Professor Robert B. O'Hara

TITLE: Comparing traditional and Bayesian approaches to ecological meta-analysis

RUNNING TITLE: Comparing methods for ecological meta-analysis

AUTHORS: P. Pappalardo[1*], K. Ogle[2,3], E.A. Hamman[1], J.R. Bence[5], B.A. Hungate[3], and C.W. Osenberg[1]

ORCID IDs:

Paula Pappalardo: 0000-0003-0853-7681

Kiona Ogle: 0000-0002-0652-8397

Elizabeth A. Hamman: 0000-0002-3494-6641

James R. Bence: 0000-0002-2534-688X

Bruce A. Hungate: 0000-0002-7337-1887

Craig W. Osenberg: 0000-0003-1918-7904

AFFILIATIONS:

[1]Odum School of Ecology, University of Georgia, Athens, GA 30602, USA

[2]School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ 86011, USA

[3]Center for Ecosystem Science and Society and Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ 86011, USA

[5]Quantitative Fisheries Center, Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI 48824, USA

* Corresponding author contact information: Department of Invertebrate Zoology, Smithsonian National Museum of Natural History, Washington, DC 20560, USA, email: paulapappalardo@gmail.com, phone: 706-308-2979

PAPER TYPE: Research article

KEY WORDS: bias, confidence interval, coverage, credible interval, effect size, global climate change, log response ratio, sample size

ABSTRACT

1. Despite the wide application of meta-analysis in ecology, some of the traditional methods used for meta-analysis may not perform well given the type of data characteristic of ecological meta-analyses.
2. We reviewed published meta-analyses on the ecological impacts of global climate change, evaluating the number of replicates used in the primary studies (n_i) and the number of studies or records (k) that were aggregated to calculate a mean effect size. We used the results of the review in a simulation experiment to assess the performance of conventional frequentist and Bayesian meta-analysis methods for estimating a mean effect size and its uncertainty interval.
3. Our literature review showed that n_i and k were highly variable, distributions were right-skewed, and were generally small (median n_i =5, median k=44). Our simulations show that the choice of method for calculating uncertainty intervals was critical for obtaining appropriate coverage (close to the nominal value of 0.95). When k was low (<40), 95% coverage was achieved by a confidence interval based on the t-distribution that uses an adjusted standard error (the Hartung-Knapp-Sidik-Jonkman, HKSJ), or by a Bayesian credible interval, whereas bootstrap or z-distribution confidence intervals had lower coverage. Despite the importance of the method to calculate the uncertainty interval, 39% of the meta-analyses reviewed did not report the method used, and of the 61% that did, 94% used a potentially problematic method, which may be a consequence of software defaults.
4. In general, for a simple random-effects meta-analysis, the performance of the best frequentist and Bayesian methods were similar for the same combinations of factors (k and mean replication), though the Bayesian approach had higher than nominal (>95%) coverage for the mean effect when k was very low (k<15). Our literature review suggests that many meta-analyses that used z-distribution or bootstrapping confidence intervals may have over-estimated the statistical significance of their results when the number of studies was low; more appropriate methods need to be adopted in ecological meta-analyses.

RESUMEN

1. A pesar del uso generalizado del meta-análisis en el área de Ecología, algunos de los métodos de análisis tradicionalmente utilizados pueden dar resultados no ideales dado el tipo de datos que los caracteriza.
2. En este trabajo se realizó una revisión de los meta-análisis publicados sobre los impactos ecológicos del cambio climático global, evaluando el número de réplicas utilizadas en las publicaciones originales (n_i) y el número de estudios o registros (*k*) que fueron agrupados para calcular un tamaño de efecto promedio. Se utilizaron los resultados de la revisión en un experimento de simulación para evaluar el desempeño de métodos frecuentistas convencionales y métodos Bayesianos para estimar un tamaño de efecto promedio y su intervalo de incertidumbre.
3. La revisión de la literatura demostró que n_i y *k* fueron muy variables, con distribuciones sesgadas, y con valores en general bajos (mediana n_i =5, mediana *k*=44). Nuestras simulaciones muestran que la elección del método para calcular un intervalo de incertidumbre fue crítica para obtener una cobertura apropiada (alrededor del valor nominal de 0.95). Cuando *k* fue bajo (<40), obtuvimos una cobertura de 95% utilizando un intervalo de confianza basado en la distribución t de student que usa un ajuste por el error estándar (llamada Hartung-Knapp-Sidik-Jonkman, HKSJ), o utilizando un intervalo de credibilidad Bayesiano, mientras que los intervalos de remuestreo o con una distribución Normal tuvieron cobertura baja. A pesar de la importancia del método utilizado para calcular el intervalo de incertidumbre, 39% de los meta-análisis revisados no reportaron el método utilizado y, de los 61% que si lo reportaron, 94% usaron uno de los métodos potencialmente problemáticos, lo que puede ser una consecuencia de la configuración por defecto de los programas informáticos utilizados para meta-análisis.
4. En general, para un meta-análisis simple con efectos aleatorios, el desempeño del mejor método frecuentista y el método Bayesiano fueron similares para las mismas combinaciones de factores (*k* y número de réplicas promedio), aunque el método Bayesiano tuvo cobertura mayor de la nominal (>95%) para el efecto promedio cuando *k* fue muy bajo (*k*<15). Nuestra revisión sugiere que muchos de los meta-análisis que utilizaron una distribución Normal o intervalos de remuestreo pueden haber sobreestimado la significancia estadística de sus

resultados cuando el número de estudios fue bajo. Otros métodos más apropiados deberían ser usados para meta-análisis en Ecología.

INTRODUCTION

Meta-analysis uses statistical techniques to quantitatively summarize information from different studies and is highly influential in the contemporary practice of science. To conduct a meta-analysis an investigator gathers summary statistics from each study to calculate an effect size, with the goal of computing an overall effect size (and its uncertainty) and exploring the factors contributing to variation in effect sizes (Nakagawa, Noble, Senior, & Lagisz, 2017). The use of meta-analysis in ecology has been growing rapidly since the 1990s, and has proven particularly useful in discerning general patterns by comparing information from different species, study sites, and systems (Cadotte, Mehrkens, & Menge, 2012). Advice on best methodological practices for meta-analysis is widespread in disciplines with a longer history of meta-analytic research (e.g. medical sciences) but is lagging behind in ecology (Gates, 2002). This can be problematic because ecological meta-analyses have specific challenges not necessarily typically in other disciplines.

One pervasive characteristic of ecological meta-analyses is the high heterogeneity (i.e., large among-study variation in effect sizes). Senior et al. (2016) analyzed 86 meta-analyses in ecology and evolution and found that the among-study variation averaged 92% of the total variance. In contrast, a review of 509 meta-analyses in medicine found that there was no detectable among-study variation in 50% of the studies (Higgins, Thompson, & Spiegelhalter, 2009). Ecological studies also differ from many other disciplines in the typical level of within-study replication, which is fewer than 10 replicates per study (Hillebrand & Gurevitch, 2014). Such low levels of replication will influence the precision of the estimates of effect size from the primary studies (Langan, Higgins, & Simmons, 2016). Importantly, the low level of replication typical of ecological studies is outside the range used in most simulation studies designed to assess meta-analytic methods, which typically range from dozens to hundreds (Langan et al., 2016). Thus differences between ecology and other disciplines potentially limit the insights ecologists can gain from existing simulations that compare different meta-analytic methods.

Specific advice for conducting ecological meta-analyses include suggestions on the type of meta-analytic model and effect size calculation to use (Gurevitch & Hedges, 1999; Osenberg, Sarnelle, Cooper, & Holt, 1999; Lajeunesse, 2015), and how to deal with non-independence (Gurevitch & Hedges, 1999; Noble, Lagisz, O'dea, & Nakagawa, 2017; Song, Peacor, Osenberg, &

Bence, 2020). For example, a random-effects model is often recommended for ecological meta-analysis over a fixed-effects model (Gurevitch & Hedges, 1999), and multi-level models are increasingly being used to incorporate the non-independence commonly found in ecological meta-analyses (Nakagawa & Santos, 2012). A topic addressed in the medical literature that has received little attention in ecology (but see Adams, Gurevitch, & Rosenberg, 1997) is the choice of confidence interval (CI) used to estimate the mean effect size in a meta-analysis (Hartung & Knapp, 2001; Sidik & Jonkman, 2003, Sánchez-Meca & Marín-Martínez, 2008).

Simulation studies have shown that when the number of studies (k) in the meta-analysis is low, the CIs for a mean effect size calculated using a normal approximation are too narrow, leading to coverage below the nominal level (i.e., a 95% CI should include the true value 95% of the time) (Brockwell & Gordon, 2001; Sánchez-Meca & Marín-Martínez, 2008). To avoid this problem, meta-analyses in the medical literature often use the HKSJ (Hartung-Knapp-Sidik-Jonkman; Hartung & Knapp, 2001; Sidik & Jonkman, 2003) method, which is based on a t-distribution and can achieve good coverage even when k is small (Inthout, Ioannidis, & Borm, 2014). Bootstrap techniques have been recommended for estimating CIs for means in ecological meta-analyses, due to its robustness to departures from normality (Adams et al., 1997). On the other hand, boot-strapped CIs can lead to poor coverage when estimating the among-study variance (Viechtbauer, 2007).

Bayesian methods, and the credible interval, offer an alternative approach to estimating uncertainty in meta-analyses. Although Bayesian methods may have a steep learning curve, they offer advantages in handling hierarchical models, for incorporating prior information, and for dealing with missing data (Ogle, Barber, & Sartor, 2013). Bayesian meta-analytic techniques produce a posterior distribution of the mean effect size and associated variance terms. Estimates of uncertainty, including credible intervals, can be directly obtained from the posterior distributions, offering an easier to interpret alternative to the frequentist-based CI (Kruschke & Liddell, 2008).

Our main goal is to compare the performance of traditional and Bayesian methods to measure the uncertainty around the estimation of a mean effect in the context of ecological meta-analysis. To achieve this goal, we conducted a two-pronged study. First, we reviewed published ecological meta-analyses to characterize the types of confidence interval used in ecological meta-analyses, the number of replicates used in the primary studies (n_i) included in published meta-analyses, and the number of studies (k) that were aggregated to calculate a mean effect size. Second, we used the n_i and k found in our literature review to inform the range of parameter values to use in conducting simulation experiments relevant to ecological meta-analyses. In particular, we determined the typical levels of n_i, k, and the among-study variance and then varied them systematically in our simulation studies. We then evaluated performance of frequentist and Bayesian meta-analysis methods when applied to the simulated data, especially with respect to their ability to estimate the true mean effect and among-study variance, and their quantification of uncertainty intervals (i.e., confidence or credible intervals). Based on our findings, we generate recommendations on the methods to measure uncertainty that perform best for ecological meta-analysis and highlight how simple choices (sometimes overlooked by the investigators) can affect the results of meta-analyses.

MATERIALS AND METHODS

Literature review to assess characteristics of ecological datasets

Literature search. We searched the Core Collection of the ISI Web of Science database in March 2017; the search string for TOPIC included ([“meta-analy*” OR “metaanaly*” OR “meta analy*”] AND [“climate change” OR “global change”]). We only included articles and reviews within the “Ecology”, “Environmental Sciences”, “Biodiversity Conservation” and “Plant Sciences” categories. The search resulted in 581 citations; the PRISMA diagram detailing the screening process is provided in Figure S1. After abstract screening, we checked the full text of the 205 articles published between 2013 and 2016. Of these, 96 papers satisfied the inclusion criteria for the final analysis.

Criteria for inclusion. We focused on narrow sense meta-analyses: i.e., those that used a quantitative meta-analytic method to combine effect sizes that compared a control and a treatment group. We excluded studies that 1) only cited published meta-analyses, 2) reviewed meta-analytic methods, but did not perform a meta-analysis, 3) were identified as meta-analysis by the authors but did not use a meta-analytic model or did not calculate effect sizes, 4) used the correlation between two variables as an effect size, and 5) were not “biological meta-analyses” (as defined in Nakagawa et al., 2017), such as studies related to human health or human behavior.

Information extracted. For each paper we extracted the number of studies (k) from the text, figure captions, figures, and supplementary materials. Here we define a "study" as yielding an estimate of an effect, so that a given primary paper could generate multiple effects and thus multiple studies. The k values were determined at three levels, 1) overall: i.e., the total k collected by the authors (e.g., if they conducted meta-analyses on different response variables, then we summed the k across these variables); 2) analysis: i.e., the total k used in a particular analysis (e.g., if an analysis examined variation among four levels of a moderator, then we summed up the number of studies in each level); and 3) category: i.e., the k included in each category of a categorical analysis. In some cases, authors calculated mean effect sizes for different categories separately and only compared the categories using confidence intervals (i.e., there was no integrated analysis incorporating a category effect). In this case, we considered each's categories' k to apply at the "analysis" level.

When available, we also recorded the number of replicates (n_i) in the original studies. If the level of replication was unequal for the control and treatment groups, we recorded the average. Finally, from each meta-analysis, we also recorded the inferential paradigm used (frequentist vs. Bayesian) and the method used to obtain confidence intervals for the frequentist approaches (e.g., non-parametric bootstrap, normal-based, KHSJ, etc.).

Simulation Experiments

Our literature review showed that 67% of the reported primary studies had less than ten replicates. In addition, the review of meta-analyses in ecology and evolution by Senior et al. (2012) showed that among-study variation was important, and typically large, in ecological studies. Given these characteristics of ecological data, we simulated data in a full-factorial design that considered the following levels: mean number of replicates n = {3, 5, 10, 15, 20, 30}, number of studies k = {5, 10, 15, 25, 35, 50}, and among-study variance σ^2_{among} = $\{0.1, 0.25, 0.5, 1, 2, 5\}$. We simulated 2,000 replicated meta-analyses for each combination of n, k, and σ^2_{among}. We then evaluated the performance of four meta-analytic methods applied to the simulated data: three frequentist approaches that differed in how they calculated confidence intervals for a mean effect and a Bayesian approach.

Simulating raw data for a study. We first determined the number of replicates for study i (n_i) based on a random draw from a Poisson distribution:

$$n_i^* \sim Poisson(n-2) \qquad \text{(Eq. 1)}$$

$$n_i = n_i^* + 2 \qquad \text{(Eq. 2)}$$

where n is the mean number of replicates representative of ecological meta-analyses. We subtracted 2 to sample from the Poisson and added 2 to the simulated n_i^* to make the minimum number of replicates for each simulated study equal 2 rather than 0. For each study, we assumed equal number of replicates for the control and treatment groups.

Individual observations ($j = 1, 2, \ldots, n_i$) for the control and treatment groups were generated from a lognormal distribution (*LN*) such that for study *i* and observation *j*:

$$y_{C_{ij}} \sim LN(0, \sigma_{rep}^2) \qquad \text{(Eq. 3)}$$

$$y_{T_{ij}} \sim LN(0 + \mu + \varepsilon_i, \sigma_{rep}^2) \qquad \text{(Eq. 4)}$$

where σ_{rep}^2 is the among-replicates variation, μ is the true overall effect, and $y_{C_{ij}}$ and $y_{T_{ij}}$ are the simulated observations for study *i* and observation *j* of the control and treatment group, respectively. We set the among-replicate variation equal to 1 for both the control and treatment. For convenience, we set the location parameter for the control group equal to zero, resulting in median (y_C) = 1. For the treatment group in study *i*, we set median (y_T) $= \mu + \varepsilon_i$, where μ is the overall true treatment effect (hereafter, true effect size) and ε_i is the random effect associated with study *i*. We simulated ε_i as:

$$\varepsilon_i \sim N(0, \sigma_{among}^2) \qquad \text{(Eq. 5)}$$

Thus, the true effect size from any given study departs from μ due to its random effect (determined by ε_i), while the estimated effect size differs from the true effect size due to within-study sampling error (i.e., as influenced by n_i and σ_{rep}^2). The range of values used for σ_{among}^2 were chosen to produce a similar distribution of I^2 (the proportion of variation among effect sizes not explained by sampling error) to that reported by Senior et al. (2016) for meta-analyses in ecology and evolution (I^2 simulation results are presented in Figure S2).

Estimating the effect size and within-study variance. Using the raw data simulated from each study, we computed the observed effect size for study *i* as the log response ratio ($lnRR_i$), which is widely used in ecology (Nakagawa & Santos, 2012) and it is often a reasonable approximation of ecological phenomena (Osenberg, Sarnelle, & Cooper, 1997):

$$lnRR_i = \ln\left(\frac{\bar{y}_{T_i}}{\bar{y}_{C_i}}\right) \qquad (6)$$

where $\bar{y}_{T_i}$ and $\bar{y}_{C_i}$ are the sample means of the treatment and control groups, respectively.

The expected sample means for each treatment in a simulated study are $E\left(y_{C_{ij}}\right) = \exp\left(\frac{\sigma_{rep}^2}{2}\right)$ and E $\left(y_{T_{ij}}\right) = exp\left(\mu + \varepsilon_i + \frac{\sigma_{rep}^2}{2}\right)$. Thus, the log of the ratio of the expected values for the treatment and control groups is $\mu + \ \varepsilon_i$, corresponding to what we call the true study-specific effect size.

We calculated the estimated within-study variance of the log ratio (Eq. 1 in Hedges, Gurevitch, & Curtis, 1999) ($\sigma_{within_i}^2$) as:

$$\sigma_{within_i}^2 = \frac{{SD_{T_i}}^2}{n_{T_i} \cdot \bar{y}_{T_i}^2} + \frac{{SD_{C_i}}^2}{n_{C_i} \cdot \bar{y}_{C_i}^2} \qquad (7)$$

where SD_T and SD_C are the sample standard deviations of the treatment and control groups, respectively, and $n_{T_i} = n_{C_i} = n_i$ are the simulated number of replicates in study i.

Meta-analytic approaches

Given that we simulated independent data to highlight how the choice of uncertainty interval affects the estimation of a mean effect, we used a standard random-effects model (Gurevitch & Hedges, 1999). We comment on how our results may change with a multi-level (hierarchical) model in the Discussion section. We assume the simulated effect size for study *i* (*lnRR*$_i$, calculated from Eq. 6) follows a normal distribution with mean θ_i (the true effect for study i) and within-study variance $\sigma_{within_i}^2$:

$$lnRR_i \sim N(\theta_i, \sigma_{within_i}^2) \qquad (8)$$

$$\theta_i \sim N\left(\mu, \sigma_{among}^2\right) \qquad (9)$$

We assume $\sigma_{within_i}^2$ is known, as calculated via Eq. 7. Likewise, the true study-specific effect size, θ_i, is assumed to follow a normal distribution with mean μ (the true overall effect) and among-study variance, σ_{among}^2 (which is sometimes referred to as τ^2 in other meta-analytic papers).

We compared different methods to construct confidence intervals (CIs) for a mean effect (at the analysis level) within the frequentist methods versus Bayesian credible intervals. For the frequentist-based analyses, we compared: a) a CI based on a *z*-distribution, which is a large sample approximation, b) a weighted CI based on the Hartung-Knapp-Sidik-Jonkman (HKSJ) method, which does not assume a large sample and instead uses a *t*-distribution, and c) bootstrap methods. For the Bayesian-based analysis, we calculated the highest posterior density (HPD) credible interval.

Frequentist approaches. We applied the random-effects model described by Eqs. 8 and 9 with inverse variance weights using the "rma" function in the R package *metafor* (Viechtbauer, 2010), and estimated σ^2_{among} with the default REML method. To calculate the *z*-distribution CI, we used the default settings for the random-effects model in *metafor*, which returns a 95% CI for μ based on the normal distribution. To apply the HKSJ CI, we set the option knha=T in *metafor*. The resulting CI for μ is based on both a refined estimate of σ^2_{among} and a Student's t-distribution (Hartung & Knapp, 2001; Sidik & Jonkman, 2003), which accounts for the fact that σ^2_{among} is estimated and not known. For the bootstrapped CI, we estimated the bias-corrected non-parametric bootstrapped 95% CI for both μ and σ^2_{among} via the *boot* package in R (Canty & Ripley, 2017). Since the choice of HKSJ or z-distribution for the μ CI does not affect the estimation of σ^2_{among}, in both cases we used *metafor*'s function "confint" to obtain the CI for σ^2_{among} ("confint" applies a Q-profile method in combination with REML).

Bayesian approach. We used a "hybrid" Bayesian framework to implement the random-effects model (Eqs. 8 and 9) in which we treat σ^2_{within} as known; whereas a fully Bayesian model may treat σ^2_{within} as unknown (this hybrid model is comparable to the "empirical Bayes" method discussed in Schmid & Mengersen, 2013). Initial explorations with full and hybrid models gave qualitatively similar results and we only include the hybrid model in our analysis.

We specified relatively non-informative priors for the unknown quantities (e.g., μ and σ^2_{among}). For the mean effect size, μ, we specified a conjugate normal prior with a mean of zero and large variance: *N*(0, 10000). Given that even diffuse priors for σ^2_{among} can influence the posterior for σ^2_{among}, particularly under small group size (Gelman, 2006), we explored five different priors for σ^2_{among} (Supporting Information Figures S12-15). For the final analysis, convergence statistics and computational speed led us to focus on the *Uniform*(0,10) prior for the standard deviation (σ_{among}).

The Bayesian meta-analyses were implemented in JAGS with the *rjags* R package (Plummer, 2018). For each model, we ran three parallel Markov chain Monte Carlo (MCMC) sequences for 200,000 iterations, and discarded the first 100,000 iterations as the burn-in period. We used the $\hat{R}$ convergence diagnostic (Gelman & Rubin, 1992) to evaluate convergence of the MCMC sequences to the posterior. For the final simulations, we only included runs that had $\hat{R}$ < 1.1, and checked that the proportion of discarded runs was lower than 1%. Using post-burn-in MCMC samples, we computed posterior means for quantities of interest (e.g., μ and σ^2_{among}) as point estimates. We computed 95% credible intervals as HPD intervals for both μ and σ^2_{among} using the "HPDinterval" function in the *coda* package (Plummer, 2006).

Implementation and Assessment of the Meta-analysis Approaches

We ran all the analyses and simulations in the R environment (R Core Team, 2019); code is provided in the Supporting Information. For each simulated dataset, we estimated μ and σ^2_{among} via the frequentist and Bayesian methods described above. We summarized the results from the 2,000 replicated meta-analyses for each combination of factors (*n*, *k*, σ^2_{among}) and modeling approaches (i.e., frequentist and Bayesian methods to measure uncertainty). The results for the model performance associated with estimating σ^2_{among} are presented in Figures S7-10.

We evaluated model performance using: coverage, width of the uncertainty intervals, bias, and efficiency. We estimated *coverage* for both μ and σ^2_{among} as the proportion (out of the 2,000 simulation replicates) of calculated 95% uncertainty intervals (CIs for the frequentist methods and credible interval for the Bayesian approach) that included the corresponding true value. Ideally, coverage should equal the nominal value of 0.95 (95%). CIs for these "coverage proportions" were computed using the "binom.confint" function in the R *binom* (Sundar, 2014) package, with the method "wilson" (Agresti & Coull, 1998).

We summarized the perceived uncertainty for μ and σ^2_{among}as the mean *width of the 95% uncertainty intervals* for the 2,000 intervals for each scenario, and assessed how well the mean width was estimated using a 95% CI based on a *t*-distribution. All else being equal, smaller uncertainty is a desirable feature, but not if it is accompanied by a reduction in coverage below the nominal level.

To evaluate *bias*, we calculated the mean difference between the point estimates for μ and σ^2_{among} and their true values based on the 2,000 simulation replicates, and report a 95% CI for this estimate based on the *t*-distribution. Ideally, bias should be centered on zero.

Finally, to quantify the *efficiency* of the point estimates, we calculated the root mean squared error (RMSE) between the estimated and true values for μ and σ^2_{among}as:

$$RMSE = \sqrt{\frac{\sum_{s=1}^{N_{sim}} (\hat{a}_s - a_{true_s})^2}{N_{sim}}} \ , \tag{10}$$

where $a = \mu$ or σ^2_{among}, $\hat{a}$ is the point estimate from each model, a_{true} is the true value used in the simulations, and N_{sim} is the number of simulations.

RESULTS

Literature review to assess characteristic of ecological datasets

Of the 96 meta-analyses that satisfied our criteria (Table S1), 95 and 26 provided information on the number of studies (*k*) and number of replicates (n_i) associated with the original dataset, respectively. Only three meta-analyses used a Bayesian approach. The majority of meta-analyses were published in *Global Change Biology* (23), followed by *Agriculture Ecosystems & Environment* (7) and *Ecology* (6) (Figure S3 displays the full list). The quality of reporting varied, and is discussed in more detail in the Supporting Information. We also provide additional information on *k* and n_i (by taxa, environment, and topic) in the Supporting Information (Table S2, Figures S4-S5).

Number of studies. The number of studies (*k*) used to estimate an effect was highly skewed at the three levels we considered: overall, analysis, and category (Figure 1). The overall *k* ranged from 25 to 32,567 (Figure 1A upper panel), with a median of 273 and with relatively few (12%) including more than 1,000 studies. For most papers, however, analyses were performed for different response variables or different moderators, and the *k* used for a particular analysis was considerably lower (Figure 1A middle panel), ranging from *k* = 1 (for a paper that presented all possible comparisons, even when one potential analysis was represented by only a single study) to *k* = 8,474, with a median of *k* = 44 (i.e., 50% of meta-analysis included 44 or fewer studies); 16% had *k* ≤ 10. The number of studies included within categories ranged from $k = 1$ to 1,430, with a median of 16; 36% had $k \leq 10$ (Figure 1A lower panel).

Number of replicates. The distribution of the reported number of replicates in the original studies (n_i) cited by the climate change meta-analyses was highly skewed, ranging from $n_i = 1$ to 21,600, with most studies having only a few replicates; the median was 5 (Figure 1B). The strong skewness in these data led us to inspect some of the original publications from which exceptionally large n_i values were reported. We found publications in which n_i values were likely misreported or greatly inflated by pseudoreplication (details in Table S3 and Figure S6).

Analytic method to estimate the uncertainty interval for a mean effect. In 38.5% of the papers reviewed, the method used to calculate the frequentist-based CI for the mean effect was not mentioned (Figure 2). Of the papers reporting how the CI was calculated, the majority used bootstrapped or *z*-distribution CIs; only three papers used credible intervals (Bayesian method), and a few used a combination of methods (Figure 2). No papers reported using HKSJ method. Of the papers that did not specify the method, nine used Metawin (which defaults to a *t*-distribution for the parametric CI, without the KHSJ refinement); 12 papers used the packages *meta* or *metafor* in R (which default to a *z*-distribution); and two used the Comprehensive Meta-Analysis software (which defaults to a *z*-distribution). Assuming these 23 papers used the software defaults, then 31 papers used a *z*-distribution, and nine used a *t*-distribution but without the KHSJ refinement. Thus, bootstrapped and *z*-distribution CIs likely comprise the vast majority of approaches, with KHSJ CIs being entirely absent from our dataset.

Simulation experiments: estimation of a mean effect

The number of studies, k, used to estimate a mean effect size, μ, substantially affected the coverage of the frequentist methods, but this effect of k depended on the type of method used to estimate the 95% CIs (Figure 3A). For example, *z*-distribution CIs for μ had coverage lower than the nominal level when $k < 40$, and coverage was appreciably lower for $k < 20$ (Figure 3A). Similarly, bootstrapped CIs had lower than nominal coverage when k < 40 (Figure 3A). In contrast, KHSJ CIs had close to nominal coverage over all values of k (Figure 3A). The Bayesian credible interval generally showed coverages around 95%, but when k = 5, coverage was >95% (Figure 3A).

Coverage can be smaller than nominal levels either because of bias or because the width of the uncertainty interval is inappropriately narrow (i.e., uncertainty is underestimated). The three frequentist methods for computing CIs for μ used the same approach for obtaining point estimates and had minimal bias centered on zero (Figures S11 A,C,E). Thus, the observed differences in coverage for μ resulted from differences in the width of the uncertainty interval (Figure 3B). The Bayesian credible interval was generally wider than the frequentist-based CIs, and of the frequentist CIs, the KHSJ CI tended to be the widest; when k was small, the z-distribution and boot-strapped CIs were ~1/3 smaller than they should be based upon the more appropriate KHSJ CI (Fig. 3B).

Increasing the mean number of replicates (n) in the primary studies did not greatly affect coverage (Figure 3B), the width of the uncertainty interval (Figure 3E), bias (Figure S11C), or RMSE (Figure S11D) for μ . Our results were likely produced because the among-study variation dominated within-study variation over the range of levels considered for the simulation factors (as determined by the review by Senior et al., 2016).

Increasing the among-study variance (σ^2_{among}) increased the width of the uncertainty interval for μ (Figure 3F), but had only small effects on coverage (Figure 3C). Bias in the estimation of μ was negligible and unaffected by an increase in σ^2_{among} (Figure S11E), but the error in the estimation increased with the increase in heterogeneity (RMSE, Figure S11F).

DISCUSSION

Our literature review shows that ecological meta-analyses are highly variable in terms of how many studies (k) are included in the meta-analysis and the number of replicates reported in the original publications (n_i). Despite this high variability, both across and within meta-analyses, k and n_i tend to be low. The high frequency of meta-analyses with comparatively few studies ($k \leq 44$ in 50% of meta-analyses reviewed) is not unique to ecology; even lower number of studies are pervasive in medical research (Kontopantelis, Springate, & Reeves, 2013) where there has been an effort to develop methods that improve the performance of meta-analyses in such scenarios (Inthout et al., 2014). Furthermore, our simulations show that the method used to calculate an uncertainty interval greatly influences how often the interval includes the true mean effect and is very important for producing intervals with close to correct coverage when *k* is low. Despite its importance, a large proportion of the ecological meta-analyses we reviewed (38%) did not report the type of uncertainty interval used, and the ones that did report their methods used intervals that are problematic when *k* is low.

Low coverage of the z-distribution confidence interval (CI) when the number of observations (in the meta-analysis context, the number of studies, *k*) are low is well known in classical statistical contexts as well as in meta-analyses (Hedges et al., 1999; Brockwell & Gordon, 2001; IntHout et al., 2014). In meta-analyses, however, approaches typically default to assuming large *k* and thus justify the application of the *z*-distribution. In ecology, this large-sample approach is often unwarranted (Figure 1A). Furthermore, bootstrapped CIs are also well known to be problematic with small *k* (Hesterberg, 2015), although ecological meta-analyses tend to prioritize the potential for non-normal distributions over concerns about small *k* (Adams et al., 1997) – based upon our results, such prioritization may be misplaced.

When *k* is low, the CI for a mean effect size (μ) based on the *z*-distribution is too narrow. Some practitioners have addressed this problem by not calculating CIs when *k* is very small (e.g.: Augusto, Delerue, Gallet-Budynek, & Achat, 2013). Others have resorted to using bootstrapped CIs (e.g.: Thébault, Mariotte, Lortie, & MacDougall, 2014). Given that bootstrapped CIs also had poor coverage when $k < 40$, this approach appears to be ill-advised. In our review, nearly half of the mean effect sizes used in an individual analysis were calculated with $k < 40$ effect sizes, where the choice of method for computing uncertainty intervals matters. As a result, many effects declared as significant probably should not have been. This is exemplified in a review of medical meta-analyses from the Cochrane Database, where of the 315 meta-analyses that yielded significant effects with the *z*-distribution CI, only 79 were significant using the HKSJ CI (Inthout et al., 2014).

The default option for frequentist CIs for μ varies among software packages. For example, a *t*-distribution CI (but without the HKSJ refinement) is Metawin's default, whereas the *z*-distribution is the default in the Comprehensive Meta-Analysis software and in the R packages *meta* and *metafor*

(metafor is one of the most common software packages currently in use by ecologists). For those planning to conduct a random-effects meta-analysis using frequentist methods, we advise use of the HKSJ CI, which employs both a weighted estimator of the variance for the overall effect size and a *t*-distribution for its associated CI (this can be set up in *metafor* using the option knha= T). Sánchez-Meca and Marín-Martínez (2008) report that the HKSJ method outperforms the simple CI-based on the *t*-distribution. However, in some scenarios, coverage could be as low as 90% even using the HKSJ CI, for example, when heterogeneity is high, $k < 10$, and the number of replicates varies greatly among studies (Inthout et al., 2014). In our simulations that did not include highly uneven number of replicates, we showed that HKSJ CI's and the Bayesian credible intervals provide accurate (or at least conservative, >95%) coverage and performed best. We encourage researchers to be aware of the software defaults when calculating an uncertainty interval, and to report the method used.

The climate change meta-analyses showed exceedingly high variation in the number of replicates reported (n_i), spanning five orders of magnitude, but the majority of values were low. In fact, $n_i < 10$ in 67% of the cases, and $n_i \leq 5$ in 51% of the cases we reviewed. This pattern may be similar in other fields of ecology (Table S2, Figures S4, S5). For example, a competition meta-analysis found n_i ranging from 1 to 1,455, with a median of 10 (Gurevitch et al., 1992). To obtain a more accurate estimate of μ, some authors specify a minimum n_i to calculate mean effect sizes (Gurevitch et al., 1992; Schirmel et al., 2016). Such censuring might improve confidence interval performance by reducing variation in replication among studies (Inthout et al. 2014) but at the high cost of discarding important information. While one would in general expect better estimates with more replication, our simulation experiment did not show important effects of the mean number of replicates on the estimation of and inferences about μ. A similar insensitivity to the number of replicates has been observed in other studies (Sánchez-Meca & Marín-Martínez 2008), although we included fewer replicates than most other simulations. Variation in replication among studies, should produce variation in within-study variance, especially when the number of replicates is small. However, in our simulations among-study variation was much larger than within-study variation, consistent with the characteristics of ecological meta-analyses (Senior et al., 2016), minimizing the role of variation in the number of replicates.

When the number of replicates reported (n_i) was unusually high, we checked a few of the original papers cited in each meta-analysis. Upon revisiting 17 of the original publications, we found at least 15 cases in which n_i was misreported (Table S3). This manifested in different ways. Some meta-analyses reported the total n_i in an experiment instead of the number of replicates per treatment. In other cases, authors reported the total n_i from repeated measurements or the numbers of individuals rather than the number of true replicates (e.g., plots or cages). There were also cases in which we were unable to verify the origin of the number reported in the meta-analysis. An incorrect n_i decreases the sampling variance for that effect size, which affects the weights and also the estimation of the overall heterogeneity (Noble et al., 2017). Researchers conducting a meta-analysis should be cautious when extracting data from the original studies to avoid misreporting (or inflating) the number of replicates. Publication of the data and code used to conduct a meta-analysis would also be useful to inform research on best practices for meta-analysis.

In our simulations using a random-effects model, the performance in the estimation of the among-study variance (σ^2_{among}) was better when the true $\sigma^2_{among}$was high (Figures S4-7). In agreement with Viechtbauer (2007), we observed that the Q-profile CI method for σ^2_{among} performed better than the bootstrap method (Figures S7-10). The Bayesian method performed best, but had coverage above the nominal level when the number of studies was low ($k < 20$). Bayesian methods led to higher perceived uncertainty in such cases, which could be real, but this could also be a consequence of positive bias in the σ^2_{among} estimates, which was more pronounced for the Bayesian methods when $k < 20$. In this scenario, one approach to improve coverage is to use priors for σ^2_{among} that perform better when k is low (Gelman, 2006). Another solution is to specify more informative priors for σ^2_{among}based on a synthesis of past publications (Higgins et al., 2009). One reason to desire good estimation of σ^2_{among} is because overestimation of this variance component can lead to higher perceived uncertainty in the estimate of μ. An additional reason is that the estimates of σ^2_{among} represent real variation in effects and could be of importance in risk assessment.

In the initial explorations with the full Bayesian model, the MCMC chains for μ converged quickly, but they converged more slowly for σ^2_{among}, often falling into a “zero variance trap” (Gelman, 2004) when the true among-study variance was close to zero. In general, convergence and mixing problems were most frequent for low k and low σ^2_{among}. While low $\sigma^2_{among}$is rare in ecology, low k is not. Of the priors we explored (Supporting Information Figures S12-15), the folded-t and the uniform prior for the standard deviation performed best when k was low (we chose the uniform prior for the final simulations because it ran slightly faster). In our simulations, the hybrid Bayesian model exhibited the practical advantages of the Bayesian methods (e.g., produces full posteriors and direct evaluation of uncertainty without approximating assumptions, among others), and was easy (and faster) to implement than the full model. On the other hand, a full Bayesian approach may be more useful for multi-level models that include missing data, hierarchical structures, and/or covariate effects (Ogle et al., 2013), and could benefit from informative priors for σ^2_{among}, particularly when k is low.

Our study simulated independent effect sizes. Often though, observed effect sizes are not independent (e.g., multiple observed effect sizes might be obtained from a single published article). As observed effect sizes within a group might respond similarly (due to similar methods, or similar environmental conditions), some of the among-study variation could be common to all members of a group or subgroup. Multi-level (hierarchical) models can be used to account for this. We believe that our results, including the insensitivity of our results to *n*, would not be materially altered in such situations, assuming the among-study variation still dominates the within-study variation. There are some challenges to be faced, however, when applying our results to more complex multi-level models. In particular, although the R package *metafor* has a function that handles multi-level models (rma.mv), the KHSJ adjustment is not available in this context, and the best that can be done with *metafor* is to construct *t*-based confidence intervals of the mean (also referred to as conditional *t*-test). For multi-level models, these *t*-based confidence intervals have inflated error rates (Luke, 2017; Song et al., in press), although they do outperform normal-based confidence intervals (Song, personal communication). Song et al. (in press) speculated that the inflated error rates of *t*-based confidence intervals resulted from not accounting for uncertainty in estimated variances. Methods exist for adjusting tests and confidence intervals to account for uncertainty in estimated variances in multi-level models, such as the Kenward-Rogers adjustment, or simulation of null distributions (Halekoh & Hojsgaard, 2014), but to our knowledge these have not been implemented in any readily available software for conducting meta-analyses.

AUTHORS CONTRIBUTIONS

All authors conceived the idea; PP collected and analyzed the data with contributions from CWO, EAH, JRB, and KO. PP led the writing; all authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY

The data compiled in the literature review, the R code for the simulation experiment, and the results from the simulation experiments are deposited in Dryad repository: https://doi.org/10.5061/dryad.zw3r22863.

ACKNOWLEDGMENTS

This research was funded by the U.S. Department of Energy (DE-SC-0010632), National Science Foundation (DEB-1655426 and DEB-1655394), and utilized Georgia Advanced Computing Resource Center resources. This is publication 2020-17 of the Quantitative Fisheries Center. We thank W. Viechtbauer for feedback on *metafor* computations, Natasja vanGestel and Kees Jan vanGroenigen for feedback during early stages of the project, Chao Song for helpful discussion, and Sergio Estay for his feedback on the abstract in Spanish.

REFERENCES

Adams, D. C., Gurevitch, J., & Rosenberg, M. S. (1997). Resampling tests for meta-analysis of ecological data. *Ecology*, *78*(4), 1277. https://doi.org/10.2307/2265879

Agresti, A., & Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. https://doi.org/10.1080/00031305.1998.10480550

Augusto, L., Delerue, F., Gallet-Budynek, A., & Achat, D. L. (2013). Global assessment of limitation to symbiotic nitrogen fixation by phosphorus availability in terrestrial ecosystems using a meta-analysis approach. *Global Biogeochemical Cycles*, *27*(3), 804–815. https://doi.org/10.1002/gbc.20069

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, *20*(6), 825–840. https://doi.org/10.1002/sim.650

Cadotte, M. W., Mehrkens, L. R., & Menge, D. N. L. (2012). Gauging the impact of meta-analysis on ecology. *Evolutionary Ecology*, *26*(5), 1153–1167. https://doi.org/10.1007/s10682-012-9585-z

Canty, A., & Ripley, A. (2017). boot: Bootstrap R (S-Plus) Functions (Version R package version 1.3-20).

Gates, S. (2002). Review of methodology of quantitative reviews using meta-analysis in ecology. *Journal of Animal Ecology*, *71*(4), 547–557. https://doi.org/10.1046/j.1365-2656.2002.00634.x

Gelman, A. (2004) Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99(466), 537–545

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gurevitch, J., & Hedges, L. V. (1999). Statistical issues in ecological meta-analyses. *Ecology*, *80*(4), 1142–1149.

Gurevitch, J., Morrow, L. L., Wallace, A., & Walsh, J. S. (1992). A meta analysis of competition in field experiments. *The American Naturalist*, *140*(4), 539–572. https://doi.org/10.1086/285428

Halekoh, U. & Hojsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. *Journal of Statistical Software*, *59*(9), 1-32. https://doi.org/10.18637/jss.v059.i09

Hartung, J., & Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, *20*(24), 3875–3889. https://doi.org/10.1002/sim.1009

Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, *80*(4), 1150. https://doi.org/10.2307/177062

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum, *The American Statistician, 69*(4), 371-386

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 137–159. https://doi.org/10.1111/j.1467-985X.2008.00552.x

Hillebrand, H., & Gurevitch, J. (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. https://doi.org/10.1111/oik.01288

IntHout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, *14*(1). https://doi.org/10.1186/1471-2288-14-25

Kontopantelis, E., Springate, D. A., & Reeves, D. (2013). A re-analysis of the Cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *PLoS ONE*, *8*(7), e69930. https://doi.org/10.1371/journal.pone.0069930

Kruschke, J.K. & Liddell, T.M. (2018). The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178-206. https://doi.org/10.3758/s13423-016-1221-4

Lajeunesse, M. J. (2015). Bias and correction for the log response ratio in ecological meta-analysis. *Ecology*, *96*(8), 2056–2063. https://doi.org/10.1890/14-2402.1

Langan, D., Higgins, J. P. T., & Simmonds, M. (2016). Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Research Synthesis Methods 8*(2), 181-198. https://doi.org/10.1002/jrsm.1198

Luke, S.G. (2017). Evaluating significance in linear mixed-effect models in R. *Behavior Research Methods*, *49*, 1494-1502. https://doi.org/10.3758/s13428-016-0809-y

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

Nakagawa, S., Noble, D. W., Senior, A.M. & Lagisz, M. (2017). Meta-evaluation of meta-analysis: ten appraisal questions for biologists. *BMC Biology*, 15:18. https://doi.org/10.1186/s12915-017-0357-7

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

Ogle, K., Barber, J., & Sartor, K. (2013). Feedback and Modularization in a Bayesian Meta–analysis of Tree Traits Affecting Forest Dynamics. *Bayesian Analysis*, *8*(1), 133–168. https://doi.org/10.1214/13-BA806

Osenberg, C. W., Sarnelle, O., & Cooper, S. D. (1997). Effect size in ecological experiments: the application of biological models in meta-analysis. *The American Naturalist*, *150*(6), 798–812.

Osenberg, C. W., Sarnelle, O., Cooper, S. D., & Holt, R. D. (1999). Resolving ecological questions through meta-analysis: goals, metrics, and models. *Ecology*, *80*(4), 1105–1117.

Pappalardo, P. K. Ogle, E.A. Hamman, J.R. Bence, B.A. Hungate, & C.W. Osenberg. (2020). Data from: Comparing traditional and Bayesian approaches to ecological meta-analysis. *Methods in Ecology and Evolution* doi:/10.5061/dryad.zw3r22863

Plummer, M. (2018). rjags: Bayesian Graphical Models using MCMC (Version R package version 4-8.).

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria.: R Foundation for Statistical Computing.

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, *13*(1), 31–48. https://doi.org/10.1037/1082-989X.13.1.31

Schirmel, J., Bundschuh, M., Entling, M. H., Kowarik, I., & Buchholz, S. (2016). Impacts of invasive plants on resident animals across ecosystems, taxa, and feeding types: a global assessment. *Global Change Biology*, *22*(2), 594–603. https://doi.org/10.1111/gcb.13093

Schmid, C. H., & Mengersen, K. (2013). Bayesian meta-analysis. In *Handbook of meta-analysis in ecology and evolution* (pp. 145–173). Princeton, New Jersey: Princeton University Press.

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. (2003). On constructing confidence intervals for a standardized mean difference in meta-analysis. *Communications in Statistics - Simulation and Computation*, *32*(4), 1191–1203. https://doi.org/10.1081/SAC-120023885

Sundar, D.-R. (2014). binom: binomial confidence intervals for several parameterizations (Version R package version 1.1-1).

Song, C., Peacor, S.D., Osenberg, C.W., & Bence, J.R. (2020). An assessment of statistical methods for non-independent data in ecological meta-analyses. *Ecology* (under review).

Thébault, A., Mariotte, P., Lortie, C. J., & MacDougall, A. S. (2014). Land management trumps the effects of climate change and elevated CO $_2$ on grassland functioning. *Journal of Ecology*, *102*(4), 896–904. https://doi.org/10.1111/1365-2745.12236

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, *26*(1), 37–52. https://doi.org/10.1002/sim.2514

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

FIGURE LEGENDS

Figure 1. Results from the literature review of ecological meta-analyses: A) distribution of the number of studies (k) reported for overall, analysis, and category levels; the median k is indicated in each panel; B) distribution of the number of replicates used in the original studies (n_i), as reported in each meta-analysis; the median n_i is indicated with a dashed line. Note that the x-axes are on a log scale.

Figure 2. Types of uncertainty intervals reported by the ecological meta-analyses. In some cases, more than one type of uncertainty interval was reported.

Figure 3. Coverage and the width of the 95% uncertainty interval for different methods used to estimate the mean effect size (μ) in a meta-analysis as a function of the number of studies (A, D), the mean number of replicates (B, E), and the among-study variance (C, F). The dashed horizontal line in panels A, B, and C indicates the nominal value of 95%. Different colors denote the method used to estimate the uncertainty interval. Error bars provide the 95% CI.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. PRISMA diagram.

Figure S2. Mean I^2 as a function of the true (simulated) among-study variance for different combinations of the mean number of replicates, n_i, and number of studies, k, in the simulated datasets.

Figure S3. Number of climate-change meta-analyses reviewed, summarized by journal in which each was published, between 2013 and 2016.

Figure S4. Results from the exploratory literature search on sub-disciplines of ecological meta-analyses. A) distribution of the number of studies (k) by sub-discipline; B) distribution of the number of replicates (n_i) used in the primary papers, as reported in each meta-analysis. Replication was not reported in any meta-analyses for ocean acidification. Note that the x-axes are on a log scale.

Figure S5. Additional results for the climate/global change meta-analysis. Variability on the median number of studies at the analysis level (A) and the median number of replicates (B) by type of organism (or variable) measured, type of environment, and meta-analysis topic.

Figure S6. Distribution of the number of replicates, n_i, in the original studies for each of the 26 meta-analysis publications in our review that provided the original data. The boxplots represent the median (thick vertical line), the 25th and 75th percentiles (box), the upper whisker extends from the box to the larger value no further than 1.5xIQR, and the lower whisker extends from the box to the smallest value at most 1.5xIQR. Extreme values that exceed the whiskers are plotted individually as solid points.

Figure S7. Performance measures of the estimation of the among-study variance as a function of the number of studies (left column), the number of replicates in the original studies (middle column) and the simulated among-study variance (right column). Performance was assessed using coverage (A, B, C), perceived uncertainty (width of the uncertainty interval) (D, E, F), bias (G, H, I), and RMSE (J, K, L). Error bars provide the 95% CI for panels A-I. Please note different scales in the y-axis for bias and width of the uncertainty interval. Simulation parameters: $n = 5, k = 25$, $\sigma^2_{\text{among}} = 0.5$, except for the cases in which that parameter was varied.

Figure S8. Performance measures of the estimation of the among-study variance as a function of the number of studies (left column), the number of replicates in the original studies (middle column) and the simulated among-study variance (right column). Performance was assessed using coverage (A, B, C), perceived uncertainty (width of the uncertainty interval) (D, E, F), bias (G, H, I), and RMSE (J, K, L). Error bars provide the 95% CI for panels A-I. Please note different scales in the y-axis for bias and width of the uncertainty interval. Simulation parameters: $n = 5, k = 25, \sigma^2_{\text{among}} = 2$, except for the cases in which that parameter was varied.

Figure S9. Performance measures of the estimation of the among-study variance as a function of the number of studies (left column), the number of replicates in the original studies (middle column) and the simulated among-study variance (right column). Performance was assessed using coverage (A, B, C), perceived uncertainty (width of the uncertainty interval) (D, E, F), bias (G, H, I), and RMSE (J, K, L). Error bars provide the 95% CI for panels A-I. Please note different scales in the y-axis for bias and width of the uncertainty interval. Simulation parameters: $n = 20, k = 25, \sigma^2_{\mathrm{among}} = 2$, except for the cases in which that parameter was varied.

Figure S10. Performance measures of the estimation of the among-study variance as a function of the number of studies (left column), the number of replicates in the original studies (middle column) and the simulated among-study variance (right column). Performance was assessed using coverage (A, B, C), perceived uncertainty (width of the uncertainty interval) (D, E, F), bias (G, H, I), and RMSE (J, K, L). Error bars provide the 95% CI for panels A-I. Please note different scales in the y-axis for bias and width of the uncertainty interval. Simulation parameters: $n = 20, k = 25, \sigma^2_{\mathrm{among}} = 0.5$, except for the cases in which that parameter was varied.

Figure S11. Bias and RMSE from the estimation of a mean effect in 2000 replicated meta-analyses as a function of the number of studies (A, B), the mean number of replicates in the original studies (C, D), and the among-study variance (E, F). Simulation parameters: $n = 5, k = 25, \sigma^2_{\mathrm{among}} = 2$, except for the cases in which that parameter was varied. Error bars provide the 95% CI for panels A-E.

Figure S12. Number of replicates yielding bad $\hat{R}$ ($\hat{R} \geq 1.1$) for different combinations of priors, true among-study variance, mean number of replicates, and number of studies.

Figure S13. Median of the posterior distribution of the among-study variance for all the different priors tested, number of replicates, number of studies, and true among-study variance. A) $n = 5$; B) $n = 25$. The vertical dashed line in each panel indicates the true among-study variance.

Figure S14. Median of the posterior distribution of the among-study variance for the four priors with the best performance (i.e., Uniform (0, 10), Uniform (0, 100), Gamma, Folded-t), number of replicates, number of studies, and true among-study variance. A) $n = 5$; B) $n = 25$. The vertical dashed line in each panel indicates the true among-study variance.

Figure S15. Median of the posterior distribution of the among-study variance for the four priors with the best performance (i.e., Uniform (0, 10), Uniform (0, 100), Gamma, Folded-t), when the number of studies was low (*k* = 5). A) $n = 5$; B) $n = 25$. The vertical dashed line in each panel indicates the true among-study variance.

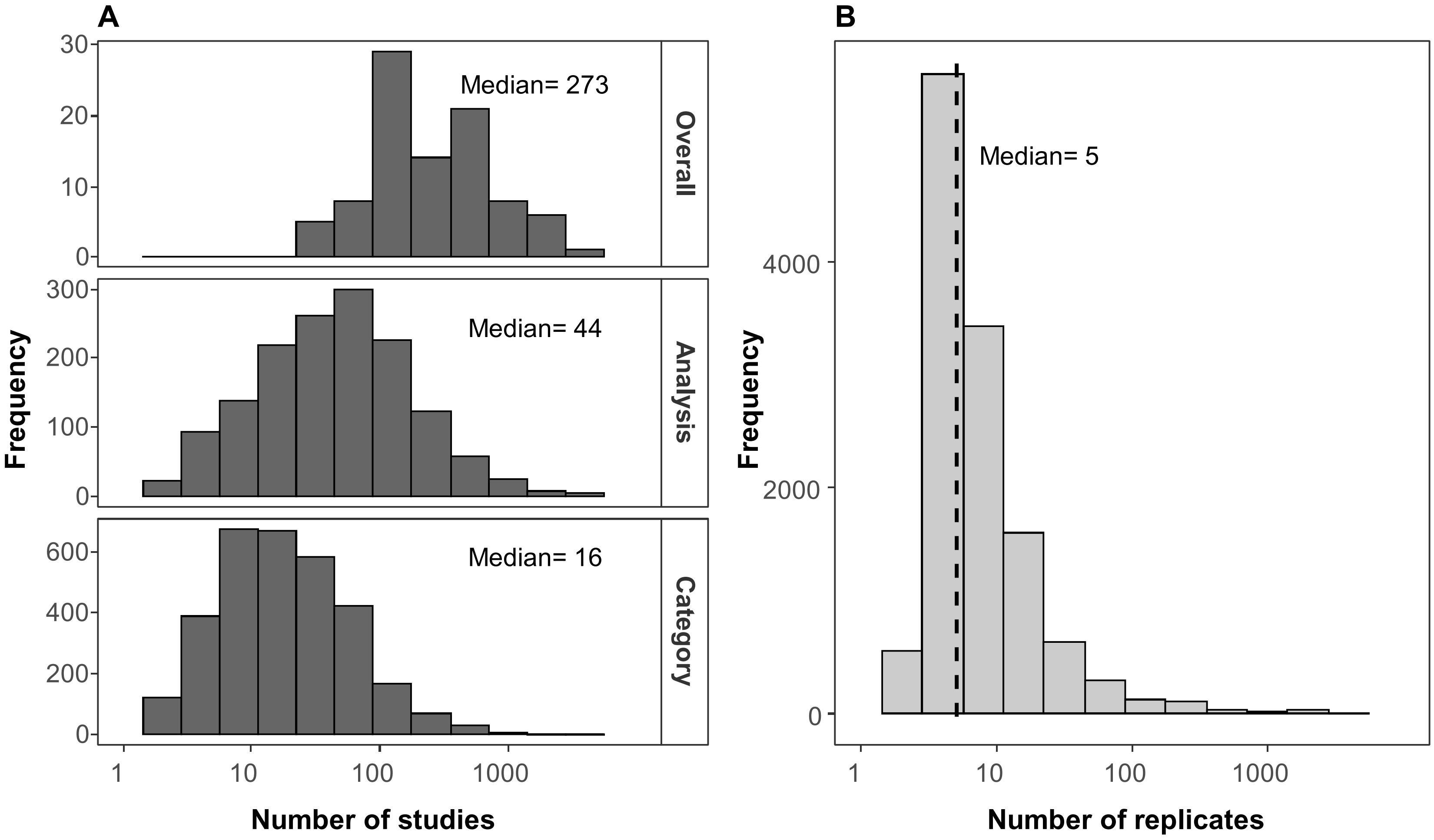
A
Median= 273
Overall
Median= 44
Analysis
Median= 16
Category
30
20
10
0
300
200
100
0
600
400
200
0
Frequency
1
10
100
1000
Number of studies
B
Median= 5
4000
2000
0
Frequency
1
10
100
1000
Number of replicates

Figure 1

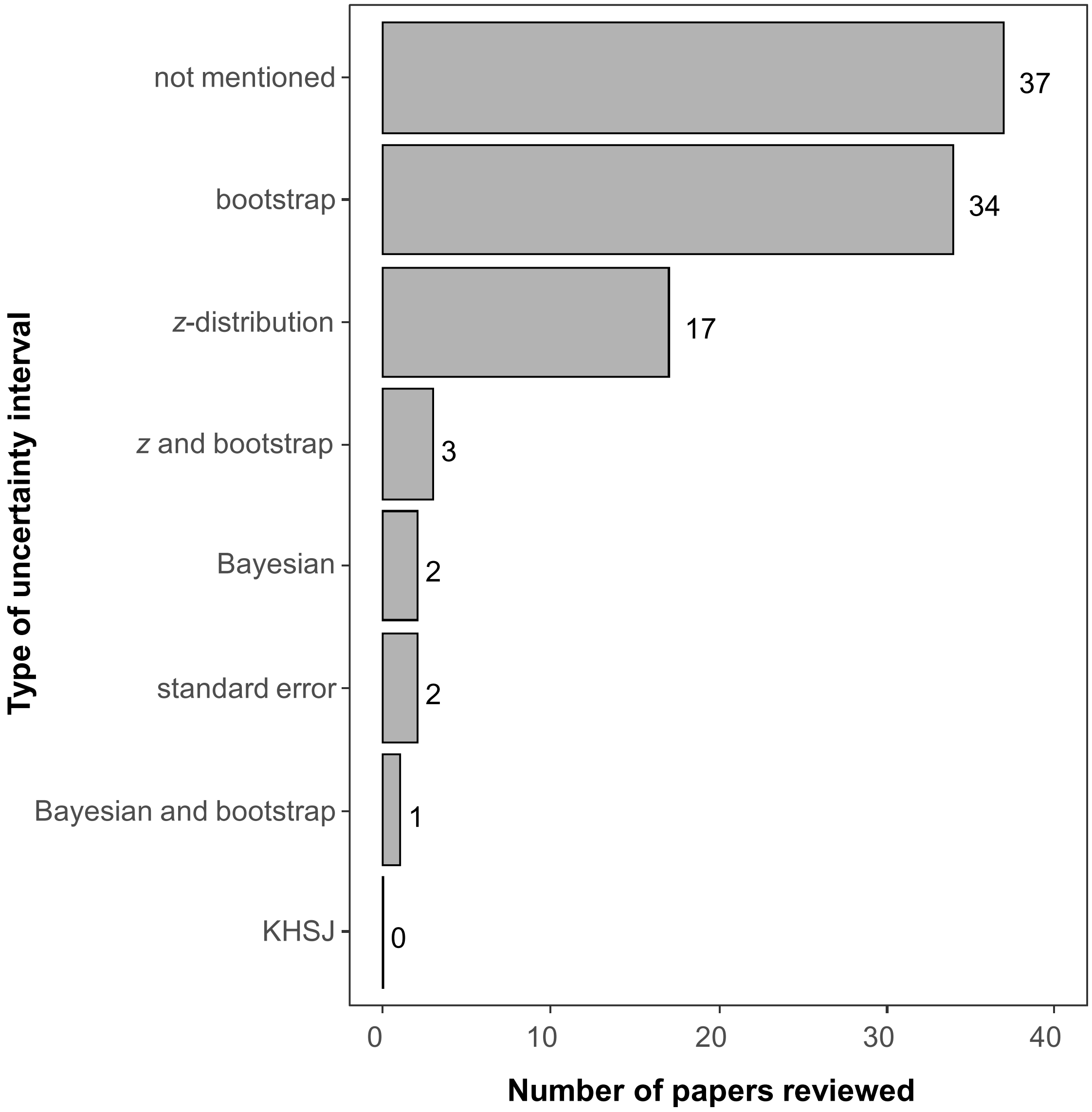
Type of uncertainty interval
not mentioned
bootstrap
z-distribution
z and bootstrap
Bayesian
standard error
Bayesian and bootstrap
KHSJ
37
34
17
3
2
2
1
0
0
10
20
30
40
Number of papers reviewed

Figure 2

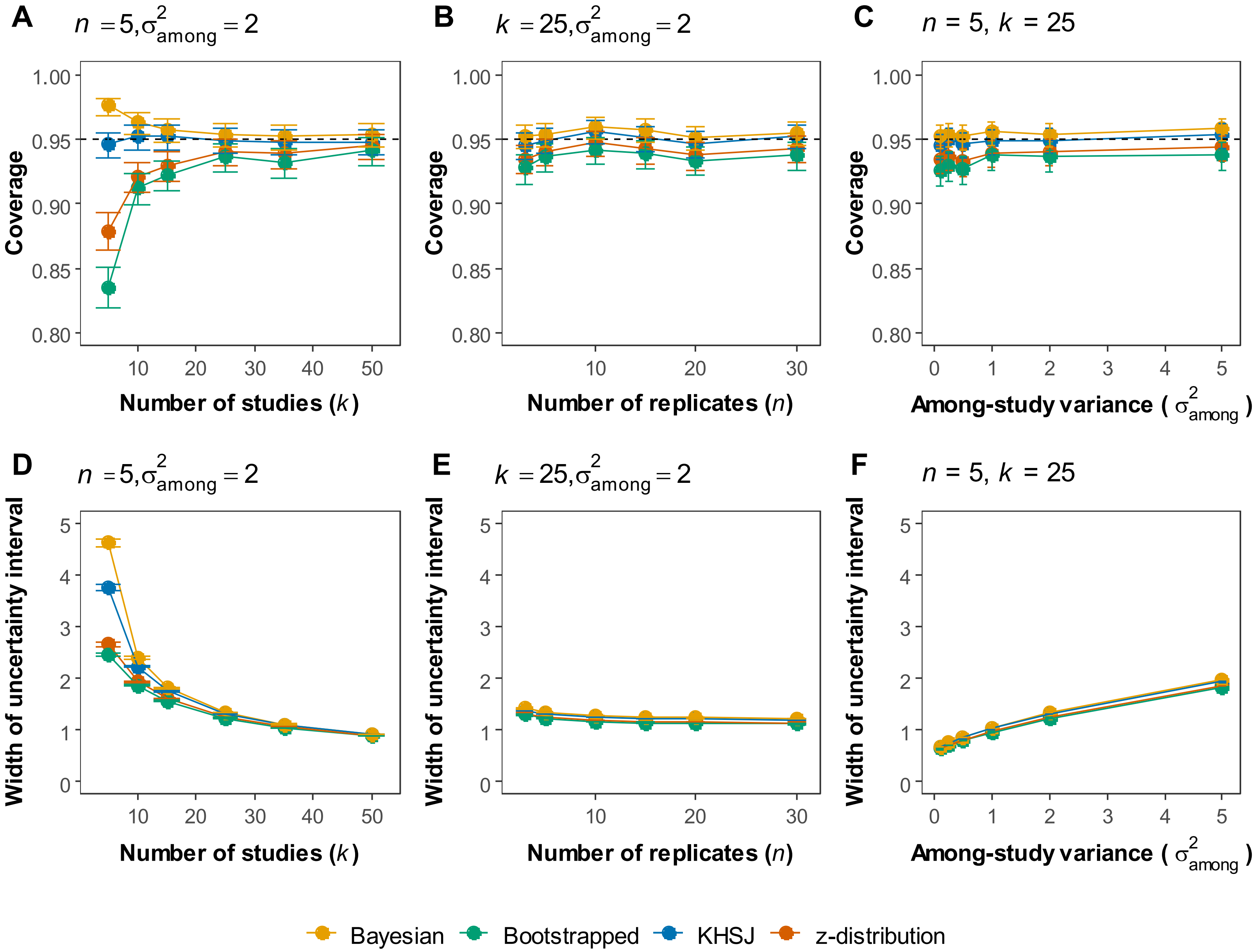
A
$n = 5, \sigma^2_{among} = 2$
B
$k = 25, \sigma^2_{among} = 2$
C
$n = 5, k = 25$
Coverage
1.00
0.95
0.90
0.85
0.80
Number of studies (k)
Number of replicates (n)
Among-study variance (σ^2_{among})
10
20
30
40
50
0
1
2
3
4
5
D
E
F
Width of uncertainty interval
Bayesian
Bootstrapped
KHSJ
z-distribution

Figure 3