Received: 30 May 2018 | Accepted: 19 November 2018

DOI: 10.1111/2041-210X.13139

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Should we be concerned about multiple comparisons in hierarchical Bayesian models?

Kiona Ogle[1,2,3] | Drew Peltier[1,3] | Michael Fell[1] | Jessica Guo[2,3] | Heather Kropp[4] | Jarrett Barber[1]

[1]School of Informatics, Computing & Cyber Systems, Northern Arizona University, Flagstaff, Arizona

[2]Department of Biological Sciences, Northern Arizona University, Flagstaff, Arizona

[3]Center for Ecosystem Science & Society, Northern Arizona University, Flagstaff, Arizona

[4]Department of Geography, Colgate University, Hamilton, New York

Correspondence
Kiona Ogle
Email: Kiona.Ogle@nau.edu

Handling Editor: Chris Sutherland

Abstract

1. Ecologists increasingly use hierarchical Bayesian (HB) models to estimate group-level parameters that vary by, for example, species, treatment level, habitat type or other factors. Group-level parameters may be compared to infer differences among levels. We would conclude a non-zero pairwise difference, separately, for each pair in the group, when the respective 95% credible interval excludes zero. Classical procedures suggest that the rejection procedure should be adjusted to control the family-wise error rate (FWER) for a family of differences. Adjustments for FWER have been considered unnecessary in HB models due to partial pooling whereby increased pooling strength – group-level parameters become more alike – could lead to decreased rejection rates (Type I error, FWER, or Power) and increased false acceptance rates (Type 2 error and its family-wise analogue).
2. To address this, we conducted a simulation experiment with factors of sample size, group size, balance (missingness), overall mean and ratio of within- to between-group variances, resulting in 2016 factor-level combinations ('scenarios'), replicated 100 times, producing 201,600 pseudo datasets analysed in a Bayesian framework. We evaluated the results in the context of a new partial pooling index (PPI), which we show is also applicable to more complex model structures based on four real-data examples.
3. Simulation results confirm intuition that rejection rates (false and true) decrease and false acceptance rates increase with increasing PPI or pooling strength (scenario-level R^2 = 0.81–0.97). The relationship with PPI differed greatly for balanced versus unbalanced designs and was affected by group size, especially for family-wise errors. Critically, an HB model does not guarantee that the FWER will follow a set significance level (α); for example, even minor imbalance can lead to FWER > α for weak to moderate pooling. These results are confirmed by the real-data examples, suggesting that ecologists need to consider FWER when applying HB models, especially for large group sizes or incomplete datasets.
4. Contrary to current thought, HB models are not immune to issues of multiplicity, and our proposed PPI offers a method for evaluating if a particular HB analysis is likely to produce FWER ≤ α (no adjustment or alternative solution required).

KEYWORDS
borrowing of strength, effective number of parameters, family-wise error, hierarchical model, imbalance, partial pooling, type I error

1 | INTRODUCTION

Application of hierarchical Bayesian (HB) models to ecological data has been rapidly increasing over the past few decades (Hooten & Hobbs, 2015; Ogle & Barber, 2008). The flexibility of the HB approach has facilitated research involving increasingly complex statistical models, applied to increasingly diverse and complex datasets (Clark, 2005; Clark & Gelfand, 2006), often involving many different factors or groups (e.g. Clark et al., 2010; McMahon & Diez, 2007; Peltier, Fell, & Ogle, 2016). Here, we are concerned with issues associated with estimating group-level parameters or effects and conducting posterior tests to infer differences among – potentially many – group levels. While multiple comparison adjustment procedures are common in applied statistics textbooks focusing on frequentist methods (e.g. Kutner, Nachtsheim, Neter, & Li, 2005), they are generally neglected in applied Bayesian textbooks. Does this mean that issues of multiplicity are not important for HB analyses?

One might use an HB model to estimate group-level parameters that could represent, for example, simple additive treatment effects (akin to ANOVA), covariate effects (e.g. treatment- or individual-level coefficients in a regression model), or parameters in a theory-inspired process model (e.g. growth rate, maximum photosynthetic rate, tissue life span, reproductive potential). These parameters or effects may vary by factors such as species, experimental treatment level, habitat type, site and so forth, and one may be interested in determining if there are differences among the group-level parameters. Within the HB framework, this is accomplished by computing pairwise differences between levels j and k ($D_{j,k}$), and one would typically reject H_o: $D_{j,k} = 0$ at the '5% level' if the 95% posterior credible interval (CI) for $D_{j,k}$ excludes zero. Should this rejection criteria be adjusted if the family of pairwise comparisons is 'large'?

For a group size of, say, 10, there are '10 choose 2' = 45 unique pairwise comparisons. In a frequentist analysis or fixed effects formulation, one would be concerned about conducting this many comparisons. For example, if the significance level (or Type 1 error rate) for an individual test is set at α, then the significance level for a family of F tests is only $1 - (1 - \alpha)^F$ (Kutner et al., 2005); assuming that the tests are independent, which can be much greater than the desired α value. For example, for $F = 45$ and $\alpha = 0.05$, there is a high probability (*c.* 90%) that at least one test in the family leads to false rejection of H_o: $D_{j,k} = 0$ (i.e. there is a very good chance of committing a family-wise error). To address this multiplicity problem, one could employ, for example, the Tukey, Scheffé or Bonferroni multiple comparison procedures to achieve a family-wise significance level that is less than or equal to the individual rate, α (Kutner et al., 2005). Each of these procedures effectively inflates the confidence interval and/or reduces the p-value threshold for each individual test (Kutner et al., 2005; Westfall, Johnson, & Utts, 1997). The choice of the specific procedure often depends on how many tests are to be conducted (size of F), the type of test to be performed (e.g. pairwise differences or contrasts involving multiple effects), whether the design is balanced or not, and/or which procedure yields the narrowest confidence intervals or larger adjusted p-value threshold (Kutner et al., 2005).

Within the HB framework, we lack a clear procedure to address multiplicity. Recent work, however, suggests that multiplicity and inflated family-wise error rates may be of little concern in HB models, due to the effects of partial pooling (Gelman, Hill, & Yajima, 2012). For example, for an effect or parameter that varies by group level k, an HB model would assume that the group-level parameters, θ_k, come from a parent distribution defined by population-level parameters (e.g. a global mean and a variance that describe variability among levels within a group). The pooling property essentially 'pulls' each θ_k towards the global mean, reducing differences among the levels (Gelman & Hill, 2007; Gelman, Hwang & Vehtari 2014). Others, however, suggest modifications to HB models to account for multiplicity (e.g., Li & Shang, 2015; Nashimoto & Wright, 2008; Shang, Cavanaugh, & Wright, 2008; Westfall et al., 1997), but the approaches can greatly increase the complexity of the model, require specification of informative priors that may not be supported by existing information or require an a priori ranking of the group-level means, which may be difficult to determine in practice. Thus, the suggestion that partial pooling diminishes the need to adjust for multiplicity is attractive but has not be rigorously tested against the types of 'messy' data that ecologists routinely work with.

Our intuition tells us that rejection rates (e.g., Type I error, family-wise error rate and Power) and the potential need for a multiplicity adjustment should increase as the partial pooling strength decreases (see also, Gelman et al., 2012). To explore this intuition, we conducted a simulation experiment with factors of sample size, group size, global mean, ratio of within- to between-group variances, degree of imbalance (or missingness that results in unequal sample sizes among group levels), and the distribution from which group-level means arise. The experiment resulted in 2,016 factor-level combinations ('scenarios'), replicated 100 times ('replicates') to produce 201,600 pseudo datasets that we analysed in a Bayesian framework. For each scenario, we used the 100 replicates to evaluate rejection rates and false acceptance rates, allowing us to address the questions: (a) How can we quantify the degree of partial pooling? (b) How are rejection and false acceptance rates – of individual comparisons and families of comparisons – affected by the degree of partial pooling? (c) In the context of such rejection rates and error rates, is an HB model advantageous over a non-hierarchical model?

Finally, (d) are HB models immune to multiple comparison issues, such as inflated family-wise significance levels?

Additionally, we evaluated if the results from the simulation experiment are supported by analyses that are more representative of the types of complex data and models encountered by ecologists. We tested this by drawing upon four diverse, real-data examples that were analysed in an HB framework. For each example, we computed the partial pooling index (PPI) based on the original data, and we simulated pseudo data to evaluate error rates, allowing us to evaluate the applicability of the aforementioned simulation results to more complex models.

2 | MATERIALS AND METHODS

2.1 | Simulation experiment

To address our research questions, we simulated pseudo data given known ('true') parameter values. Data were simulated from a normal distribution such that for group level $k = 1, 2, \ldots, K$ and observation $i = 1, 2, \ldots, N$:

$$y_i \sim \text{Normal}\left(\mu_{k(i)}, \sigma^2\right), \quad (1)$$

$k(i)$ denotes group level k associated with observation i, and σ^2 is the residual error variance or, here, the within-group variance.

Data were generated under different combinations of factors (see Figure S1, Supporting Information). We refer to a particular combination of factor levels as a 'scenario'. We explored three factor levels of group size (K = 5, 10, or 20; Figure S1B, Supporting Information), resulting in 10, 45 or 190 pairwise comparisons respectively. Let n_k denote the sample size associated with group level k such that the total sample size for a particular scenario is $N = \sum_{k=1}^{K} n_k$. We varied the maximum group-level sample size, $\max(n_k)$, from 3 to 1,000, with a total of seven factor levels (Figure S1A, Supporting Information). Under a completely balanced design ($n_k = \max(n_k)$ for all k), N varied from 15 (3 observations × 5 group levels) to 20,000 (1,000 observations × 20 group levels). We also explored three levels for the global mean (m = 0, 10, 100) and between-group variation (standard deviation = s) relative to within-group variation (σ, Equation (1); Figure S1C, Supporting Information); we set $\sigma = 1$ for all scenarios and adjusted s to vary within one order of magnitude of σ (s = 0.1, 1 or 10).

Group-level means were drawn from a normal distribution: $\mu_k \sim \text{Normal}(m, s^2)$. For $m = 0$, we also simulated μ_k from a uniform distribution, $\mu_k \sim \text{Uniform}(-3s, 3s)$, that covered approximately the same range of potential values while allowing for greater representation of more 'extreme' group-level means relative to the normal distribution. We intentionally set some of the group-level means equal to other group-level means, allowing us to evaluate rejection and/or error rates. That is, for K = 5, 10 and 20, we drew 3, 6 and 12 unique μ_k values, respectively, from the corresponding normal or uniform distributions; the remaining 2, 4 and 8 μ_k values, respectively, were set equal to one of the simulated μ_k values. This was done randomly, without replacement, such that for K = 5, 10 and 20, there were 2, 4 and 8 pairs, of 10, 45 and 90 pairs, respectively, with corresponding true $D_{i,j} = 0$.

The above factors – sample size, group size, overall mean, among group variance and distribution – were combined with eight levels of imbalance (Figure S1E,F, Supporting Information). One level represented a completely balanced design (same n_k for all k), and the others represented different levels of imbalance, ranging from about 10% to 60% 'missing data'. For the scenarios defined by some degree of imbalance, we randomly selected a subset of group levels to receive a 'small' sample size of $n_k = \min(n_k)$, and the remaining group levels were assigned the maximum sample size of $n_k = \max(n_k)$. The number of groups with $n_k = \min(n_k)$ was varied to achieve different levels of missingness (Figure S1F, Supporting Information).

For normally distributed μ_k, the total number of scenarios that we simulated was 1512; the number of scenarios associated with uniformly distributed μ_k was 504 (i.e. 1,512/3 since there was only one level for m). This resulted in a total of 2,016 scenarios. Furthermore, we simulated 100 independent (random) datasets ('replicates') per scenario, resulting in 201,600 pseudo datasets and over 2.9 million pseudo observations of y.

2.2 | Bayesian analysis of the pseudo data

We analysed the pseudo datasets in an HB framework (see Figure 1a) to obtain posterior estimates of the group-level means (μ_k), variance components (σ and s) and global mean (m). For each scenario d and replicate r, we computed all pairwise differences such that for levels j and k ($1 \le j < k \le K_d$), $D_{j,k,d,r} = \mu_{j,d,r} - \mu_{k,d,r}$. The null hypothesis ($H_o$: $D_{j,k,d,r} = 0$) was rejected if the central 95% posterior credible interval (CI) for $D_{j,k,d,r}$ did not contain zero, suggesting that group levels j and k differed. We simultaneously analysed all pseudo observations, since no parameters are shared across datasets; the analysis is equivalent to implementing separate models for each dataset since there is no pooling across the different levels of d and r.

2.3 | Partial pooling index

Our definition of a PPI (Equation 2) requires that we implement the above HB model (Figure 1a) and two additional models, similar to Gelman, Carlin, et al. (2014). One version assumes no pooling (Figure 1b); the other imposes complete pooling (Figure 1c). The three model variants are fit to the pseudo data, and their posterior results are used to compute the PPI of the focal HB model:

$$\text{PPI} = \frac{p_{nh} - p_h}{p_{nh} - p_{cp}}, \quad (2)$$

where p_{nh}, p_h and p_{cp} represent the effective number of parameters in the non-hierarchical, hierarchical (HB) and complete pooling models respectively. In theory, $0 \le \text{PPI} \le 1$, where PPI = 0 if there is no pooling among the group levels, and PPI = 1 if there is complete pooling among the group levels. Note that p_{nh}, p_h and p_{cp}, and thus

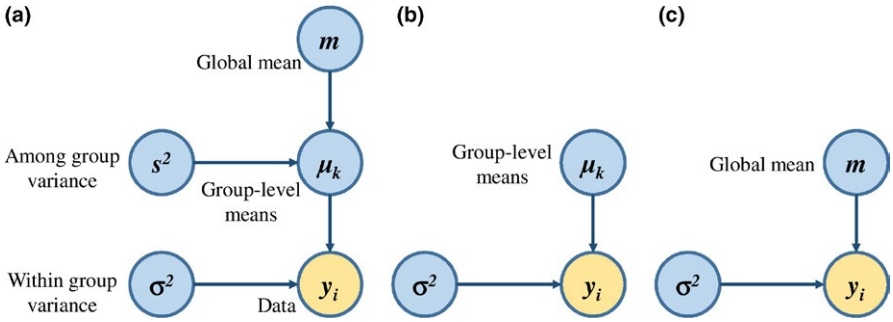

FIGURE 1 Directed acyclic graphs (DAGs) summarizing the three Bayesian models fit to each dataset (for $y_{i,d,r}$, $i = 1, 2, ..., N_d$ observations), associated with each scenario (d) and replicate (r) (d and r subscripts not shown in DAGs for simplicity). (a) The HB model assumes a normal likelihood, $y_{i,d,r} \sim Normal(\mu_{k,d,r}, \sigma^2_{d,r})$, with group-level means modelled hierarchically around a global mean (m), $\mu_{k,d,r} \sim Normal(m_{d,r}, s_{d,r}^2)$, for $k = 1, 2,..., K_d$ levels. (b) The non-hierarchical model assumes the same likelihood, but specifies independent, vague priors for the group-level means, $\mu_{k,d,r} \sim Normal(0, 10{,}000)$. (c) The complete pooling model does not estimate separate means for each group such that $y_{i,d,r} \sim Normal(m_{d,r}, \sigma^2_{d,r})$. In (a) and (c), m is assigned a vague prior, $m_{d,r} \sim Normal(0, 10{,}000)$. In (a) and (b), σ^2 is the within-group variance, whereas in (c), it captures the combined within- and among-group variance. In (a), s^2 is the among-group variance; σ and s are assigned vague priors, $\sigma_{d,r}, s_{d,r} \sim Uniform(0, 100)$

PPI, are computed for each r and d; we avoid subscripting by r and d in Equation (2) for clarity. We also computed the scenario-level PPI (PPI_d) by averaging the replicate-level PPI ($PPI_{d,r}$, Equation 2) across all r for each d.

There are multiple ways the p terms (p_{nh}, p_h and p_{cp}) can be computed, and we used the formula associated with the Watanabe–Akaike Information Criterion (WAIC) (Gelman, Hwang, et al., 2014; Gelman, Carlin, et al. 2014). We explored using the Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, & van der Linde, 2002; Gelman, Carlin, et al., 2014) formulas, but found that WAIC gave notably more realistic values of p and PPI (Supporting Information, section S1 and Figure S2, Supporting Information). Gelman, Carlin, et al. (2014) and the Supporting Information (section S1) give the WAIC expression for p.

2.4 | Rejection rates and error rates

We evaluated individual and family-wise rejection (false and true) and false acceptance rates. For individual comparisons, the Type 1 error and Power (1 – Type 2 error) were evaluated with respect to rejection of the null hypothesis, H_o: $D_{j,k,d,r} = 0$. A Type 1 error is the probability of *rejecting* H_o given that H_o is *true* (false rejection), a Type 2 error is the probability of *accepting* H_o given that H_o is *false* (false acceptance), and Power is the probability of *rejecting* H_o given that H_o is *false* (correct rejection) (e.g. Kutner et al., 2005).

We grouped the pairwise differences into those having a true difference of zero ($D_{j,k,d} = 0$); a total of $n_{0,d}$ pairs fall in this group. The remaining $n_{D,d}$ pairs have non-zero true differences ($D_{j,k,d} \neq 0$). Note that $n_{0,d} + n_{D,d} = K_d\ (K_d - 1)/2$, the total number of pairwise comparisons. Subsetting by the first group (true $D_{j,k,d} = 0$), for each replicate, we determined the number of pairwise differences ($n_{R,d,r}$) out of the $n_{0,d}$ pairs whose 95% CI excluded zero (reject H_o); the ratio $n_{R,d,r}/n_{0,d}$ provides a replicate-level estimate of the Type 1 error rate. Subsetting by the second group (true $D_{j,k,d} \neq 0$), we determined the number of pairwise differences ($n_{A,d,r}$) out of the $n_{D,d}$ pairs whose 95% CI contained zero (accept H_o); the ratios $n_{A,d,r}/n_{D,d}$ and $(n_{D,d} - n_{A,d,r})/n_{D,d}$ provide replicate-level estimates of the Type 2 error rate and Power, respectively. We averaged these ratios across replicates to obtain scenario-level error rates and Power for visualization of the results.

We also considered family-wise error rates. The family-wise error rate (FWER) has typically been used in the context of Type 1 errors, and thus, we define FWER as the probability of falsely rejecting at least one individual comparison (e.g. a single H_o) among a family of comparisons (e.g., Li & Shang, 2015). Less commonly, we may be interested in evaluating a family-wise Type 2 error rate, which we denote, FWER2. By analogy, FWER2 is the probability of falsely accepting at least one individual comparison. For each d and r, we commit a Type 1 family-wise error if $n_{R,d,r} > 0$, and we commit a Type 2 family-wise error if $n_{A,d,r} > 0$. Thus, we define $I_{R,d,r} = 1$ if $n_{R,d,r} > 0$ ($I_{R,d,r} = 0$ if $n_{R,d,r} = 0$) and $I_{A,d,r} = 1$ if $n_{A,d,r} > 0$ ($I_{A,d,r} = 0$ if $n_{A,d,r} = 0$). Averaging $I_{R,d,r}$ and $I_{A,d,r}$ across all r for each d provides estimates of the scenario-level FWER and FWER2, respectively.

2.5 | Implementation of models and processing of posterior results

Pseudo data were generated in R (R Core Team, 2017) and the Bayesian models (Figure 1) were implemented in JAGS 4.0.0 (Plummer, 2003, 2015) via R using the RJAGS package (Plummer, 2013) (code for implementing the three Bayesian model variants and for computing PPI is provided in Supporting Information section S2, along with example pseudo datasets). For each model, three parallel MCMC chains were run for a burn-in period; after convergence, each model was updated to obtain 3,000 relatively independent samples, which were subsequently used to compute posterior summary statistics. Convergence was verified using the Brooks–Gelman–Rubin diagnostic (Brooks & Gelman, 1998; Gelman & Rubin, 1992) via the gelman.diag function in the coda package ('RJAGS') (Plummer, Best, Cowles, & Vines, 2006). Data

simulation, Bayesian models, storage and processing of posterior samples (coda) and error calculations were performed in R using supercomputing resources at Northern Arizona University (nau.edu/hpc).

2.6 | Analysis of error rates and partial pooling strength

First, we evaluated the factors causing variation in the scenario-level PPI (PPI_d) and the variance in PPI across replicates within each scenario ($VPPI_d$). We conducted linear models using the 'lm' function in R, with response variables logit(PPI_d) and log($\sqrt{VPPI_d}$). For logit(PPI_d), we conducted a weighted regression, with weights = $1/VPPI_d$. We explored a suite of models and used partial R^2 to select the final, most parsimonious model. The final models for both response variables included the log-transformed, scenario-level covariates of group size, sample size and the ratio of among- to within-group variation – log(K_d), log(N_d) and log(s_d/σ), respectively – and the categorical (binary) covariates indicating the distribution from which group-level means were drawn (normal or uniform) and whether the dataset is balanced or not (M_d). We focus on the final models, which included the main effects and all relevant two-way interactions among the aforementioned covariates. (Quadratic effects were consistently non-significant, as were the covariates representing the global mean, m_d, and percent missing, PM_d.) We excluded scenarios (<0.5%) with $PPI_d < 0$; all records were retained for $VPPI_d$. To identify the factors explaining the greatest amount of variation in PPI_d and $VPPI_d$, we computed partial R^2 values, which were obtained by comparing a reduced model – excluding a particular factor (main effect and all interactions involving that factor) – to the full (final) model that included all factors (main effects and interactions) of interest. Partial $R^2 = (SSE_{reduced} - SSE_{full})/SSE_{reduced}$ (Kutner et al., 2005), where SSE is the sum of squared errors, obtained for both the reduced and full models.

Next, we explored how the error rates – Type 1, Type 2, FWER and FWER2 – varied with PPI. To evaluate Type 1 errors, we conducted binomial (logistic) regressions using the 'glm' or 'glm2' functions in R, with the replicate-level counts that tabulate the number of comparisons incorrectly rejected ($n_{R,d,r}$) (number of 'successful trials') and correctly accepted ($n_{0,d} - n_{R,d,r}$) (number of 'failures') (i.e. $n_{0,d}$ represents the 'total number of trials'). Likewise, to evaluate Type 2 errors, binomial regressions were conducted with the replicate-level number of comparisons incorrectly accepted ($n_{A,d,r}$) ('successes') and correctly rejected ($n_{D,d} - n_{A,d,r}$) ('failures'). Binomial regressions were also implemented to evaluate FWER and FWER2, using the replicate-level binary indicators that denote if at least one Type 1 error ($I_{R,d,r}$) or at least one Type 2 error ($I_{A,d,r}$) was committed.

In all binomial regressions, we used a logit link function for the probability ($q_{d,r}$) of committing an error. Visual inspection of the scenario-level error rates (e.g. Figure 3) suggested models of the form logit($q_{d,r}$) = $a_d + b_d$ $PPI_{d,r}$, given replicate-level PPI ($PPI_{d,r}$). We explored a suite of models to identify the final, most parsimonious model. We specified linear models for the a_d and b_d terms, which included the scenario-level continuous covariates of log(K_d), log(N_d), PM_d and log(s_d/σ), and the binary indicator of M_d. All models included the main effects of these covariates, and some included all two-way interactions among the covariates and/or quadratic effects of the continuous covariates. We selected the final model based on differences in AIC and the proportion of the deviance explained by the model (R^2). The final models for a_d and b_d included the main effects of log(K_d), log(N_d), PM_d and M_d and all two-way interactions.

Similar to the PPI analysis, to identify the factors explaining the greatest amount of variation in the error rates, we computed partial R^2 values by comparing the residual deviance obtained from a reduced model, which excluded a particular factor, to the full (final) model that included all factors of interest. Here, partial $R^2 = (deviance_{reduced} - deviance_{full})/deviance_{reduced}$, where 'deviance' is the residual deviance from the full or reduced model. We also evaluated the importance of PPI via two approaches. One approach eliminated $PPI_{d,r}$ from the model and retained the other simulation factors (covariates) to obtain the partial R^2 associated with $PPI_{d,r}$. The other approach included $PPI_{d,r}$ as the only explanatory variable (all simulation factors were excluded), such that a and b are scalar coefficients, and we computed the proportion of the deviance explained by $PPI_{d,r}$ as R^2 = (null deviance – residual deviance)/(null deviance).

Using the final binomial regression results, we obtained predicted FWER as a function of PPI for different levels of N, K, PM and M. We also computed the critical PPI (PPI_{crit}) that is expected to yield FWER = α (for illustration, we set $\alpha = 0.05$). In particular, $PPI_{crit} = (\text{logit}(\alpha) - a)/b$, and we obtained predictions for PPI_{crit} for different levels of N, K, PM and M given the coefficient estimates associated with the a and b models for FWER.

2.7 | Real-data examples

Our simulation experiment does not capture the complexity of most ecological data and associated HB models. Thus, we drew-upon four real-data examples to evaluate if more complex model structures align with the simulation experiment results. We summarize the four examples in Table 1 and provide additional details in the Supporting Information (section S3). The examples represent diverse Bayesian applications, including: (a) a multivariate regression that evaluates allometric scaling relationships among plant mass, length and diameter, and that compares parameter estimates among 49 plant species (Price, Ogle, White, & Weitz, 2009), (b) a mixed effects regression that evaluates drivers of plant water stress, from which we compute contrasts involving eight subjects (shrubs) to determine if model parameters differ among two treatment groups (Guo & Ogle, 2018), (c) a nonlinear mixed effects regression for orange tree growth over time, with a hierarchical model for tree-level growth parameters, and (d) a generalized linear model involving Bernoulli data and a nontraditional link function to determine if dogs learn from repeated stimuli. The latter two are taken from the OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) examples volumes, and we evaluate differences in parameters among individuals (trees or dogs). We note that the comparisons that we conducted were not the focus of the original four studies.

For each example, the focal HB model was modified to give the complete pooling and non-hierarchical versions. The three versions were fit to each dataset, allowing us to compute PPI, Equation (2), for each example. To evaluate potential error rates, we simulated 100 representative psuedo datasets per example. To do this, we obtained the posterior means for the model parameters and pairwise differences or contrasts for each example's HB model as applied to the real data. We treated these posterior means as the 'true' parameter values and simulated data given these true values and the original covariate data. If the 95% CI for a pairwise difference, based on the real data, indicated that a particular parameter did not differ between the two group levels being compared, then we set their 'true' values equal to each other (i.e. the average of the two posterior means). We then fit the HB, complete pooling and non-hierarchical model variants to each of the pseudo datasets and computed error rates and PPI. The real data and code for implementing the three model variants and computing PPI for each example are provided the Supporting Information (section S4).

3 | RESULTS

3.1 | Variation in the PPI

We developed PPI, Equation (2), to quantify the pooling strength of the focal, HB model (Figure 1a). Most PPI values were between 0 (no pooling) and 1 (complete pooling); only <0.5% of the scenarios produced scenario-level PPI < 0 (Figure 2). We also identified the factors explaining variation in PPI. The final model explained *c*. 74% (adjusted $R^2 = 0.744$) of the variation in logit(PPI_d); log(N), log(s/σ) and the group-level sampling distribution were the most important predictors of logit(PPI_d) (partial $R^2 = 0.852$, 0.827 and 0.625, respectively), followed by the binary indicator for balance, M (partial $R^2 = 0.276$). The final model for the scenario-level variation in PPI explained 75% of the variation in log($\sqrt{VPPI_d}$), and log(N) and log(s/σ) were the most important factors (partial $R^2 = 0.601$ and 0.479, respectively), followed by the group-level sampling distribution (partial $R^2 = 0.270$). In general, smaller N and smaller s/σ lead to higher and more precise PPI values or stronger pooling (Figure 2b).

3.2 | Error rates versus PPI

The family-wise Type 1 error rate (FWER) produced by the HB model was never greater – and in many cases, notably lower – than the FWER produced by the non-hierarchical model (Figure S3C, Supporting Information). With respect to individual comparisons, the Type 1 error rate, Power and Type 2 error rate were slightly affected by an HB specification (Figure S3A,B); in particular, rejection rates (false and true) were slightly lower for the HB model. However, the family-wise Type 2 error rate (FWER2) was nearly identical among the HB and non-hierarchical models (Figure S3D).

Focusing on the error rates produced by the HB model, we evaluated how they varied with PPI (Figure 3). Greater pooling (higher PPI) is associated with lower rejection rates (Type 1, Power, FWER; Figure 3a–c) and higher false acceptance rates (Type 2 and FWER2; Figure 3b,d). Here, we are most interested in understanding how variation in PPI affects the Type 1 error rate and especially the FWER. The final binomial regressions explained 69.1% (or 92.0%) and 62.0% (or 95.7%) of the variation in the replicate-level (or scenario-level) Type 1 error rate and FWER respectively (Table 2). Based on partial R^2 values (Table 2), PPI was the most important predictor of both Type 1 error rates and FWER, followed by the binary indicator for balance, M. While log(K), log(N) and PM had comparatively little influence on the Type 1 error rate, and similarly for the influence of log(N) and PM on FWER (Table 2), they were often

TABLE 1 Summary of the four real-data examples

	Allometric scaling	Plant water stress	Orange trees	Dog learning
Source	Price et al. (2009)	Guo and Ogle (2018)	OpenBUGS Example Vol II	OpenBUGS Example Vol I
Overview of model (all Bayesian)	Multivariate, linear regression with a hierarchical model for species-level parameters and a stochastic covariate model.	Nonlinear mixed effects regression model with a hierarchical model for shrub-level parameters.	Nonlinear regression model with a hierarchical model for tree-level growth parameters.	Hierarchical, generalized linear model (GLM) with binary (Bernoulli) data and a non-standard link function.
Comparisons conducted	4,704 (1,176 species pairs × 4 parameters)	7 ([watered vs. control] × 7 parameters)[a]	30 (10 tree pairs × 3 parameters)	870 (435 dog pairs × 2 parameters)
N (sample size)	1,162 plants, 3 variables each	112 measurements	35 measurements	750 binary records
K (group size)	49 species	8 shrubs, 2 treatment levels	5 trees	30 dogs
Balanced?	No	No	Yes	Yes

[a]Contrasts conducted for watered versus control treatment levels; treatment-level means were computed across four shrubs each (shrub-level parameters were modelled hierarchically around global parameters that do not vary by treatment level).

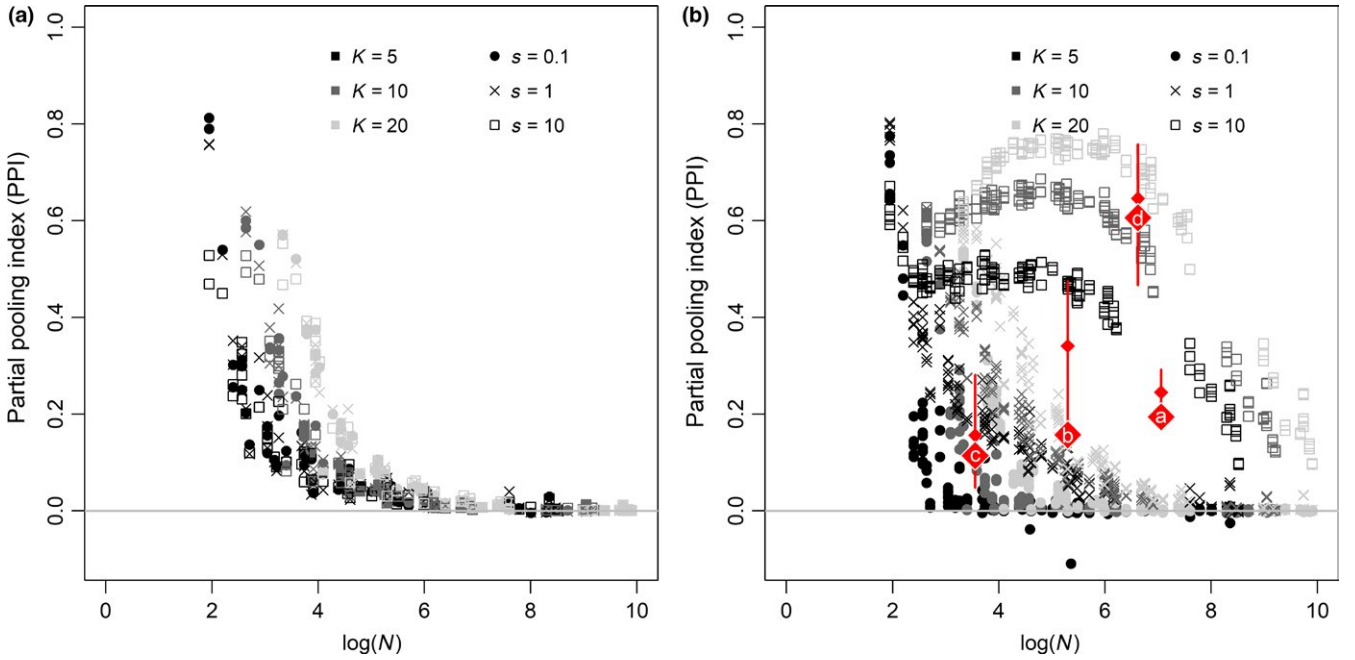

FIGURE 2 Scenario-level partial pooling index (PPI_d) versus $\log(N_d)$, where N = sample size, for group means (μ_k) that were simulated from a (a) uniform versus (b) normal distribution. Symbol shading reflects group size (K) and symbol shape reflects the among-group standard deviation (s); $\sigma = 1$ (within group standard deviation). In both panels, each point represents the mean PPI value of 100 replicates, for each of (a) 504 and (b) 1,512 scenarios. The horizontal grey lines indicate the expected lower bound on PPI. The large red diamonds in (b) correspond to the PPI based on the four real-data examples; the small red diamonds and 95% uncertainty intervals are based on each example's 100 pseudo datasets that were simulated given the posterior results for the real data. Examples are denoted by letters inside the symbols: (a) allometric scaling, (b) plant water stress, (c) orange trees and (d) dog learning (Table 1)

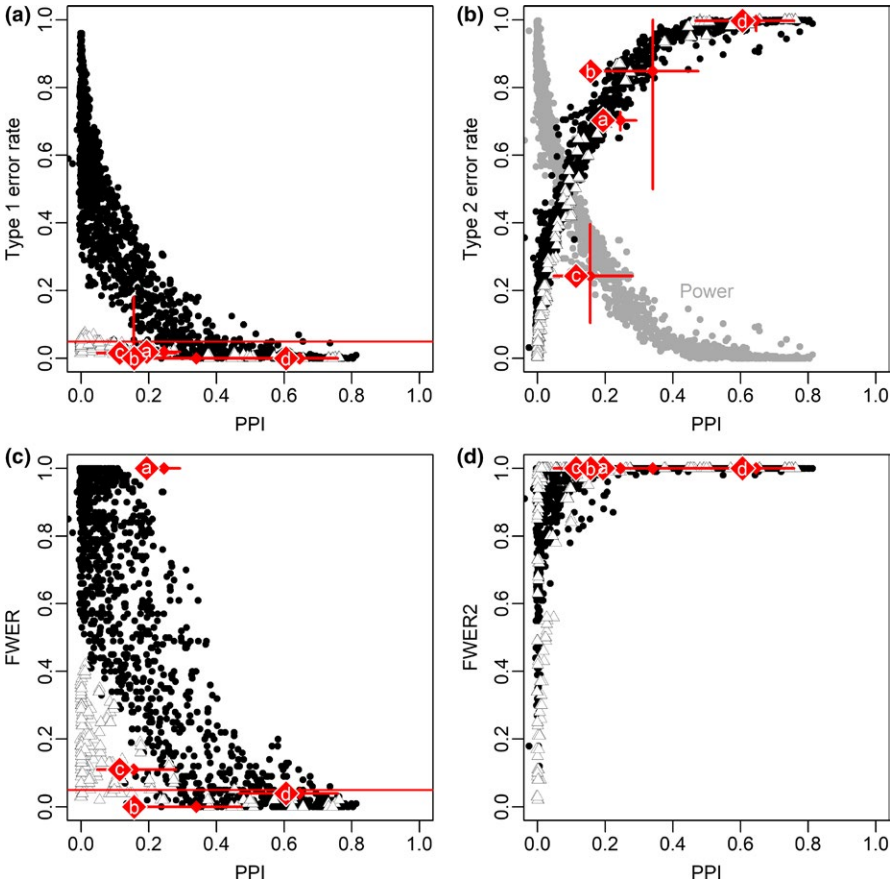

FIGURE 3 Scenario-level error rates derived from 100 replicates for each of 2016 scenarios versus scenario-level partial pooling index (PPI_d). Error rates often differ depending on whether a scenario produced a balanced (open triangles, $M = 0$) or unbalanced (black circles, $M = 1$) dataset. (a) Type 1 error rate for individual comparisons, (b) Type 2 error rate for individual comparisons, overlaid with Power (1 – Type 2 error rate; grey symbols), (c) family-wise Type 1 error rate (FWER) and (d) family-wise Type 2 error rate (FWER2). Red horizontal lines in (a) and (c) denote the nominal comparison-wise error rate ($\alpha = 0.05$). The large and small red diamonds correspond to the results for the four real-data examples (see Figure 2b); uncertainty intervals are not relevant to FWER and FWER2, and Power is not shown for the real-data examples

TABLE 2 Summary of the binomial regression fits for each error type based on the final models (see Supporting Information section S5 for details), and the relative importance of each covariate or simulation factor

Error type	[a]Scenario-level R^2	[b]Rep.-level R^2	[c]R^2 for partial pooling index (PPI) only	[d]ΔR^2	[e]Partial R^2 for each covariate				
					PPI	log(*K*)	log(*N*)	*PM*	*M*
Type 1	0.920	0.691	0.479	0.212	0.547	0.023	0.081	0.008	0.313
Type 2	0.971	0.935	0.893	0.042	0.905	0.016	0.326	0.037	0.058
FWER	0.957	0.620	0.382	0.238	0.507	0.161	0.055	0.006	0.208
FWER2	0.805	0.453	0.189	0.264	0.178	0.256	0.067	0.001	0.090

[a]The binomial regressions were used to predict scenario-level error rates (q_d), and empirical scenario-level mean error rates were regressed on predicted q_d to evaluate scenario-level model fit (R^2). [b,c]Replicate-level R^2 was computed for each binomial regression as R^2 = (null deviance − residual deviance)/(null deviance). [c]R^2 based on a simple model that only included $PPI_{d,r}$ in the model for $q_{d,r}$. [d]ΔR^2 = (Rep.-level R^2) − (R^2 for PPI only). [e]Each covariate was individually eliminated, producing a 'reduced' model, to compute its relative importance as partial R^2 = (deviance$_{reduced}$ − deviance$_{full}$)/deviance$_{reduced}$, where 'deviance' is the residual deviance.

significant predictors of *a* and/or *b* (Supporting Information, section S5 and Table S1), which describe the relationship between the error rates and PPI.

How do the factors of balance (*PM* and *M*), *N* and *K* affect the relationship between FWER (or Type 1 error rate) and PPI? In the final models (Table S1, Supporting Information), the PPI-intercept (*a*) gives the predicted error rate at PPI = 0 (no pooling), and the PPI-slope (*b*) describes how 'quickly' these error rates change (here, decline) with increasing PPI. An unbalanced dataset (*M* = 1) increases the intercept and steepens the slope (Table S1, Supporting Information, $a_4 > 0$ and $b_4 < 0$; Figure 3a,c, and Figure S4, Supporting Information). A larger group size (*K*) increases the intercept, but *K* has a minimal effect on the slope ($a_1 > 0$ and $b_1 \cong 0$, Table 2; Figure S4B, Supporting Information). While *PM* and *N* significantly affect the slope and/or intercept either as main effects or via interactions with other factors (Table S1, Supporting Information), their overall effect on FWER is minimal (Figure S4A,C, Supporting Information). In general, FWER is primarily elevated by unbalanced designs, regardless of the degree of imbalance, and larger *K*, which determines the number of pairwise comparisons (Figure S4, Supporting Information).

Given the predictable relationship between FWER and PPI (Table S1, Supporting Information), we solved for the critical PPI (PPI_{crit}) leading to FWER = α = 0.05. PPI_{crit} provides an indication of when one might be concerned about adjusting for multiple comparisons; if the HB analysis produces PPI > PPI_{crit}, then the partial pooling effect would negate the need for multiplicity adjustments. Based on the coefficient estimates in Table S1, Supporting Information, we computed PPI_{crit} for combinations of *K*, *N*, *PM* and *M* used to simulate the pseudo data (Figure 4). For balanced datasets, PPI_{crit} spanned 0.499 (*K* = 5, *N* = 10,000) to 0.639 (*K* = 20, *N* = 10,000); for unbalanced datasets (3%–64% missing), PPI_{crit} spanned 0.568 (3% missing, *K* = 5, *N* = 100) to 0.653 (3% missing, *K* = 20, *N* = 10,000). In general, unbalanced designs, larger *N* and larger *K* lead to higher PPI_{crit} values (Figure 4).

While the Type 2 error rate and FWER2 were generally the same for the HB and non-hierarchical models (Figure S3, Supporting Information), they do vary with PPI (Figure 3b,d). The final binomial regressions for these error rates follow the same form as those for the Type 1 error rate and FWER (see Supporting Information section S5). The final models explained 93.5% (or 97.1%) and 45.3% (or 80.5%) of the variation in replicate-level (or scenario-level) Type 2 error rates and FWER2, respectively (Table 2). PPI was the best predictor of the Type 2 error rate (partial R^2 = 90.5%), followed by *M* (partial R^2 = 31.3%); conversely, log(*K*) was the best predictor of FWER2 (partial R^2 = 25.6%), followed by PPI (partial R^2 = 17.8%) (Table 2).

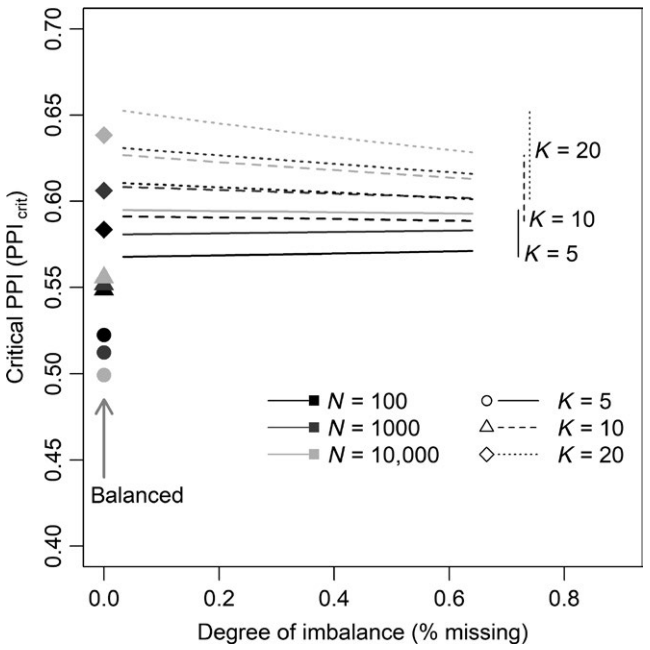

FIGURE 4 Predicted partial pooling index (PPI) leading to a family-wise error rate (FWER) of α = 0.05 (PPI_{crit}). The coefficient estimates in Table S1 were used to compute PPI_{crit} = (logit(α) − *a*)/*b* for different sample sizes (*N*), group sizes (*K*) and degree of imbalance (% missing) that reflect values used in the simulation experiment (Figure S1). The binary balance indicator (*M*) was a significant predictor of FWER, leading to a discontinuous (step function) relationship between PPI_{crit} and the degree of imbalance

3.3 | Application to real-data examples

The PPI and error rates computed for the four real-data examples generally align with the patterns produced by the simulation experiment (Figures 2 and 3). The real-data PPI values ranged from 0.11 (orange trees) to 0.61 (dogs learning), and all four examples yielded low Type 1 error rates (<0.05; Figure 3a). The orange trees and dog learning examples represent balanced designs and their Type 1 and FWER values fall within the range of values estimated for the balanced simulation experiments (Figure 3a,c). The allometric scaling example yielded a comparatively high FWER given its associated PPI (Figure 3c), but this is expected given its large number of comparisons (4,704, Table 1). Conversely, the plant water stress example produced Type 1 error rates and FWER exactly equal to zero (Figures 3a and 4c), but this example was also associated with a small number of comparisons (7). The Type 2 error rate versus PPI predictions for each example also follow the simulation experiment results (Figure 3b). All four examples produced FWER2 values exactly equal to one (Figure 3d), which is not surprising for the three examples that yielded large numbers of comparisons.

4 | DISCUSSION

4.1 | Quantifying the degree of partial pooling

While the terms 'partial pooling' and 'borrowing of strength' are often used when referring to HB models (Carlin & Louis, 2008; Gelman, Carlin, et al., 2014; Qian, Cuffney, Alameddine, McMahon, & Reckhow, 2010), quantitative methods for defining these attributes are generally lacking. We offer the PPI as a quantitative measure of the pooling strength. The ratio of within- to among-group variation (e.g. s/σ) has been informally used an index of potential pooling strength (e.g. Gelman & Hill, 2007), where higher s/σ (comparatively little variability among group levels) is expected to lead to stronger pooling. Higher s/σ was in fact associated with higher PPI (stronger pooling) (Figure 2). However, PPI is not solely determined by s/σ; sample size (N) and the distribution from which the group-level means arise (e.g. normal, uniform) also notably affected PPI. The combined effects of s/σ and N on the degree of partial pooling are qualitatively discussed in Gelman and Hill (2007). That is, larger N is expected to result in more information to inform each group-level effect, which should reflect the group-level sample means. However, if N is relatively small, then pooling strength is primarily governed by s/σ.

The effective number of parameters (p) used to compute model comparison indices – such as DIC (Spiegelhalter et al., 2002) or WAIC (Gelman, Carlin, et al., 2014 Gelman, Hwang, et al., 2014) – has also been used in informal assessments of pooling strength (e.g. Gelman, Hwang, et al., 2014; Plummer, 2008). Motivated by this, we use p based on WAIC to compute PPI (Equation 2). In practice, computing PPI requires implementing three Bayesian models: (1) the focal HB model (Figure 1a), (2) a non-hierarchical version (Figure 1b) and (3) a complete pooling version (Figure 1c). While we successfully implemented all three models with our simulated data, issues may be encountered in real-data applications (see below). However, if all three can be successfully implemented, then the calculation of PPI should be possible. Inferences about effects, parameters and pairwise differences, however, would likely be limited to the HB model.

4.2 | Rejection/error rates and the degree of partial pooling

Based on our intuition and Gelman et al. (2012), we would expect rejection rates – Type 1 error rate, FWER and Power – to generally be lower under an HB versus non-hierarchical model, due to the effect of partial pooling. This was supported by our simulation experiment, especially for FWER (Figure S3C, Supporting Information). However, specification of an HB model does not imply that the Type 1 error rate or FWER will be less than or equal to a set significance level (α). In fact, the scenario-level Type 1 error rate and FWER were greater than $\alpha = 0.05$ in *c.* 60% and *c.* 73%, respectively, of the 2,016 scenarios. Thus, in general, HB models are not immune to inflated rejection error rates associated with multiple comparisons.

However, both the Type 1 error rate and FWER are strongly related to PPI (Figure 3 and Table 2): as pooling strength (PPI) increases, both error rates quickly drop (Figures 3 and S4). The maximum error rates and the greatest variation in the error rates occurred at PPI = 0 (no pooling), and this variation is primarily controlled by measures of imbalance (M) and group size (K) (Table 2; Figures 3a,c, and 4). In the context of multiple comparisons, we are most interested in ensuring that FWER ≤ α, and there appears to be a critical PPI (PPI_{crit}) at which this is achieved. Based on our simulation experiment, PPI_{crit} spanned a relatively narrow range, from *c.* 0.5 to 0.64 (Figure 4). Under a balanced design with small K (comparatively few pairwise comparisons), PPI_{crit} is relatively low ($PPI_{crit} < 0.53$, Figure 4), indicating that even weak pooling can achieve FWER ≤ α. Moreover, PPI_{crit} was constrained between *c.* 0.57 and 0.65 (Figure 4) for unbalanced designs, regardless of the degree of imbalance (or percent missing data). This narrow range provides a useful diagnostic for determining if the partial pooling of an HB model is sufficiently strong (PPI greater than, say, 0.65) to render multiplicity adjustments unnecessary.

One may also be interested in the Type 2 error rate and its family-wise analogue (FWER2); both varied predictably (increased) with PPI (Figure 3 and Table 2). The relationship between these error rates and PPI was primarily governed by K and N, and measures of imbalance (M and PM) exerted comparatively little influence (Table 2). We expected these false acceptance rates to increase with PPI, indicating that with greater pooling strength, we tend to accept H_o: $D_{j,k} = 0$, even when H_o is false. That is, the (estimated) difference becomes indistinguishable from zero as the (estimated) group-level means, μ_j and μ_i, are pulled more towards each other.

4.3 | Advantages of a hierarchical model

Implementing an HB model is advantageous for reducing false rejection rates (Figures S3A and S3C, Supporting Information), especially as PPI increases (Figure 3). In fact, FWER is notably reduced under an HB compared to a non-hierarchical model, across the entire range of potential PPI values (Figure S3C), supporting the expectation that the partial pooling effect of an HB model reduces FWER. If one is concerned about reducing the false acceptance rate, an HB model provides a slight disadvantage by yielding slightly higher Type 2 error rates and lower Power (for individual comparisons) compared to a non-hierarchical model (Figure S3B). However, these differences disappear when considering the family-wise version (FWER2, Figure S3D). Collectively, these results suggest that an HB model is generally advantageous over a non-hierarchical Bayesian model, and presumably over a frequentist analysis (e.g., Gelman et al., 2012), because of its effect on false rejection rates.

4.4 | Hierarchical Bayesian models and multiple comparisons

Our simulation results indicate that HB models are not necessarily immune to problems of inflated family-wise error rates. The partial pooling property of an HB model can lead to FWER values that satisfy a set α level, but the pooling strength required to do so can vary among datasets (Figure 4). However, in general, it appears that HB models yielding PPI ≥ 0.65 are likely to achieve FWER < 0.05, assuming the group size is generally $K \leq 20$, the maximum explored in our simulations. While we cannot provide a specific, quantitative rule that is applicable to all analyses, one could use our predictions of PPI_{crit} (Figure 4) as a general guide for determining if inflated FWER could be a problem.

4.5 | Application to real-data examples

The application of our proposed PPI to four real-life examples provides further support for the utility of PPI. These examples differ greatly in their data and model structures relative to each other and to the simulation experiment (Table 1), yet their error rates versus PPI relationships were consistent with the simulation experiment results (Figure 3). Moreover, all four examples yielded PPI values (based on the real data) that were generally less than the predicted PPI_{crit} (Figure 4), which aligns with FWER values that were generally less than $\alpha = 0.05$ (Figure 3c).

Differences between balanced (dog learning and orange trees) versus unbalanced (allometric scaling and plant water stress) datasets were preserved (Figure 3), as was the effect of group size. In particular, the allometric scaling and dog learning examples supported large group sizes that yielded 4,704 and 870 pairwise comparisons, respectively, many more than the maximum (190) considered in the simulation experiment. The allometric scaling example's large group size clearly impacted the FWER (Figure 3c), and the high FWER is consistent with the predictions generated by the simulation experiment for large K (Figure S4B, Supporting Information). In practice, however, it is probably unlikely that one would compare multiple parameters among all 49 plant species – this was not a goal of the original study (Price, Enquist, & Savage, 2007; Price et al., 2009) – but this example is valuable for demonstrating the generality of the simulation results.

We explored the application of the PPI to two other real-data examples, but they proved problematic because the non-hierarchical models did not converge. Our experience suggests that such mixing and convergence issues are especially likely to occur if some group levels are associated with small sample sizes (small n_k), combined with a large within-group variance (σ^2). These issues tend to disappear when the group-level effects are modelled hierarchically. Thus, computation of PPI for real-data applications will depend on whether or not the non-hierarchical version can be successfully implemented.

4.6 | Other considerations

It is likely that our approach of using posterior coverage probabilities – that is, evaluate if the 95% CI for a difference contains zero – to evaluate different hypotheses may correspond to an equivalent selection procedure using a loss function within a Bayesian decision theoretic framework (Berger, 1985). While it may be useful to introduce a more formal procedure, our approach is commonly and easily employed by ecologists. Likewise, some may consider alternative methods that offer an explicit accounting of model (hypotheses) uncertainty, such as Bayesian model averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999; Raftery, Madigan, & Hoeting, 1997) or transdimensional methods that accommodate changing parameter dimension across models (Sisson, 2005). However, these procedures obscure parameter interpretation (Banner & Higgs, 2017; Cade, 2015), which we are generally interested in maintaining. The potential utility of using formal decision theoretic or model selection approaches in place of a more traditional multiple comparisons procedure is just one of several potential solutions. Other solutions have been suggested in the context of Bayesian models (e.g., Li & Shang, 2015; Nashimoto & Wright, 2008; Shang et al., 2008; Westfall et al., 1997), but as noted previously, they introduce additional concerns. We emphasize that our goal is to lend insight into whether or not HB models are immune to multiplicity issues – they are not – and to offer PPI as a tool for evaluating when alternative modelling approaches or adjustment procedures should be explored.

AUTHORS' CONTRIBUTIONS

K.O. conceived of the study and performed the post-analysis of the simulation output; D.P., M.F., J.G., H.K. and J.B. co-conceived of the study; D.P., M.F., J.G. and H.K. wrote code for simulating data, implementing models and computing PPI and error rates; J.B. assisted with the post-analysis of the simulation output. K.O. wrote the manuscript. All authors contributed to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

Code and example pseudo datasets generated in the simulation experiment are provided through GitHub (https://github.com/kropheather/multiple_comp_ex/tree/v1.0) (https://doi.org/10.5281/zenodo.1244289) as well as the four real-data examples and associated code for implementing the Bayesian models and computing PPI (https://github.com/kionaogle/BayesianMultipleComparisons) (https://doi.org/10.5281/zenodo.2207449). For more details, see Supporting Information (sections S2 and S4).

ORCID

Kiona Ogle https://orcid.org/0000-0002-0652-8397
Drew Peltier https://orcid.org/0000-0003-3271-9055

REFERENCES

Banner, K. M., & Higgs, M. D. (2017). Considerations for assessing model averaging of regression coefficients. *Ecological Applications*, *27*, 78–93. https://doi.org/10.1002/eap.1419

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*, 2nd ed. New York, NY: Springer. https://doi.org/10.1007/978-1-4757-4286-2

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Cade, B. S. (2015). Model averaging and muddled multimodel inferences. *Ecology*, *96*, 2370–2382. https://doi.org/10.1890/14-1639.1

Carlin, B. R., & Louis, T. A. (2008). *Bayesian methods for data analysis*, 3rd ed. Boca Raton, FL: CRC Press.

Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecology Letters*, *8*, 2–14.

Clark, J. S., Bell, D., Chu, C. J., Courbaud, B., Dietze, M., Hersh, M., … Wyckoff, P. (2010). High-dimensional coexistence based on individual variation: A synthesis of evidence. *Ecological Monographs*, *80*, 569–608. https://doi.org/10.1890/09-1541.1

Clark, J. S., & Gelfand, A. E. (2006). A future for models and data in environmental science. *Trends in Ecology & Evolution*, *21*, 375–380. https://doi.org/10.1016/j.tree.2006.03.016

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*, 3rd ed. Boca Raton, FL: CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*, 189–211. https://doi.org/10.1080/19345747.2011.618213

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 997–1016. https://doi.org/10.1007/s11222-013-9416-2

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

Guo, J. S., & Ogle, K. (2018). Antecedent soil water content and vapor pressure deficit interactively control water potential in *Larrea tridentata*. *New Phytologist*, *221*, 218–232. https://doi.org/10.1111/nph.15374

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–401.

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, *85*, 3–28. https://doi.org/10.1890/14-0661.1

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*, 5th ed. Boston, MA: McGraw-Hill Irwin.

Li, Q., & Shang, J. F. (2015). A Bayesian hierarchical model for multiple comparisons in mixed models. *Communications in Statistics-Theory and Methods*, *44*, 5071–5090. https://doi.org/10.1080/03610926.2013.813042

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine*, *28*, 3049–3082. https://doi.org/10.1002/sim.3680

McMahon, S. M., & Diez, J. M. (2007). Scales of association: Hierarchical linear models and the measurement of ecological systems. *Ecology Letters*, *10*, 437–452. https://doi.org/10.1111/j.1461-0248.2007.01036.x

Nashimoto, K., & Wright, F. T. (2008). Bayesian multiple comparisons of simply ordered means using priors with a point mass. *Computational Statistics & Data Analysis*, *52*, 5143–5153. https://doi.org/10.1016/j.csda.2008.05.015

Ogle, K., & Barber, J. J. (2008). Bayesian data-model integration in plant physiological and ecosystem ecology. *Progress in Botany*, *69*, 281–311. https://doi.org/10.1007/978-3-540-72954-9

Peltier, D. M. P., Fell, M., & Ogle, K. (2016). Legacy effects of drought in the southwestern United States: A multi-species synthesis. *Ecological Monographs*, *86*, 312–326. https://doi.org/10.1002/ecm.1219

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In H. Kurt, L. Friedrich & Z. Achim, (Ed.). *Proceedings of the 3rd international workshop on distributed statistical computing* (pp. 125). Vienna, Austria.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, *9*, 523–539. https://doi.org/10.1093/biostatistics/kxm049

Plummer, M. (2013). *Rjags: Bayesian graphical models using MCMC*. R Package Version 3–10. CRAN. Retrieved from http://CRAN.R-project.org/ package=rjags

Plummer, M. (2015). JAGS version 4.0.0 user manual. pp. 43. Retrieved from: https://sourceforge.net/projects/mcmc-jags/files/Manuals/, SourceForge

Plummer, M., Best, N. G., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*, 7–11.

Price, C. A., Enquist, B. J., & Savage, V. M. (2007). A general model for allometric covariation in botanical form and function. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 13204–13209. https://doi.org/10.1073/pnas.0702242104

Price, C. A., Ogle, K., White, E. P., & Weitz, J. S. (2009). Evaluating scaling models in biology using hierarchical Bayesian approaches. *Ecology Letters*, *12*, 641–651. https://doi.org/10.1111/j.1461-0248.2009.01316.x

Qian, S. S., Cuffney, T. F., Alameddine, I., McMahon, G., & Reckhow, K. H. (2010). On the application of multilevel modeling in environmental and ecological studies. *Ecology*, *91*, 355–361. https://doi.org/10.1890/09-1043.1

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: https://www.R-project.org/

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*, 179–191. https://doi.org/10.1080/01621459.1997.10473615

Shang, J. F., Cavanaugh, J. E., & Wright, F. T. (2008). A Bayesian multiple comparison procedure for order-restricted mixed models. *International Statistical Review*, *76*, 268–284. https://doi.org/10.1111/j.1751-5823.2008.00051.x

Sisson, S. A. (2005). Transdimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association*, *100*, 1077–1089. https://doi.org/10.1198/016214505000000664

Spiegelhalter, D. J., Best, N. G., Carlin, B. R., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *64*, 583–616. https://doi.org/10.1111/1467-9868.00353

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, *84*, 419–427. https://doi.org/10.1093/biomet/84.2.419

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Ogle K, Peltier D, Fell M, Guo J, Kropp H, Barber J. Should we be concerned about multiple comparisons in hierarchical Bayesian models? *Methods Ecol Evol*. 2019;10:553–564. https://doi.org/10.1111/2041-210X.13139