Ecological Applications, 30(7), 2020, e02159

Ensuring identifiability in hierarchical mixed effects Bayesian models

Kiona Ogle [1], and Jarrett J. Barber

School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, Arizona 86011 USA

Citation: Ogle, K., and J. J. Barber. 2020. Ensuring identifiability in hierarchical mixed effects Bayesian models. Ecological Applications 30(7):e02159. 10.1002/eap.2159

Abstract. Ecologists are increasingly familiar with Bayesian statistical modeling and its associated Markov chain Monte Carlo (MCMC) methodology to infer about or to discover interesting effects in data. The complexity of ecological data often suggests implementation of (statistical) models with a commensurately rich structure of effects, including crossed or nested (i.e., hierarchical or multi-level) structures of fixed and/or random effects. Yet, our experience suggests that most ecologists are not familiar with subtle but important problems that often arise with such models and with their implementation in popular software. Of foremost consideration for us is the notion of effect identifiability, which generally concerns how well data, models, or implementation approaches inform about, i.e., identify, quantities of interest. In this paper, we focus on implementation pitfalls that potentially misinform subsequent inference, despite otherwise informative data and models. We illustrate the aforementioned issues using random effects regressions on synthetic data. We show how to diagnose identifiability issues and how to remediate these issues with model reparameterization and computational and/or coding practices in popular software, with a focus on JAGS, OpenBUGS, and Stan. We also show how these solutions can be extended to more complex models involving multiple groups of nested, crossed, additive, or multiplicative effects, for models involving random and/or fixed effects. Finally, we provide example code (JAGS/OpenBUGS and Stan) that practitioners can modify and use for their own applications.

Key words: crossed effects; equifinality; fixed effects; hierarchical model; identifiability; MCMC; multi-level model; nested effects; prior distribution; random effects; sum-to-zero; sweeping.

Introduction

Complex data often suggest models with crossed or nested (hierarchical or multi-level) structures of fixed or random effects. Ecological analyses of such data are increasingly common (Fig. 1), including, in particular, the use of Bayesian models and associated Markov chain Monte Carlo (MCMC) methodology for implementing such models (Ellison 2004, Clark 2007, McCarthy 2007, Ogle and Barber 2008, Hobbs and Hooten 2015, Dorazio 2016, Touchon and McCoy 2016). Implementation of Bayesian methods has been facilitated by popular and fairly user-friendly software (e.g., Kruschke 2014, McElreath 2016, Monnahan et al. 2017), such as JAGS (Plummer 2003, 2012), WinBUGS or OpenBUGS (Lunn et al. 2000, Lunn et al. 2009), and Stan (Stan Development Team 2018, Carpenter et al. 2017). Yet, our experience also suggests that ecologists are relatively unfamiliar with subtle and important identifiability problems that often arise with the implementation of such models (Gelfand and Sahu 1999, Gelman 2004, Gelman and Hill 2007, Hines et al. 2014). These problems can lead to poor mixing and convergence behavior of the numerical sampling algorithm, potentially producing biased parameter estimates. In turn, this can mislead inference about interesting quantities (Raue et al. 2013), despite otherwise reasonable models and informative data. Here, we focus on the diagnosis and remediation of identifiability problems that can arise during the numerical implementation of seemingly reasonable hierarchical Bayesian models.

While a Bayesian model may be relatively straightforward to specify, its implementation is more subtle, with potential pitfalls that can mislead inference about effects. In particular, as alluded to above, implementation may lead to identifiability problems (e.g., Omlin and Reichert 1999, Rannala 2002, Gelman 2004, Gelman and Hill 2007, Raue et al. 2013, Holand and Steinsland 2016). In the strict sense, non-identifiability of parameters refers to a constancy in the posterior probability or likelihood with changes in the parameters (e.g., Raue et al. 2013), but we broadly consider (non) identifiability as the (in)ability of models, data, or implementations to inform about effects of interest. For example, a model may be over-parameterized, whereby a change in one parameter compensates exactly for the change in posterior probability or likelihood caused by a change in another parameter (e.g., Rannala 2002, Swartz et al. 2004, Raue et al. 2013); thus, such parameters are strictly non-identifiable (Casella and Berger

Manuscript received 30 August 2019; revised 29 January 2020; accepted 17 March 2020. Corresponding Editor: Jarno Vanhatalo.
[1]E-mail: kiona.ogle@nau.edu

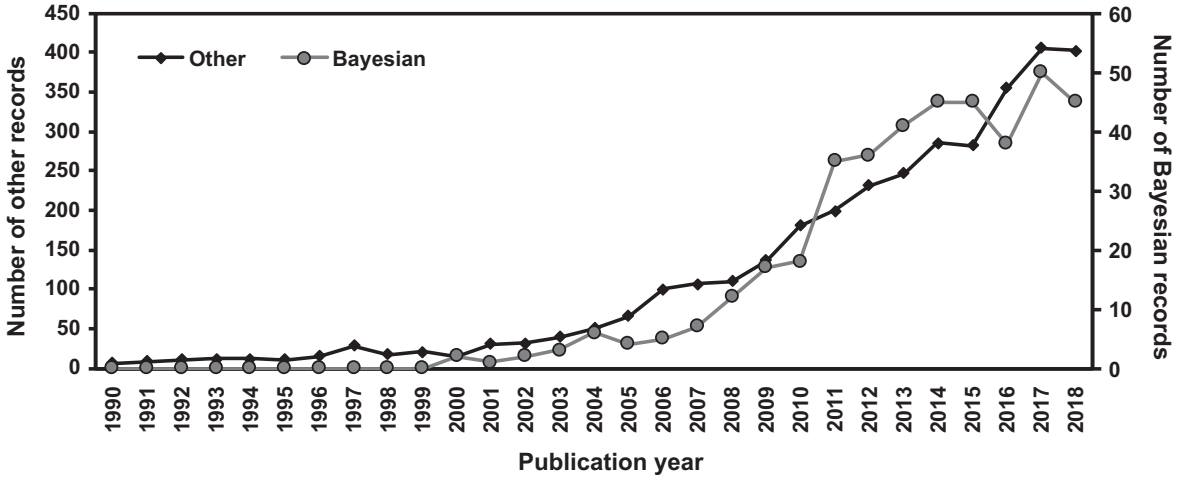

FIG. 1. Web of Science search results (search conducted 20 June 2019) illustrate the rise in popularity of random and mixed effects models in ecological data analysis, especially Bayesian methods. Search keywords = ((random NEAR/1 effect*) OR (mixed NEAR/1 effect*) OR (mixed NEAR/1 model*)), limited to the period 1990–2018 and refined by "Web of Science Categories" = "Ecology," yielded 3,829 total records (gray circles + black diamonds). The search was repeated to partition the 3,829 records into those that included "Bayes*" in the topic or keywords, yielding 412 publications (gray circles) that we presume to have employed Bayesian methods closely related to those discussed herein; the remaining 3,417 "other" records (black diamonds) likely employed non-Bayesian methods or did not explicitly use Bayesian-related terms. It is worth noting that over 50% of the Bayesian-focused publications occurred during the last five years (2014–2018).

2002). However, we consider mostly cases wherein parameters are "weakly" identifiable (e.g., Gelfand and Sahu 1999, Gimenez et al. 2009), wherein the parameters are correlated within an MCMC chain; that is, one parameter's values tend to track another's values with little change to posterior probabilities or likelihoods (Rannala 2002, Carlin and Louis 2009). As a simple example, if a linear model is specified with a random effect that is added to an overall intercept term, a jump in the intercept by one unit may be compensated for by a jump in one unit, in the opposite direction, of the random effect (see the specific example associated with Eq. [1]).

In this paper, we specifically focus on cases that arise largely from model implementation despite otherwise reasonable models and data. Typical frequentist implementations or software packages solve such identifiability problems by implementing constraints within the analysis and software. For example, the lm and glm functions for fitting linear and generalized linear models, respectively, in R, employ "treatment coding" (also referred to as "cell reference coding" or "treatment contrasts") for fixed effects, whereby the effect associated with a factor's first level is constrained to zero (e.g., $\theta_1 = 0$). This default coding achieves identifiability of the fixed effects associated with a factor, and, incidentally, results in a particular interpretation of effects parameters whereby the factor's first level is interpreted as a reference level. Alternatively and commonly, constraining the factor level effects to sum to zero achieves identifiability ("sum-to-zero coding" or "sum-to-zero contrasts") and a different interpretation of parameters. These solutions partly motivate the implementation of similar constraints within a Bayesian model.

While we have been aware of the aforementioned identifiability problems, in our own work and from working with fellow ecologists as well as from the statistical literature and colleagues, we are unable to recommend a single reference that addresses these problems in a concise manner, accessible to ecologists. Together, these issues motivate our current article, which collects results from the broader scientific literature, tempered by our experiences working with our own data and with our ecological colleagues. In the course of our discussion, we review common concepts and terminology, primarily associated with linear mixed models, and offer advice for more general situations. We use simulation experiments to illustrate issues and demonstrate solutions. In doing so, we wish to make ecologists aware of important identifiability issues associated with implementing hierarchical Bayesian models involving random, fixed, or mixed effects, and existing methods for addressing these issues. Thus, this article is aimed at ecologists that have some experience implementing, or anticipate implementing, hierarchical or multi-level Bayesian models.

A Bayesian Perspective on Fixed Vs. Random Effects

Whether frequentist or Bayesian, the essential statistical nature of random effects stems from their specification as arising from a common probability distribution whose parameters, to be estimated in some manner, often, but not always, include just a single variance parameter (Kutner et al. 2004, Gamerman and Lopes

2006, Gelman and Hill 2007, Ramsey and Schafer 2013, Gelman et al. 2014). In the context of random effects, the units or group levels, i.e., experimental units, observational units, individuals, subjects, etc., such as a subset of trees randomly selected from a multi-hectare plot, are often viewed as exchangeable (Draper et al. 1993, O'Neill 2009). In particular, the observed units (e.g., trees) are typically treated as conditionally independent, arising from a common probability distribution described by (conditional on), for example, variance and/or covariance parameters that quantify variability among the units. The assumption of exchangeability and a common distribution often results in shrinkage (Gelman and Hill 2007, Qian et al. 2010, Ogle et al. 2019) of a group of random effects (toward some value, often zero) or, equivalently, partial pooling or borrowing of strength (Gelman and Hill 2007, Carlin and Louis 2009, Qian et al. 2010, Ogle et al. 2019) among a group of effects (e.g., among individual trees, the group levels or units, within the plot). And, the degree of partial pooling among units is related to among unit (within group) variability, with smaller variances allowing for the possibility of stronger pooling (Gelman and Hill 2007, Ogle et al. 2019). The exchangeability assumption allows for inference about individuals (units or levels; e.g., individual trees) via individual-specific (random) effects or about the population from which they came (e.g., the forest represented by the plot) via a common distribution's variance or covariance parameters (e.g., the within-group variance terms). These different levels of inference are often referred to as *individual-based* (or *conditional*) inference or *population-based* (or *marginal*) inference, respectively (in the case of linear statistical models, at least) (Reid 1995, Wakefield 2013: Chapters 8 and 9).

Fixed effects are different. For frequentists, these are fixed quantities to be estimated, with uncertainty being inherited from the specification of a likelihood for the data (e.g., McCulloch and Searle 2001). In many cases, the number of fixed effects levels may be small and chosen for specific reasons, as might occur in a manipulative experiment (e.g., two levels of CO_2: ambient vs. elevated). In our experience, most Bayesians view fixed effects as fixed, too, but characterize uncertainty more directly via probability distributions, which are completely specified a priori (e.g., Gelman and Hill 2007). There is often no notion of a larger unobserved population of units; hence it often does not make sense to estimate population-level variance parameters to characterize variability among such a nonexistent set of units (or levels). Consequently, the notion of exchangeability does not apply to fixed effects and they do not exhibit shrinkage or borrowing of strength, frequentist or Bayesian. As an aside, we acknowledge a connection to shrinkage priors that are often employed to regularize a problem (see Part IV in Wakefield 2013). With random effects, however, we (should) specify an exchangeable prior based on our beliefs, whereby shrinkage and borrowing of strength are a consequence of our beliefs, updated with data via the likelihood, and are not necessarily a means to somehow regularize a problem.

While we can estimate fixed or random effects associated with observed units, with random effects, the scope of inference extends beyond observed units to a population of units, characterized by estimated variance/covariance components or predictions of (random) effects of unobserved units. For fixed effects, inference is generally limited to the observed units (e.g., effect of ambient vs. elevated CO_2 on some response variable of interest). We refer the reader to Gelman (2005) for additional discussion about fixed vs. random effects from frequentist and Bayesian perspectives.

To help make the above notions about fixed and random effects more concrete, let us consider observations y_i ($i = 1, 2, \ldots, N$) for which we specify a probability distribution, conditional upon μ_i, which is modeled as a function of covariates and their effects, and μ_i is linked to the mean of y_i. To illustrate, let the y_i be normally distributed with mean given exactly by μ_i, which we simplify as a linear model of a single covariate, x_i, with its (slope) effect (β_1) and an overall (intercept) effect (β_0), a simple linear regression model, so far. Further, assume observations are obtained for different species, $s = 1, 2, \ldots, S$, across different plots, $p = 1, 2, \ldots, P$. In this context, we consider species and plots to be units or levels for which effects are considered for modeling as fixed or random. It seems reasonable to remodel β_0 and β_1 to reflect our sampling scheme among plots and species. For example, consider the remodeled intercept to reflect *additive main effects* of species and plot: $\beta_{0,s(i)}$ and $\varepsilon_{p(i)}$; and, allow the slope to vary by species: $\beta_{1,s(i)}$. The subscripts $s(i)$ and $p(i)$ indicate species and plot, respectively, associated with observation i. Thus, we write the mean as

$$\mu_i = \beta_{0,s(i)} + \beta_{1,s(i)} x_i + \varepsilon_{p(i)}. \quad (1)$$

Our model may be seen as a traditional analysis of covariance (ANCOVA), with additive main effects of a species factor, with S levels, and a plot factor, with P levels, and species-specific regression covariate effects. We may also say that observations are grouped by species and plots. Species and plots may be completely crossed in the sense that every species occurs in every plot, or vice-versa, but we make no such assumption in what follows. Further, in our example, we consider plots to be sampled from some larger population of plots, suggesting ε_p as random effects, and we consider species effects, $\beta_{0,s}$ and $\beta_{1,s}$, as fixed. Thus, we have a *mixed* model of random and fixed effects. Because μ_i is a function of random effects, it is common to say that the mean is *conditional* on the random variables, ε_p for $p = 1, 2, \ldots, P$, allowing conditional or individual-based (i.e., plot-based) inference via plot effects.

Continuing our example, we adopt the familiar normal specification for the random effects such that we may assume $\varepsilon_p \sim \text{Normal}(0, \sigma^2_\varepsilon)$. This assumes that the

plots are exchangeable, which is a reasonable assumption barring that no particular plot (or group of plots) is associated with an unusually large or small effect (Draper et al. 1993). The variance component, σ_ε, accounts for variability among plots in the population, so called *population-based* or *marginal* inference resulting from marginalizing over the random effects, given σ_ε. The specification also allows for borrowing of strength (Carlin and Louis 2009, Qian et al. 2010, Gelman et al. 2014) among the plots (ε_p) given that they are assumed to arise from a common distribution. In contrast, we would likely treat the species-level effects, $\beta_{0,s}$ and $\beta_{1,s}$, as not arising from a common distribution, and thus, they would not share population-level parameters, such as variance terms. So far, in our example, our terminology and modeling holds, whether frequentist or Bayesian, with the understanding that a Bayesian specification of a mean typically entails (implicitly) conditioning on further quantities, which are, at some level, given their own distributional specifications (e.g., priors). Toward this end, we depart from the frequentist perspective and specify a prior on the plot effects' variance component (σ^2_ε) and the species-level fixed effects, $\beta_{0,s}$ and $\beta_{1,s}$.

How do we pick priors for $\beta_{0,s}$ and $\beta_{1,s}$? First, consider a "generic" coefficient or parameter, θ, that is indexed by different units or levels of a factor, $j = 1, 2, \ldots, J$ (e.g., θ_j could represent a species- or plot-specific effect). Assume that we specify a normal distribution as a prior for this parameter; in doing so, there are three primary specifications that we may choose from:

$$\theta_j \sim \text{Normal}\,(m, v), \text{ with fixed values specified for the prior mean } (m) \text{ and variance } (v) \quad (2)$$

$$\theta_j \sim \text{Normal}\,(m, v), \text{with priors specified for the unknown } m \text{ and } v \quad (3)$$

$$\theta_j \sim \text{Normal}\,(0, v), \text{ with a prior specified for the unknown variance } (v). \quad (4)$$

(We touch on non-normal analogies to Eqs. [2–4] in subsequent sections.) We generally reserve the *prior* defined by Eq. (2) for parameters that we view as fixed effects (e.g., as might be done for treatment-level effects associated with a manipulative experiment), and/or for which the group size is exceptionally small (e.g., $J = 2$ or 3 levels); if we want a fairly non-informative prior, we may set $m = 0$ and $v =$ large value. Hence, in the context of the previous example in Eq. (1), for species-specific parameters, we would likely specify priors for $\beta_{0,s}$ and $\beta_{1,s}$ according to Eq. (2). If, however, there are many species (e.g., $S \gg 3$), then we may choose a prior following Eq. (3), whereby m would describe the mean effect (intercept or slope) across all species, and v the variability among species (e.g., Sauer and Link 2002, Kery and Royle 2008, Price et al. 2009, Zipkin et al. 2009, Ovaskainen and Soininen 2011, Ogle et al. 2013, 2014, Foss-Grant et al. 2016). However, the hierarchical models defined by Eqs. (3 and 4) are typically reserved for random effects (e.g., random site or plot effects). Eq. (4) differs from Eq. (3) in that Eq. (4) assumes a mean of exactly zero; we may refer to Eq. (4) as a *zero-centered hierarchical prior* and Eq. (3) as a *hierarchically centered prior* (e.g., Gelfand et al. 1995, Gilks and Roberts 1996). Returning to the example associated with Eq. (1), we may expect some plots to produce larger than expected values for y ($\varepsilon_p > 0$), and others to produce smaller than expected values ($\varepsilon_p < 0$), but across all plots, the plot effects should be centered on a mean of zero, motivating our choice of Eq. (4) for modeling ε_p.

The Identifiability Problem

We used the previous example to motivate relevant terminology. However, to illustrate identifiability problems that can arise even in simple linear models such as Eq. (1), let us first consider an even simpler model. As before, assume that each y_i arises from a normal distribution, with mean μ_i and variance σ^2, but assume a scalar intercept and slope such that

$$\mu_i = \beta_0 + \beta_1 x_i + \varepsilon_{j(i)}. \quad (5)$$

We interpret β_0 as the overall intercept and ε_j as a random effect for each group level j (for $j = 1, 2, \ldots, J$, and $J < N$), which we model according to Eq. (4): $\varepsilon_j \sim$ Normal(0, σ^2_ε). Assume that relatively non-informative priors are specified for β_0 and β_1, e.g., according to Eq. (2), and that the two variance terms (σ^2 and σ^2_ε) are assigned relatively non-informative, conjugate priors (Gelman et al. 2014, Kruschke 2014) or semi-informative priors that reduce the probability of unrealistically large values (e.g., Gelman 2004, 2006, Lemoine 2019).

While we may be able to obtain analytical solutions for the posterior distributions of the parameters in the above model (e.g., β_0, β_1, ε_1, ε_2, …, ε_J, σ^2, and σ^2_ε), most real applications, however, involve models of greater complexity, for which analytical solutions are not easily derived. Thus, we typically use numerical simulation methods such as Markov chain Monte Carlo (MCMC) to sample from, thus estimating, the joint and marginal posterior distributions of the parameters (Gamerman and Lopes 2006). However, Eq. (5) is useful for illustrating potential issues that plague models of varying complexity. For example, if this model is implemented in a software package such as OpenBUGS (Lunn et al. 2009) or JAGS (Plummer 2003, 2012), then the behavior of the MCMC chains can potentially reveal an underlying identifiability problem (Gelfand et al. 1995, Eberly and Carlin 2000, Gelman and Hill 2007, Hines et al. 2014). We illustrate this via simulations, which we elaborate upon below.

The non-identifiability of groups of random or fixed effects, our focus here, may occur in tandem with the

potential non-identifiability of the slope and intercept in a linear model, which we touch on briefly. For example, upon implementing the model in Eq. (5), it may be difficult to individually estimate β_0 (intercept) and β_1 (slope) if the observed values of x_i are "far from" zero, resulting in potentially strong posterior correlation among β_0 and β_1 (see Fig. 2). To address this problem, centering of x_i about its sample mean ($\overline{x}$) and potentially standardizing by its sample standard deviation (SD), giving $z_i = x_i - \overline{x}$ or $z_i = (x_i - \overline{x})/\text{SD}$, respectively, and regressing y_i on z_i results in a posterior correlation between β_0 and β_1 of approximately zero (Gilks and Roberts 1996, Gelman et al. 2014), enabling us to identify β_0 and β_1 (e.g., Gelfand et al. 1995; Fig. 2). That is, with covariate centering or standardization, the MCMC chains for β_0 and β_1 move independently of each other.

Returning to the issue of the non-identifiability of β_0 and the ε_j's, it is easy to see that we can add a constant to β_0 and subtract the same constant from each ε_j (only one of which contributes to the mean, μ_i, for observation *i*), thus resulting in the same value of μ_i, and, thus, the same likelihood value and posterior density. In this sense, β_0 and ε_j are not identifiable or not individually "estimable," and the mean is over-parameterized (Carlin and Louis 2009). We illustrate this identifiability problem with synthetic data based on the model in Eq. (5); in our synthetic data, $\overline{x}$ is close to zero, so covariate centering is not required (see Appendix S1: Fig. S1). We then fit the model defined by Eq. (5) to the synthetic data using JAGS, with standard and relatively non-informative priors for β_0 (intercept), β_1 (slope or x coefficient), σ^2 (measurement error variance), and σ^2_ε (random effects variance). The synthetic data and code (R and JAGS) are provided in Appendix S1 (Sections S1 and S2).

The simulation experiment demonstrates that when the random effects variance is small relative the measurement error variance (i.e., for true values of $\sigma = 1$ and $\sigma_\varepsilon = \sigma/10$), the MCMC chains exhibit "text book" behavior by showing excellent mixing and convergence for all quantities monitored (see Fig. 3A, D, G, and J for β_0, β_1, $\overline{\varepsilon}$, and one of the ε_j, respectively; where $\overline{\varepsilon}$ is the average of the ε_j's). In this case, Eq. (4) acts like an informative prior for the ε_j, such that the ε_j are estimated to be close to the prior mean of zero, again, reflecting strong borrowing of strength or shrinkage toward zero (Gelman and Hill 2007), and leading to identifiability of β_0 and the ε_j. Here, β_0 and $\overline{\varepsilon}$ (or ε_j) are only moderately correlated (Fig. 4A, D).

When the two variance terms are of similar magnitude (i.e., for true $\sigma_\varepsilon = \sigma$), the chains for some of the parameters (e.g., β_1; Fig. 3E), show similar mixing behavior as described above. The chains for β_0, $\overline{\varepsilon}$, and individual ε_j's, however, exhibit greater within chain autocorrelation, but they still converge rather quickly (see Fig. 3B, H, K). In this case, Eq. (4) acts like a moderately informative prior, and the borrowing of strength among the ε_j is somewhat weaker. For the scenario involving a large random effects variance (true $\sigma_\varepsilon = 10\sigma$), the chains for β_1 behave similar to the first two scenarios (Fig. 3F), but the chains for β_0, $\overline{\varepsilon}$, and individual ε_j's exhibit extremely poor mixing and do not converge after 5,000 iterations (Fig. 3C, I, L), despite model simplicity and otherwise no apparent problem indicated by the model specification. In fact, the chains for β_0 (Fig. 3C) look like mirror images of the $\overline{\varepsilon}$ chains (Fig. 3I). This trade-off between β_0 and $\overline{\varepsilon}$ (or ε_j) is revealed in the bivariate scatter plot (Fig. 4C, F) whereby the MCMC samples for β_0 and $\overline{\varepsilon}$ are nearly perfectly negatively correlated. That is, when σ_ε is very large, Eq. (4) acts like a non-informative prior for the ε_j, with very little to no borrowing of strength, thus allowing the MCMC chains for the ε_j to move away from the prior mean of zero. This latter scenario illustrates our broader perspective of near non-identifiability: changes in one quantity (e.g., β_0) are compensated by changes in another (e.g., $\overline{\varepsilon}$ and/or individual ε_j), while their sum (e.g., $\beta_0 + \overline{\varepsilon}$, the "overall" intercept; Fig. 3M, N, O) and the posterior probability remains relatively unchanged. Nearly identical results were obtained when the models were implemented in OpenBUGS (see Appendix S1: Fig. S2 and Table S1).

The potential non-identifiability of β_0 and the ε_j is supported by analysis of a simpler model only involving an overall intercept plus a random effect (i.e., no covariate effect) (Gelfand et al. 1995, Gilks and Roberts 1996), with a flat prior on the overall intercept; under this model, the correlation between β_0 and any particular ε_j is

$$\rho_{\beta_0,\varepsilon_j} = -\left(1 + \frac{N\sigma^2}{J\sigma^2_\varepsilon}\right)^{-\frac{1}{2}}. \quad (6)$$

If, for example, the sample size $N = 100$ and the group size $J = 10$ (as in the above simulations; e.g., Figs. 3, 4), this correlation, Eq. (6), is close to zero (never positive) if $\sigma_\varepsilon \ll \sigma$, and it approaches -1.0 as σ_ε approaches 10σ (see Appendix S1: Fig. S3). This is consistent with the bivariate plots in Fig. 4, which also suggest that the correlation is even stronger between β_0 and $\overline{\varepsilon}$ (compared to individual ε_j).

So, why does moderate to strong correlation among parameters (e.g., β_0 and ε_j) lead to poor mixing and/or near non-identifiability? Many MCMC sampling algorithms, such as most of the univariate algorithms in JAGS or OpenBUGS, move through the posterior parameter space by taking steps in the direction of each parameter's coordinate axes, one parameter at a time, to a new coordinate value that is associated with a somewhat minor change in the posterior density. Correlation and/or near non-identifiability among parameters cause long, narrow regions or "ridges" in the parameter space wherein the posterior is concentrated (Omlin and Reichert 1999, Swartz et al. 2004, Hines et al. 2014). Thus, moving too much in a coordinate axes direction can quickly send the sampling algorithm up/down a steep posterior cliff, and sampling steps are made small to avoid this, which is revealed in chains that move slowly

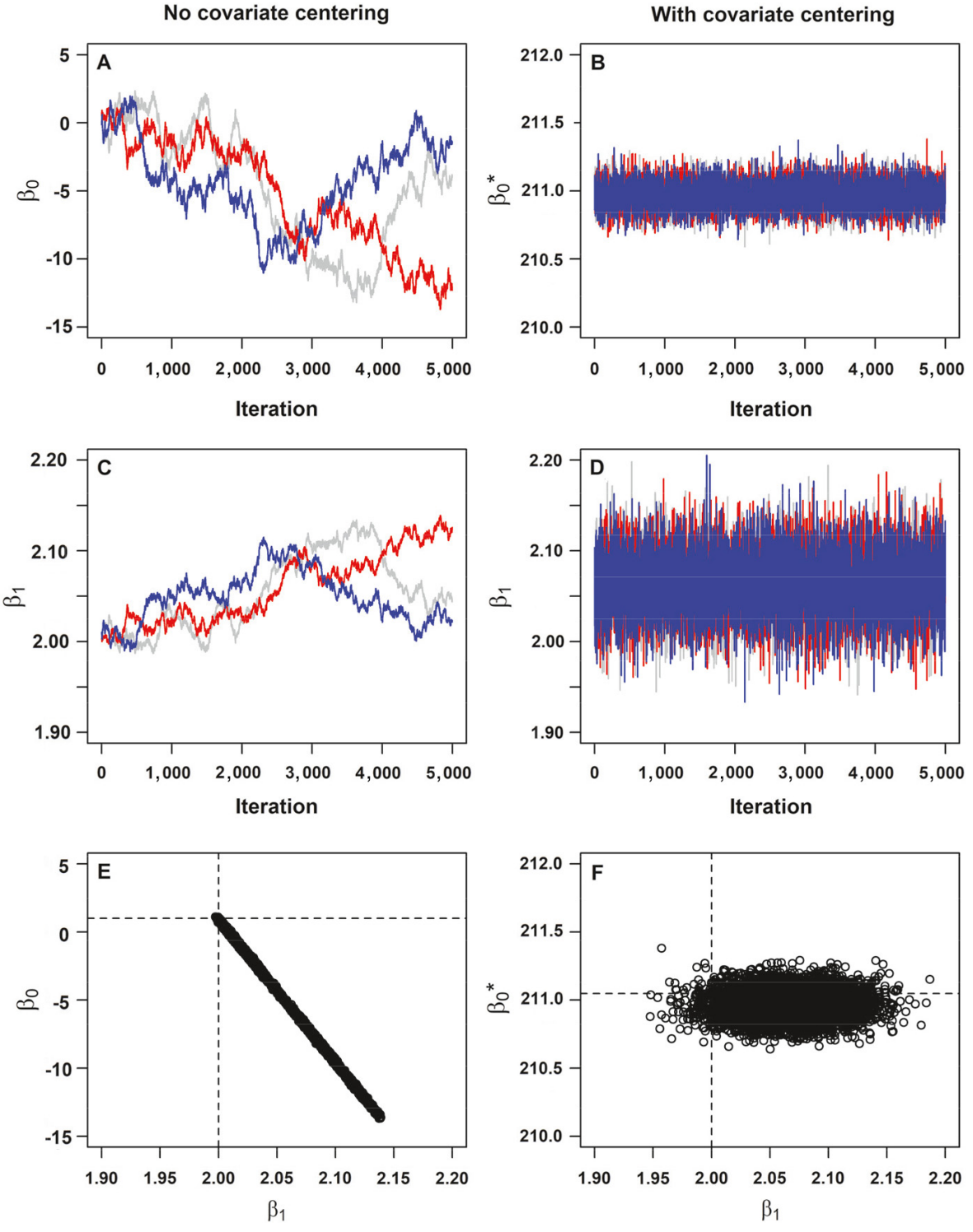

FIG. 2. (A) Data were simulated from $y_i \sim \text{Normal}(\mu_i, \sigma^2)$ for $i = 1, 2, \ldots, 100$, $\mu_i = \beta_0 + \beta_1 x_i$, and $x_i \sim \text{Uniform}(100, 110)$, with $\beta_0 = 1$, $\beta_1 = 2$, and $\sigma = 1$. A Bayesian model was applied to the synthetic data with priors $\beta_0, \beta_1 \sim \text{Normal}(0, 1 \times 10^6)$ and $\sigma^{-2} \sim \text{Gamma}(0.1, 0.1)$ using (A, C, E) the original x data with $\mu_i = \beta_0 + \beta_1 x_i$ or (B, D, F) centered x data with $\mu_i = \beta_0^* + \beta_1(x_i - \bar{x})$; note, in the covariate-centered version, the original intercept is computed as $\beta_0 = \beta_0^* - \beta_1 \bar{x}$. Without covariate centering, the Markov chain Monte Carlo (MCMC) chains for (A) the intercept (β_0) and (C) the slope (β_1) show poor mixing

(Fig. 2. *Continued*)

and (E) are highly correlated ($r \cong -1.000$); due to high within chain autocorrelation (see A and C), the Raftery and Lewis (1996) diagnostic indicates that over 635,000 MCMC samples are required for accurate 95% credible intervals (CIs). Covariate centering greatly improves mixing of (B) the "new" intercept (β_0^*) and (D) β_1 and results in (F) uncorrelated posterior samples of β_0^* and β_1 ($r \cong 0.005$), and only requires ~ 3,800 MCMC samples (due to lack of within chain autocorrelation, see B and D). The dashed vertical and horizontal lines in (E) and (F) are the "true" values used to simulate the data.

(poor mixing), as in Fig. 3C. Algorithms that move groups of parameters simultaneously, such as Hamiltonian Monte Carlo (HMC; Neal 2011, Monnahan et al. 2017) or other block-wise samplers, are expected to show improved mixing as they can usually take "larger" steps in the multivariate parameter space. While our practical experience using OpenBUGS has often revealed that mixing does not improve, or can even be worse, when block samplers are automatically selected, the HMC methods employed by Stan have generally proven more successful (e.g., Monnahan et al. 2017). In fact, when we implemented the model based on Eq. (5) in Stan (see code in Appendix S1: Section S6), mixing and convergence notably improved compared to the JAGS and OpenBUGS simulations, even for large σ_ε (see Appendix S1: Fig. S4 and Table S1). While non-identifiability of β_0 and ε_j (or $\bar{\varepsilon}$) is not obvious for large σ_ε, based on visual inspection of the history plots (see Appendix S1: Fig. S4C, I, L), the HMC chains possess greater within chain autocorrelation, requiring a greater number of iterations to effectively sample the posterior parameter space, compared to scenarios with smaller σ_ε (see Appendix S1: Table S1).

Moreover, while the HMC methods employed by Stan can greatly improve mixing and convergence, the actual posterior correlation (e.g., Eq. [6]) among pairs of parameters is unaffected, regardless of the algorithm used. However, block samplers such as HMC move parameters in accordance with this correlation structure, increasing their efficiency. For example, evaluation of the Stan output reveals that β_0 and ε_j (or $\bar{\varepsilon}$) are still non-identifiable, especially for large σ_ε. Bivariate scatter plots of the posterior samples of β_0 vs. ε_j (or vs. $\bar{\varepsilon}$) obtained from Stan reveal that β_0 and ε_j (or $\bar{\varepsilon}$) are still highly correlated for large σ_ε such that a change in ε_j (or $\bar{\varepsilon}$) can entirely compensate for a change β_0 (see Appendix S1: Fig. S5), and the range of β_0 values explored by the HMC sampling algorithm is very wide (Appendix S1: Figs. S4C, S5 and Table S1).

Note that, by specifying the zero-centered hierarchical prior for the random effects, $\varepsilon_j \sim \text{Normal}(0, \sigma_\varepsilon^2)$, this implies that we might expect or want the overall mean or average, $\bar{\varepsilon}$, to be exactly zero. Returning to our simulation results, in all three σ_ε scenarios, the posterior mean for $\bar{\varepsilon}$ is not exactly zero (see Table 1 and Fig. 3G–I), regardless of the software or sampling algorithm used (Appendix S1: Table S1), alluding to potential non-identifiability of $\bar{\varepsilon}$ and individual ε_j. For large σ_ε ($\sigma_\varepsilon = 10\sigma$), the central posterior 95% credible interval (CI) for $\bar{\varepsilon}$ spans a wide range of values, from about -5 to 5 (the simulated y data span -17 to 23; Table 1 and Appendix S1: Table S1). A small σ_ε ($\sigma_\varepsilon = \sigma/10$) results in a 95% CI for $\bar{\varepsilon}$ that only spans -0.20 to 0.19, but $\bar{\varepsilon}$ is still never exactly zero (Table 1). Why is this? The zero-centered hierarchical prior for ε_j is simply that: a prior. While the prior means are $E(\varepsilon_j) = 0$ and $E(\bar{\varepsilon}) = 0$, the marginal or conditional posterior means are not necessarily restricted to zero. For example, based on a simple model only involving an overall intercept and random effects, again, no covariate effects, such that $\mu_i = \beta_0 + \varepsilon_j$, regardless of the priors chosen for β_0, σ, and σ_ε (i.e., conditional on these quantities), the analytical solution for the conditional posterior mean of $\bar{\varepsilon}$ is

$$E(\bar{\varepsilon}|\beta_0, \sigma, \sigma_\varepsilon, \mathbf{y}) = \frac{\beta_0 - \bar{\bar{y}}}{1 + \frac{\sigma^2}{\sigma_\varepsilon^2}\left(1 + \frac{N}{J}\right)}. \quad (7)$$

Eq. (7) does not evaluate to exactly zero; it is affected by the sampled values of β_0 and the magnitude of σ_ε relative to σ. In the unlikely event that an MCMC sample gives β_0 exactly equal to the overall mean of the data (i.e., $\bar{\bar{y}}$, the average of the group-level sample means, $\bar{y}_j$), then the posterior mean for $\bar{\varepsilon}$ evaluates to zero. Otherwise, the posterior mean approaches zero only for β_0 very close to $\bar{\bar{y}}$ or for a small random effects variance, σ_ε^2 (i.e., as $\beta_0 \to \bar{\bar{y}}$ and/or $\sigma^2/\sigma_\varepsilon^2 \to \infty$, $E(\bar{\varepsilon}|\beta_0,\sigma,\sigma_\varepsilon,\mathbf{y}) \to 0$). That is, when σ_ε^2 is relatively large (weak to no borrowing of strength), the mean of the random effects can be far from zero, and for a given value of $\sigma^2/\sigma_\varepsilon^2$, the deviation from zero is controlled by the value of β_0, further pointing to the non-identifiability of β_0, ε_j, and $\bar{\varepsilon}$.

Solutions to the Identifiability Problem

The above examples and associated identifiability problems are well known among applied statisticians, but their details may not be discussed in an applied statistics course typically taken by ecologists. In the context of Eqs. (1 and 5), more information is needed to estimate or identify the parameters (i.e., β_0 [or $\beta_{0,s}$] and ε_j [or ε_p]), which typically comes from a constraint on the parameters. Frequentist analyses build-in such constraints, which can also be used within a Bayesian model to solve this identifiability problem (we elaborate on this shortly). In this section, we outline multiple solutions, including use of more informative priors and imposing constraints in the form of reparameterizing the original model and/or implementing coding solutions.

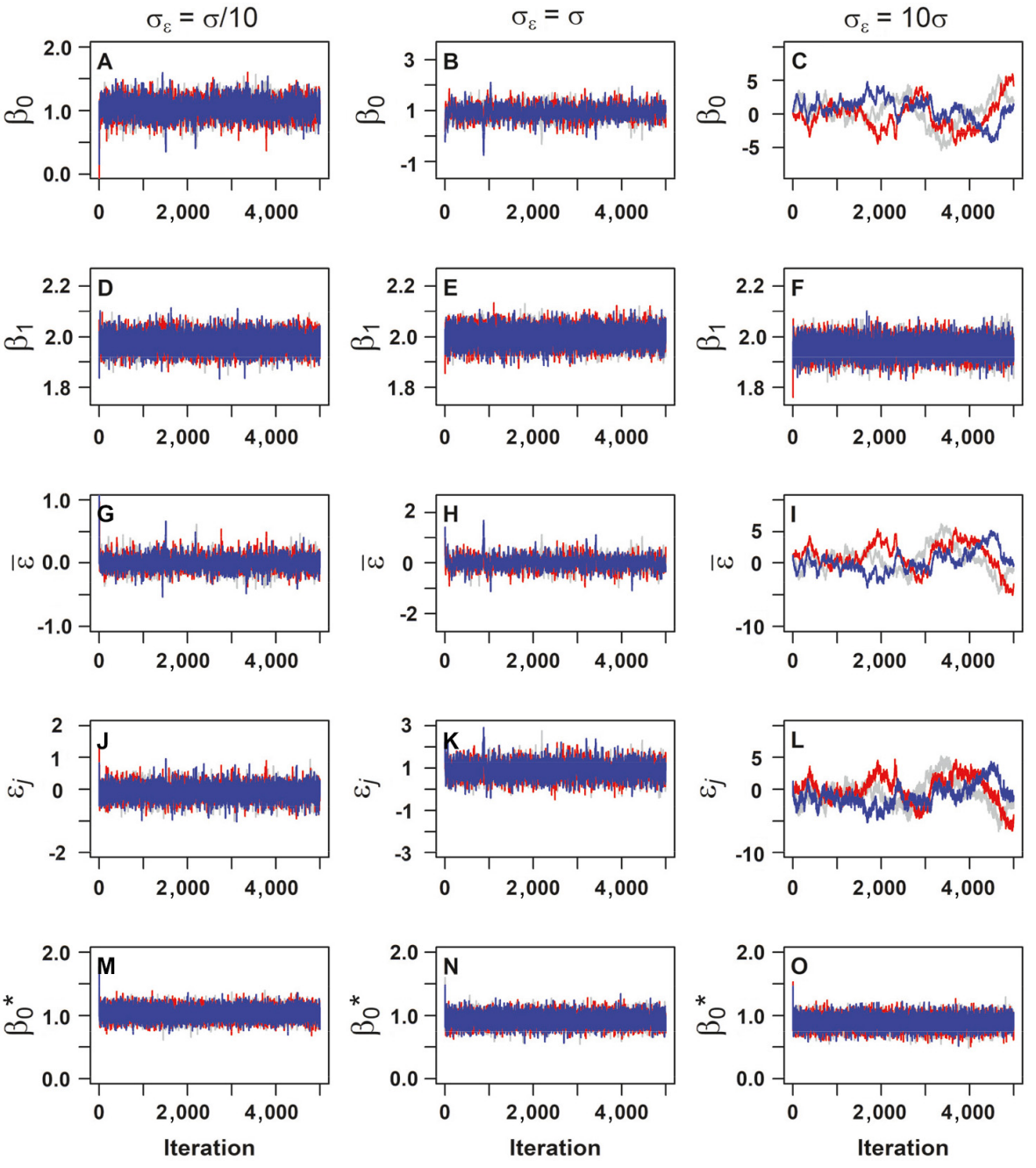

FIG. 3. History plots of the MCMC samples for parameters associated with the model in Eq. (5), fit to synthetic data. That is, for $y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i + \varepsilon_{j(i)}, \sigma^2)$ and $\varepsilon_j \sim \text{Normal}(0, \sigma_\varepsilon^2)$, data were generated from true values of $\beta_0 = 1$, $\beta_1 = 2$, $\sigma = 1$, and for $\sigma_\varepsilon = 0.1$ (left column), $\sigma_\varepsilon = 1$ (middle column), and $\sigma_\varepsilon = 10$ (right column). The random effects regression model, with a Gamma(0.1, 0.1) prior for σ^{-2} and σ_ε^{-2}, was in-turn fit to the synthetic data in JAGS to obtain posterior samples of parameters, including (A–C) β_0, (D–F) β_1, (G–I) the mean of the random effects, $\bar{\varepsilon}$, (J–L) an individual random effect (ε_j, for $j = 4$), and (M–O) the identifiable overall intercept ($\beta_0^* = \beta_0 - \bar{\varepsilon}$). The history plots for all quantities show excellent mixing and convergence for the scenario with a small random effects variance ($\sigma_\varepsilon = 0.1$; left column), but for large σ_ε, β_0, $\bar{\varepsilon}$, and ε_j exhibit very poor mixing and lack of convergence by iteration 5,000 (C, I, and L, respectively); β_1 and β_0^* exhibit excellent mixing and convergence behavior, regardless of the value of σ_ε. Differences in mixing and within-chain autocorrelation lead to differences in the number of posterior samples required for inference; based on Raftery and Lewis (1996), 13,200, 77,200 and 1,016,392 samples are required when $\sigma_\varepsilon = 0.1$, 1, and 10, respectively.

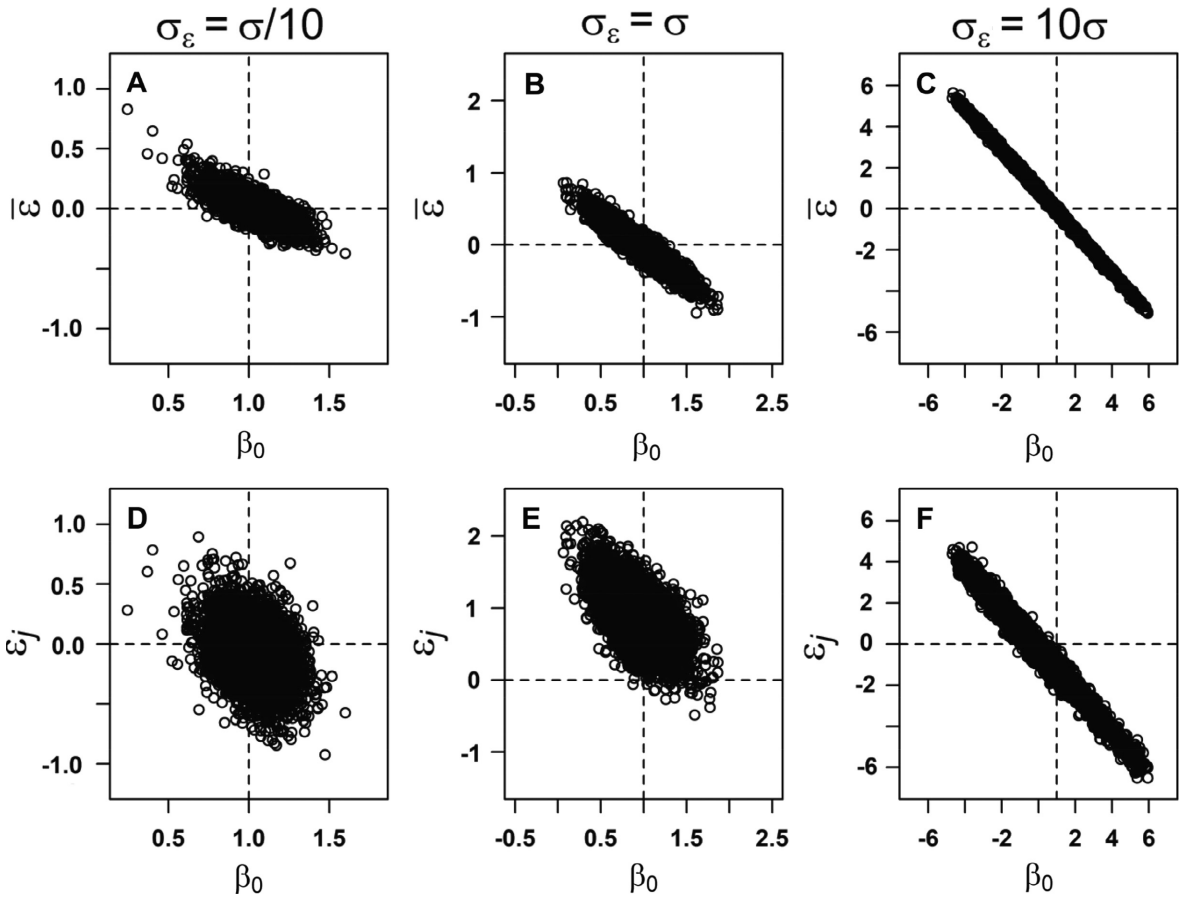

FIG. 4. Posterior results associated with the simulation described in Fig. 3. As in Fig. 3, the random effects regression model was fit to the synthetic data to obtain posterior samples of parameters. Bivariate scatterplots of the posterior MCMC samples are shown for the random effects mean ($\bar{\varepsilon}$) vs. the intercept (β_0) (top row) and for an individual random effect (ε_j, for $j = 4$) vs. β_0 (bottom row). Note that the correlation between $\bar{\varepsilon}$ (or ε_j) and β_0 becomes stronger as σ_ε (random effects variance component [SD]) increases relative to the observation variance (SD, $\sigma = 1$) such that for $\sigma_\varepsilon = 0.1$, (A) $r = -0.72$ and (D) $r = -0.32$; for $\sigma_\varepsilon = 1$, (B) $r = -0.93$ and (E) $r = -0.62$; and, for $\sigma_\varepsilon = 10$, (C) $r = -1.00$ and (F) $r = -0.99$.

Specification of more informative priors

A potentially easy solution to reducing the posterior correlation between, and thus enabling estimation of, the overall intercept (β_0) and the additive random effects (ε_j's) is to specify more informative priors for β_0 and/or σ_ε. Recent work suggests use of at least weakly informative priors for random effects variance terms (e.g., σ_ε) that result in greater shrinkage of the random effects (e.g., ε_j terms) toward their prior mean (e.g., toward 0 as per Eq. [4]) while reducing the likelihood of unrealistically large values of σ_ε (Gelman 2006, Lemoine 2019). For the simple random effects regression in Eq. (5), Lemoine (2019) recommends using a Cauchy(0,1) prior, folded at zero, for σ_ε (see also, Gelman 2006). While the use of a folded Cauchy(0,1) prior did shrink the marginal posterior for σ_ε toward smaller values (compare Table 1 vs. Appendix S1: Table S2, for scenario $\sigma_\varepsilon = 10\sigma$), it did not notably improve mixing or convergence of the MCMC chains (see Appendix S1: Fig. S6). Much more informative priors would be required for σ_ε and/or β_0 to improve MCMC behavior and to facilitate identification of β_0 and the ε_j terms (and $\bar{\varepsilon}$).

Thus, when prior information is available to construct such informative priors, we agree that such information should be leveraged (Hobbs and Hooten 2015), partly to address potential identifiability issues. For example, as demonstrated by the examples summarized in Table 1 and Fig. 3, pseudo-identifiability can be achieved if the prior(s) restrict σ_ε to small values relative to σ. It is common, however, for one to lack relevant and objective information for imposing informative priors, especially for parameters describing random effects; informative priors are more likely to be developed for population-level parameters describing biologically relevant quantities that can be directly measured (Gelman et al. 1996). Use of informative priors is not the focus of this paper, and we direct readers to other papers that focus on application of informative priors (e.g., Gelman et al. 1996, Rivot et al. 2001, Gelman 2006, Gelman et al.

TABLE 1. Posterior estimates (mean and 95% credible interval in parentheses) for the parameters in the random effects linear regression, Eq. (5), based on synthetic data described in Fig. 3, and using a relatively non-informative Gamma(0.1, 0.1) prior for σ_ε^{-2}.

	Random effects variance scenario		
Parameter, true value, and approach	$\sigma_\varepsilon = 0.1\sigma = 0.1$	$\sigma_\varepsilon = \sigma = 1$	$\sigma_\varepsilon = 10\sigma = 10$
β_0, true value = 1			
Orig.	1.042 (0.779, 1.309)	0.939 (0.439, 1.432)	0.921 (−3.991, 5.801)
HC	1.039 (0.770, 1.305)	0.939 (0.428, 1.457)	0.912 (−3.912, 5.851)
SZ	1.039 (0.858, 1.224)	0.938 (0.751, 1.129)	0.896 (0.699, 1.094)
PS	1.039 (0.854, 1.223)	0.936 (0.748, 1.124)	0.896 (0.700, 1.093)
β_1, true value = 2			
Orig.	1.978 (1.914, 2.042)	2.002 (1.935, 2.068)	1.955 (1.886, 2.025)
HC	1.978 (1.915, 2.043)	2.002 (1.935, 2.069)	1.955 (1.886, 2.025)
SZ	1.978 (1.914, 2.043)	2.002 (1.934, 2.069)	1.955 (1.886, 2.024)
PS	1.978 (1.913, 2.043)	2.002 (1.937, 2.068)	1.956 (1.887, 2.025)
σ, true value = 1			
Orig.	0.935 (0.812, 1.079)	0.959 (0.829, 1.116)	0.995 (0.860, 1.156)
HC	0.935 (0.811, 1.080)	0.958 (0.827, 1.116)	0.995 (0.861, 1.155)
SZ	0.934 (0.811, 1.079)	0.961 (0.828, 1.115)	0.996 (0.861, 1.158)
PS	0.935 (0.813, 1.082)	0.959 (0.829, 1.115)	0.995 (0.860, 1.154)
σ_ε, true value varies (see columns)			
Orig.	*0.289 (0.152, 0.529)*	0.720 (0.395, 1.261)	7.410 (4.677, 12.305)
HC	*0.289 (0.155, 0.529)*	0.721 (0.401, 1.254)	7.411 (4.683, 12.253)
SZ	*0.272 (0.148, 0.494)*	0.713 (0.386, 1.259)	*5.940 (3.745, 9.892)*
PS	*0.291 (0.155, 0.542)*	0.722 (0.398, 1.268)	7.474 (4.705, 12.410)
$\bar{\varepsilon}$, true value = 0			
Orig.	-0.001 (−0.199, 0.192)	-0.002 (−0.462, 0.466)	-0.024 (−4.902, 4.882)
HC	0.001 (−0.197, 0.198)	-0.001 (−0.487, 0.477)	-0.016 (−4.947, 4.816)
SZ	0	0	0
PS	0	0	0

Notes: The "true value" is the parameter value used to generate the synthetic data. Italicized CIs do not contain the true value, which only occurs for some instances of σ_ε. Approach is Orig., original without addressing identifiability issues; HC, hierarchical centering (Solution 1); SZ, sum-to-zero constraints for random effects (Solution 2); and, PS, post-sweeping of random effects (Solution 4, where β_0 is reported as the identifiable β_0^*). Results are not provided for reparameterization by sweeping (Solution 3) because one would typically choose one of the less technical and computationally faster solutions (i.e., Solutions 1, 2, or 4). See Appendix S1 for results obtained with OpenBUGS (Appendix S1: Fig. S2 and Table S1), Stan (Appendix S1: Fig. S4 and Table S1), and JAGS using a folded-Cauchy(0,1) prior for σ_ε (Appendix S1: Fig. S6 and Table S2).

2008, Choy et al. 2009, Delean et al. 2013, Morris et al. 2013, Morris et al. 2015, Thorson and Cope 2017, Lemoine 2019). Thus, in many cases, one may opt for approaches that impose constraints on the parameters via reparameterizing the original model and/or implementing coding solutions, which we outline in the following subsections. Though, these solutions can certainly be combined with use of informative or weakly informative priors.

Reparameterization or coding solutions

We draw upon the literature to summarize four potential solutions to the aforementioned identifiability problem. The first solution is to hierarchically center the random effects around the global intercept (Gelfand et al. 1995), effectively assigning a hierarchical prior following Eq. (3), thus abandoning Eq. (4). However, this solution is limited to models involving single or nested random effects. Thus, one may draw upon other reparameterization approaches that are more generally applicable. The second and third solutions are motivated by Gilks and Roberts (1996) and involve imposing sum-to-zero constraints on the random effects, or employing "reparameterization by sweeping" with the sum-to-zero constraint. The fourth solution involves "post-sweeping of random effects" as described in Gelman and Hill (2007). The last three solutions employ a zero-centered hierarchical prior akin to Eq. (4), but they result in the group of identifiable random effects having both a prior mean and posterior mean of zero (i.e., $\bar{\varepsilon} = 0$ exactly, for every MCMC iteration).

Solution 1: Hierarchical centering.—Consider models involving a scalar intercept and an additive random effect, such as Eq. (5). We can simply combine the intercept and random effects such that Eq. (5) can be rewritten as $\mu_i = \alpha_{j(i)} + \beta_1 x_i$. Then, we assign a hierarchically centered prior to α_j following Eq. (3) such that $\alpha_j \sim \text{Normal}(\beta_0, \sigma_\varepsilon^2)$ (Gelfand et al. 1995, Gilks and Roberts

1996). Thus, α_j is the identifiable group-specific intercept (i.e., $\alpha_j = \beta_0 + \varepsilon_j$), and β_0 and σ^2_ε are still interpreted as the overall intercept and the random effects variance, respectively. Many examples of hierarchically centered random effects can be found in the ecological literature, including, but certainly not limited to, species effects centered on higher taxonomic-level effects or global effects (Price et al. 2009, Zipkin et al. 2009, Coomes et al. 2011, Ogle et al. 2014, Tobler et al. 2015, Foss-Grant et al. 2016, Rich et al. 2017, Wooliver et al. 2017), plot effects centered on treatment-level or global effects (HilleRisLambers et al. 2009, Barker et al. 2014), and individual effects centered on global effects (Thomas et al. 2006, Kropp and Ogle 2015, Peltier et al. 2016).

Reparameterization by hierarchical centering does not alter the underlying statistical model, but it allows us to identify β_0 (compare Fig. 5B to 4A) and individual ε_j, which are computed as $\varepsilon_j = \alpha_j - \beta_0$ (results not shown). However, the 95% CI for β_0 is comparable to the original non-identifiable model (compare Orig. and HC results in Table 1), indicating that the precision of β_0 is generally not improved by this solution. Moreover, this approach only works when the random effects (e.g., species, plot, or individual effects) can be centered on the overall intercept (e.g., genus, order, treatment-level, or global effects). See *Extensions* for an overview of how to extend hierarchical centering to a situation involving multiple, nested random effects (e.g., plot within watershed random effect plus a watershed random effect). Conversely, consider the first example, Eq. (1), which involves a species-specific intercept (fixed factor) and additive plot random effects. Recall that it is unlikely that we can treat plots as being nested within species; species and plot are more likely to be crossed factors. Thus, we are forced to work with the original model specification where $\mu_i = \beta_{0,s(i)} + \beta_{1,s(i)}x_i + \varepsilon_{p(i)}$. In situations involving crossed effects, whether fixed or random, hierarchical centering is not appropriate, and we draw upon one of the other potential solutions.

Solution 2: Sum-to-zero constraint.—For the simple linear model with one group of random effects, as in Eqs. (1 or 5), this solution effectively treats $J - 1$ of the ε_j's as stochastic and assigns each a hierarchical prior according to Eq. (4); the remaining (one), say ε_J, is set equal to minus the sum of the other $J - 1$ ε_j such that

$$\begin{aligned} \varepsilon_j &\sim \text{Normal}(0,\sigma^2_\varepsilon) \text{ for } j = 1, 2, \cdots, J-1 \\ \varepsilon_J &= -\sum_{j=1}^{J=1} \varepsilon_j. \end{aligned} \tag{8}$$

Clearly, this ensures that the sum, and hence the average of the random effects ($\bar{\varepsilon}$), is always equal to zero. Because the average is fixed at zero and no longer trades-

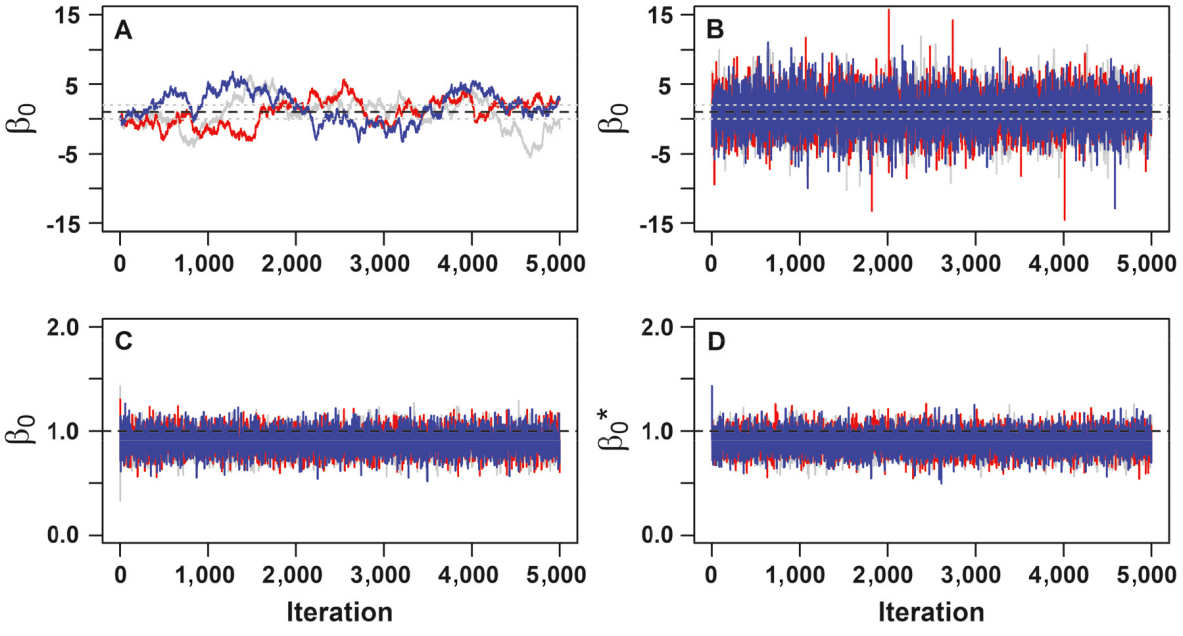

FIG. 5. History plots of the MCMC samples for the overall intercept (β_0 or β_0^*) associated with the model in Eq. (5), fit to synthetic data (see Figs. 3, 4). Results are only shown for the case involving a large random effects variance ($\sigma_\varepsilon = 10\sigma$) for (A) the original, non-identifiable parameterization (same as Fig. 3C); (B) the hierarchically centered version such that $\mu_i = \alpha_{j(i)} + \beta_1 \cdot x_i$ and $\alpha_j \sim \text{Normal}(\beta_0, \sigma^2_\varepsilon)$; (C) sum-to-zero constraints applied to the random effects, where $\mu_i = \beta_0 + \beta_1 x_i + \varepsilon_{j(i)}$, $\varepsilon_j \sim \text{Normal}(0, \sigma^2_\varepsilon)$ for $j = 1, 2, \ldots, J-1$, and $\varepsilon_J = -\sum_{j=1}^{J-1} \varepsilon_j$; and (D) post-sweeping of the random effects such that the identifiable intercept (plotted here) is computed as $\beta_0^* = \beta_0 + \bar{\varepsilon}$. The black horizontal dashed line is the true value of β_0 (or β_0^*) that was used to generate the synthetic data. The *y*-axes are scaled differently (wider) in panels A and B compared to C and D, and the horizontal gray dotted lines in A and B denote the *y*-axes range in C and D. Based on Raftery and Lewis (1996), (A) 830,500, (B) 4,300, (C) 3,800, and (D) 3,900 samples are required to obtain accurate 95% CIs for the intercept (β_0 or β_0^*).

off with the overall intercept (β_0), this leads to identifiable random effects and overall intercept. We applied this constraint to the simulated data example, and the posterior results are given in Table 1. This solution resulted in posterior estimates for all quantities that agree with the true values (the truth is contained in the 95% CIs), with the exception of σ_ε, which was slightly underestimated for true $\sigma_\varepsilon = 0.1$ and slightly overestimated for true $\sigma_\varepsilon = 10$ (though, a folded Cauchy(0,1) prior resulted in a more accurate estimate of σ_ε for true $\sigma_\varepsilon = 10$; see Appendix S1: Table S2). Importantly, the sum-to-zero constraint results in notable improvements in mixing and convergence of the MCMC chains (e.g., compare Fig. 5C to A), and the 95% CI for β_0 is notably narrower (more precise) than the original and hierarchical centering approaches (compare SZ to Orig. and HC results in Table 1). The sum-to-zero solution, Eq. (8), however, is not appropriate for small group sizes (e.g., $J < 5$ or 10), as discussed in *Solution 3: Reparameterization by sweeping*.

Solution 3: Reparameterization by sweeping.—As discussed in Gilks and Roberts (1996), the sum-to-zero constraint, Eq. (8), essentially "sweeps" the mean of the random effects ($\bar{\varepsilon}$) out of the random effects (ε_j) and into the overall mean or intercept (e.g., β_0). This simple sum-to-zero constraint works well for large group sizes (roughly, $J > 10$), but the independent normal assumption for the $J - 1$ ε_j terms is unreasonable for small J. Consider the extreme example where $J = 2$. If we employ sum-to-zero for ε_1 and ε_2, then $\varepsilon_1 = -\varepsilon_2$, exactly. That is, ε_1 and ε_2 are perfectly, negatively correlated. In general, the sum-to-zero constraint results in negative correlations among the ε_j terms, and the strength of this correlation increases with smaller J. Gilks and Roberts (1996) give an analytical solution for the correlation among pairs of such constrained random effects, which leads to modeling the *vector* of $J - 1$ ε_j terms as coming from a multivariate normal distribution with a covariance matrix ($\boldsymbol{\Sigma}$) that explicitly accounts for the induced correlation among the ε_j terms, giving Solution 3:

$$\begin{gathered}\boldsymbol{\varepsilon}_{-J} \sim \text{Normal}_{J-1}(0, \Sigma) \\ \textstyle\sum_{j,k} = -\frac{\sigma_\varepsilon^2}{J} \quad j \neq k \quad \text{and} \quad \sum_{j,j} = \sigma_\varepsilon^2 \\ \varepsilon_J = -\sum_{j=1}^{J-1} \varepsilon_j,\end{gathered} \tag{9}$$

where $\boldsymbol{\varepsilon}_{-J}$ is the vector of all $J - 1$ random effects, that is, excluding the "last" (J^{th}) random effect, and $\Sigma_{j,k}$ denotes element (j, k) of the covariance matrix; the last effect, ε_J, is obtained by the sum-to-zero constraint. As J gets very large (as $J \to \infty$), the covariance, $\Sigma_{j,k}$, among any pair of random effects, ε_j and ε_k, goes to zero (uncorrelated), and we can fall back on the simple sum-to-zero solution. Application of Eq. (9) should thus lead to unbiased estimates of σ_ε^2.

The sum-to-zero constraint and associated sweeping of the random effects mean (Solution 3, Eq. [9]) requires some additional coding steps upon implementation in software such as OpenBUGS, JAGS, or Stan (or via one's own custom MCMC routine). In particular, we must define the covariance matrix ($\boldsymbol{\Sigma}$) in addition to the sum-to-zero constraint, and evaluation of the multivariate normal prior in Eq. (9) requires inversion of the $(J - 1) \times (J - 1)$ covariance matrix, which becomes computationally burdensome for increasing J. For large J, however, we may simply use sum-to-zero, see Eq. (8), as an fast approximation, which improves with increasing J. However, if random effects are thought to be correlated, independent of correlations caused by the sum-to-zero constraint, as might occur for spatial or temporal random effects, a multivariate model, different from Eq. (9), would likely be required. Discussion of spatially or temporally correlated random effects, leading to non-exchangeability, is beyond the scope of this paper, and can be found elsewhere (e.g., Wikle 2003, Banerjee et al. 2004, Latimer et al. 2009, Finley 2011, Kang and Cressie 2011, Ver Hoef et al. 2018, Wikle et al. 2019).

Solution 4: Post-sweeping of random effects.—This solution retains the original parameterization involving the non-identifiable intercept and random effects. However, these non-identifiable quantities are only used to compute relevant identifiable quantities that we store, monitor, evaluate, summarize, and report. There is no need to monitor or store the non-identifiable quantities, and they should not be involved in our assessment of mixing and convergence. Following the example in Eq. (5), we compute the identifiable intercept (β_0^*) and random effects (ε_j^*) as

$$\begin{aligned}\varepsilon_j^* &= \varepsilon_j - \bar{\varepsilon} \text{ for} j = 1, 2, \cdots, J, \text{ where } \bar{\varepsilon} = \frac{1}{J}\sum_{j=1}^{J} \varepsilon_j. \\ \beta_0^* &= \beta_0 + \bar{\varepsilon}\end{aligned} \tag{10}$$

That is, we subtract ("sweep out") $\bar{\varepsilon}$ from the non-identifiable ε_j's to obtain the identifiable ε_j^*, and we add ("sweep in") $\bar{\varepsilon}$ to the non-identifiable intercept to obtain β_0^*. This results in adding and subtracting $\bar{\varepsilon}$ (a constant) to the model for μ_i (net change of zero) such that the mean, μ_i, is not affected. We specify the original, zero-centered hierarchical prior for the non-identifiable ε_j following Eq. (4), and we retain the original prior for the non-identifiable β_0, likely following Eq. (2). The average of the ε_j^* terms is always zero (i.e., $\bar{\varepsilon}^* = (1/J)\sum_{j=1}^{J} \varepsilon_j^* = 0$), and they thus have the typical intuitive interpretation as deviations from the global mean ($\beta_0^* + \beta_1 \cdot x_i$).

The identifiable terms are considered derived quantities and their solutions can be programmed directly into the model code (e.g., using OpenBUGS, JAGS, or Stan) or computed outside of the model code using the MCMC output (e.g., coda object; Plummer et al. 2006) that contains the non-identifiable β_0 and ε_j. Either way, we evaluate burn-in and convergence of the identifiable quantities. Based on the simulation experiment, relative to the non-identifiable model (Figs. 3, 5A), mixing and

convergence of the identifiable quantities (Fig. 5D) produced by Solution 4 are much improved and comparable to Solutions 1 (Fig. 5B) and 2 (Fig. 5C). Often, implementation of the intentionally non-identifiable model, as per Solution 4, can help to improve mixing and convergence of the desired identifiable quantities and lead to relatively precise estimates (see Table 1).

In general, introducing auxiliary quantities or nuisance variables, identifiable or not, can aid mixing (Gelfand et al. 1995, Gelman 2004). As another example of an intentionally non-identifiable model, see parameter expansion techniques (e.g., Gelman 2004, 2006, Gelman and Hill 2007), which may be employed to improve mixing and convergence of hierarchically centered effects, Eq. (3), associated with a small variance component (e.g., small σ_ε). In such situations, the MCMC chains for σ_ε can get stuck near zero ("zero variance trap"), and the corresponding random effects (e.g., ε_j or α_j terms) will likely exhibit strong within chain autocorrelation and poor mixing (e.g., "flat lining"). Parameter expansion introduces additional, intentionally non-identifiable quantities (redundant parameters) that facilitate greater movement (and mixing) of the MCMC chains (for σ_ε, ε_j, or α_j). Again, we ignore the non-identifiable quantities and focus our inference on the identifiable quantities (e.g., σ_ε and ε_j or α_j).

When to implement which solution?.—As illustrated by the simulation experiment, all three of the solutions (1, 2, and 4) highlighted in Fig. 5 and Table 1 resulted in improved mixing and convergence of the MCMC chains. Compared to the original, non-identifiable version, Solutions 2 and 4 also produced more precise (narrower 95% CIs) and more accurate (95% CIs contained the true value) estimates of the quantities that are susceptible to non-identifiability, especially when $\sigma_\varepsilon \gg \sigma$ (Table 1). The sum-to-zero constraint (Solution 2), however, yielded a posterior for σ_ε that is noticeably, but not statistically, different (here, lower and narrower 95% CI) from the posteriors produced by all other solutions (Table 1). This difference partly reflects the fact that the sum-to-zero constraint produces dependent random effects, thus producing a biased estimate of σ_ε (Gilks and Roberts 1996). Here, σ_ε is interpreted as the super-population (the population from which the sampled levels came from) standard deviation (Gelman 2005, Gelman and Hill 2007). One may also be interested in computing the finite-population (the specific levels sampled) standard deviation, s_ε, which will generally have a more precise estimate (Gelman and Hill 2007), and should be consistent among the different approaches. For example, in the simulation with a large random effects variance ($\sigma_\varepsilon = 10\sigma$), the estimates of s_ε were nearly identical among the four approaches summarized in Table 1, with a posterior mean and 95% CI of 6.86 (6.50, 7.06). It is straightforward to compute s_ε in the model code as the standard deviation of the ε_j's or ε_j^*'s (Gelman and Hill 2007), and one may want to report s_ε in addition to or in lieu of σ_ε (Gelman 2005).

So, which solution should one use? If we simply wish to improve mixing and convergence, we may opt to implement our models in software such as Stan given its ability to efficiently sample the multivariate parameter space. If we also wish to separately identify the overall intercept and groups of random effects and to interpret the random effects as deviations from the overall mean, we may want to employ one of the aforementioned solutions (1, 2, or 4), potentially in combinations with informative priors. If we employ a model with a single group of additive random effects or multiple groups of additive random effects that can be nested within each other, in general, the preferred solution is to employ hierarchical centering (Solution 1). This solution is easy to code and relatively fast compared to the other solutions. We summarize our recommendations in Fig. 6.

In more complex situations involving, for example, multiple groups of random effects that are not nested (see *Extensions*), we should consider one or more of the latter three solutions. For a particular model, we would recommend using the same solution for all groups of random effects, for consistency. And, for a particular model, we might try one of two (or both) appropriate solutions. If the group size, J_g, associated with each group $g = 1, 2, \ldots, G$ of random effects (e.g., plots, time periods, etc.) is *large* (e.g., $J_g \gg 10$ for all g), then one could employ the sum-to-zero constraint or post-sweeping for each group of random effects. Both of these solutions are easy to code and faster than reparameterization by sweeping. If all or some group sizes are comparatively small (e.g., $J_g \leq 5$ for one or more g), then the basic sum-to-zero constraint is inappropriate, leaving reparameterization by sweeping and post-sweeping as options. In practice, however, the former is more challenging to code, requires specification of the covariance matrix ($\boldsymbol{\Sigma}$) in Eq. (9), and leads to slower MCMC simulations. In summary, while we have used all four of these solutions, our experience has led us to prefer Solution 4 (post-sweeping of random effects), which, again, is easy to code, works for J (or J_g's) large or small, and MCMC simulation speed is not notably impacted. Though, the choice of which solution to use may be problem specific, and dictated by convergence and mixing diagnostics.

We provide example code illustrating application of hierarchical centering (Solution 1), sum-to-zero constraint (Solution 2), and post-sweeping of random effects (Solution 4) in Appendix S1: Section S3. We do not provide code for reparameterization by sweeping (Solution 3) given that we generally do not recommend this solution (see above discussion). Again, we note that these reparameterization or coding solutions can be combined with specification

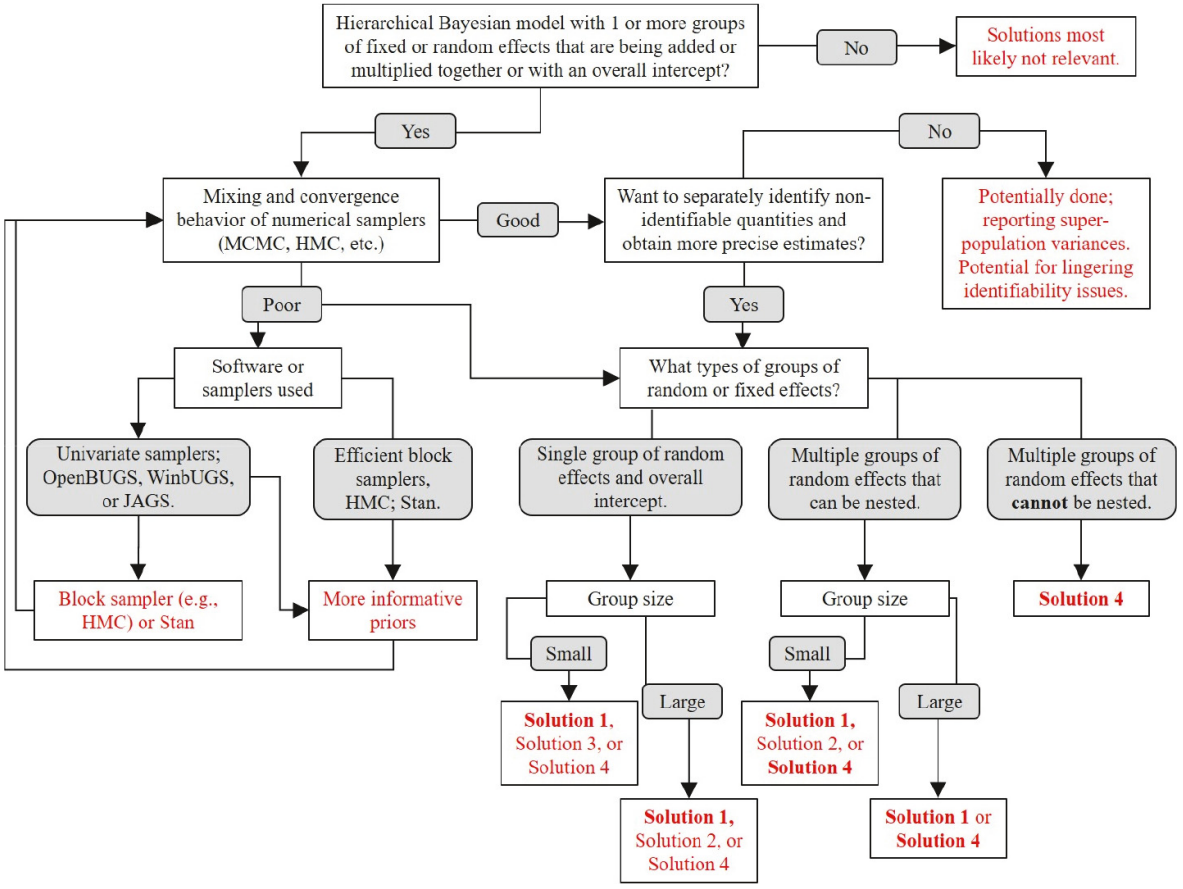

FIG. 6. Flowchart summarizing our recommendations for selecting solutions to address potential identifiability and/or mixing and convergence problems associated with a hierarchical Bayesian model involving one or more groups of fixed or random effects. Solution 1 is hierarchical centering, Solution 2 is sum-to-zero constraint, Solution 3 is reparameterization by sweeping, and Solution 4 is post-sweeping of random effects. For boxes with multiple solutions, solutions shown in boldface type are the preferred solutions. HMC, Hamiltonian Monte Carlo.

of informative priors (e.g., Lemoine 2019), when appropriate.

Extensions

Here we outline approaches to dealing with common modeling situations involving mixed effects, multiple groups of nested random effects, multiple groups of non-nested random effects, multiplicative random effects, and multiple groups of fixed effects.

Additive fixed and random effects.—Let us return to the mixed effects model in Eq. (1), which assumes an intercept that varies by some group level s (e.g., $\beta_{0,s}$ where s could refer to species) plus additive random effects for level p of another group (ε_p, where p could refer to plot). We have already discussed the types of priors that we would assign to the fixed effects, $\beta_{0,s}$, and the random effects, ε_p. In this example, since plot and species are likely crossed, we would employ Solutions 2 (sum-to-zero), 3 (reparameterization by sweeping), or 4 (post-sweeping). If we choose Solution 4, we would simply compute the identifiable random effects and fixed effect intercepts as $\varepsilon_p^* = \varepsilon_p - \bar{\varepsilon}$ (for all p) and $\beta_{0,s}^* = \beta_{0,s} + \bar{\varepsilon}$ (for all s), respectively.

Multiple groups of nested random effects.—Consider the following model that extends Eq. (5) to include two additive random effects, where one group (e.g., plot, $p = 1, 2, \ldots, P_w$) is nested in the other (e.g., watershed, $w = 1, 2, \ldots, W$). This notation indicates there are P_w plots in watershed w. (P_w can be different for each w; we are not restricted to balanced designs.) The mean model might look like

$$\mu_i = \beta_0 + \beta_1 x_i + \varepsilon_{p(i),w(i)} + \gamma_{w(i)}, \qquad (11)$$

where $p(i)$ and $w(i)$ are the plot and watershed associated with observation i, respectively. This formulation explicitly includes plot and watershed random effects, $\varepsilon_{p,w}$ and γ_w, which are added to the overall (global) intercept, β_0. Staying with the parametrization in Eq. (11), one would assign hierarchical priors to $\varepsilon_{p,w}$ and γ_w following Eq.

(4), with variances σ^2_ε and σ^2_y, respectively. This additive model, however, creates non-identifiability among β_0, $\varepsilon_{p,w}$, and γ_w; again, β_1 is identifiable because it is the coefficient on the (centered or standardized) covariate, x, which is assumed to vary among observations. Assuming plots are nested in watersheds, then the identifiability problem can be solved by hierarchical centering via a multi-level hierarchical model (Gelman and Hill 2007). That is, rewrite the mean model as $\mu_i = B_{p(i),w(i)} + \beta_1 x_i$, and specify a hierarchical prior for the plot-level intercept as $B_{p,w} \sim \text{Normal}(b_w, \sigma^2_\varepsilon)$, followed by a hierarchical prior for the watershed-level intercept, $b_w \sim \text{Normal}(\beta_0, \sigma^2_y)$. Note, σ^2_ε and σ^2_y still describe the variability among plots within a watershed and the variability among watersheds, respectively. The model specification is completed by assigning appropriate priors to β_0, β_1, σ^2_ε, σ^2_y, and any additional parameters (e.g., observation variances) introduced by the likelihood for the data.

If we are interested in making inferences about how plots or watersheds deviate from the overall response such that we wish to learn about $\varepsilon_{p,w}$ and γ_w, then we could retain the original formulation in Eq. (11) and employ Solution 2 (sum-to-zero constraints; assuming large P_w and W) or Solution 4 (post-sweeping; small or large P_w and W). Both approaches require modifications to account for plots being nested within watersheds; for example, plot random effects sum to zero *within* each watershed. Under Solution 2:

$$\begin{aligned}
&\varepsilon_{p,w} \sim \text{Normal}\,(0, \sigma^2_\varepsilon) \text{for } p = 1, 2, \cdots, P_w - 1 \text{ and} \\
&\varepsilon_{P_w,w} = -\sum_{p=1}^{P_w-1} \varepsilon_{p,w} \text{ for } w = 1, 2, \cdots, W \\
&\lambda_w \sim \text{Normal}\,(0, \sigma^2_\gamma) \text{ for } w = 1, 2, \cdots, W - 1 \text{ and} \\
&\lambda_w = -\sum_{w=1}^{W-1} \lambda_w
\end{aligned} \tag{12}$$

and the sum-to-zero constraint for γ_w follows Eq. (8). Alternatively, Solution 4 computes the identifiable quantities (e.g., Gilks and Roberts 1996):

$$\begin{aligned}
\varepsilon^*_{p,w} &= \varepsilon_{p,w} - \bar{\varepsilon}_w \text{ where } \bar{\varepsilon}_w = \tfrac{1}{P_w}\sum_{p=1}^{P_w} \varepsilon_{p,w} \\
\gamma^*_w &= \gamma_w + \bar{\varepsilon}_w - \bar{\gamma} - \bar{\bar{\varepsilon}} \text{where } \bar{\gamma} = \tfrac{1}{W}\sum_{w=1}^{W} \gamma_w \text{ and } \bar{\bar{\varepsilon}} = \tfrac{1}{W}\sum_{w=1}^{W} \bar{\varepsilon}_w \\
\beta^*_0 &= \beta_0 + \bar{\gamma} + \bar{\bar{\varepsilon}}
\end{aligned} \tag{13}$$

That is, the average of the non-identifiable plot random effects ($\bar{\varepsilon}_w$) is computed *within* each watershed, and this average is subtracted from the non-identifiable $\varepsilon_{p,w}$ terms and added to the γ_w terms, which also vary by w. The average ($\bar{\gamma} + \bar{\bar{\varepsilon}}$) of the "new" non-identifiable watershed random effects ($\gamma_w + \bar{\varepsilon}_w$) is computed across all watersheds, as done in Eq. (10), subtracted from the non-identifiable random effect, and added to β_0 to produce the identifiable global intercept (β^*_0). Again, adding and subtracting $\bar{\varepsilon}_w$, $\bar{\gamma}$, and $\bar{\bar{\varepsilon}}$ results in no net change to the mean, μ_i. Intuitively, all constants are swept from plot effects into the watershed effects and from watershed effects into the overall constant effect, creating a familiar interpretation of deviations from an overall constant, often mean or intercept, effect. Example JAGS code associated with this model, Eq. (11), is provided in Appendix S1: Section S4.

Multiple groups of non-nested random effects.—Suppose we have a model similar to Eq. (11), but the two groups of random effects are crossed rather than nested such as might occur for plots (plot $p = 1, 2, \ldots, P$) and dates (date $d = 1, 2, \ldots, D$). The model becomes

$$\mu_i = \beta_0 + \beta_1 x_i + \varepsilon_{p(i)} + \lambda_{d(i)} \tag{14}$$

where $p(i)$ and $d(i)$ indicate plot p and date d associated with observation i. It is straightforward to hierarchically center *one* group of random effects; either center the plot effects (ε_p) on the global intercept (β_0) and use one of the other solutions for the date random effects (λ_d), or vice versa. If we stick with the parameterization in Eq. (14), then the sum-to-zero constraint in Eq. (8) is applied separately to ε_p and λ_d. The post-sweeping approach computes the identifiable quantities

$$\begin{aligned}
\varepsilon^*_p &= \varepsilon_p - \bar{\varepsilon} \text{where } \bar{\varepsilon} = \frac{1}{P}\sum_{p=1}^{P} \varepsilon_p \\
\lambda^*_d &= \lambda_d - \bar{\lambda} \text{where } \bar{\lambda} = \frac{1}{D}\sum_{d=1}^{D} \lambda_d \\
\beta^*_0 &= \beta_0 + \bar{\varepsilon} + \bar{\lambda}
\end{aligned} \tag{15}$$

That is, ε^*_p and λ^*_d are computed as in Eq. (10), but the identifiable intercept is obtained by adding the means of both groups of non-identifiable random effects to β_0. As discussed, constant values are swept from all effects into the overall effect, again yielding the interpretation of deviations about an overall constant or mean effect. Example JAGS code associated with this model, Eq. (14), is provided in Appendix S1: Section S5.

Of course, there may be situations that involve both nested and non-nested random effects, such as random effects for plots (within watersheds), watersheds, and dates. For the sum-to-zero constraint, Eq. (12) would be used for the plot and watershed effects, and Eq. (8) for the date effects. For post-sweeping of random effects, Eq. (13) would be used to compute the identifiable plot and watershed effects, Eq. (15) for the identifiable date effects, and the identifiable intercept would be computed as $\beta^*_0 = \beta_0 + \bar{\gamma} + \bar{\bar{\varepsilon}} + \bar{\lambda}$ (i.e., the non-identifiable intercept is modified by the overall means for the non-identifiable watershed, $\bar{\gamma} + \bar{\bar{\varepsilon}}$, and date random effects, $\bar{\lambda}$).

Multiplicative random effects.—Multiplicative models are commonly used and may take on a form similar to

$$\mu_i = \alpha_0 f(\boldsymbol{\alpha}, \mathbf{x})\delta_{j(i)}, \quad (16)$$

where $f(\boldsymbol{\alpha},\mathbf{x})$ is some, likely nonlinear, function of potential covariates ($\mathbf{x}$) and associated parameters ($\boldsymbol{\alpha}$). Here, δ_j represents the multiplicative random effect associated with group level j. We typically expect $\delta_j > 0$ and that the δ_j terms vary around an "average" value of one, which represents no effect. Given the assumption $\delta_j > 0$, the normal priors listed in Eqs. (2–4) may not be appropriate as they would allow for $\delta_j < 0$, which would allow for an unusual and abrupt change in μ_i from positive to negative. Thus, we would likely chose a different probability distribution for the prior that aligns with the domain for δ_j, such as, but not limited to, a lognormal or gamma distribution for $\delta_j > 0$.

Note that α_0 and the δ_j terms are not identifiable; for example, we can multiply one (e.g., α_0) by a constant c and the other (e.g., δ_j) by $1/c$, which changes the parameters but not the mean, μ_i. It is "easiest" to solve this identifiability problem by first linearizing Eq. (16) such that

$$\begin{aligned}\log(\mu_i) &= \log(\alpha_0) + \log(f(\boldsymbol{\alpha}, \mathbf{x})) + \log(\delta_{j(i)}) \\ &= \beta_0 + \log(f(\boldsymbol{\alpha}, \mathbf{x})) + \varepsilon_{j(i)}\end{aligned} \quad (17)$$

where β_0 and ε_j are the global intercept and random effects, on the log scale, respectively; the priors in Eqs. (2–4) would be appropriate for ε_j, with Eq. (4) being most appropriate if viewed as a random effect. Thus, hierarchical centering (using Eq. [3] as a prior for the hierarchically centered random effect) or one of the other solutions can be directly applied to ε_j (and β_0 when relevant) in Eq. (17). If the model cannot be linearized in this way, then one could consider an approach that parallels hierarchical centering by rewriting Eq. (16) as $\mu_i = a_{j(i)} f(\boldsymbol{\alpha},\mathbf{x})$, and specifying a hierarchical prior for a_j, parameterized such that the prior mean (or mode or median) is $E(a_j) = \alpha_0$, with appropriate priors for $\boldsymbol{\alpha}$, α_0, and any other parameters introduced by the hierarchical prior for a_j. For example, for $a_j > 0$, we might model log (α_j) via a normal distribution with mean $\log(\alpha_0)$, or model α_j directly via a gamma distribution parameterized such that $E(a_j) = \alpha_0$ or $\text{mode}(a_j) = \alpha_0$. Finally, it is also possible that one could maintain the original Eq. (16) and employ constraints (similar to Solution 2) on the product of the random effects such that $\prod_{j=1}^{J} \delta_j = 1$. This involves specifying an appropriate prior distribution for $\delta_j > 0$ for $j = 1, 2, \ldots, J - 1$, and setting $\delta_J = (\prod_{j=1}^{J-1} \delta_j)^{-1}$. We have tried this constraint in a limited number of cases, but ultimately, we have been able to linearize the model and employ more "standard" solutions (our preference).

We note that nonlinear models may result in non-identifiability of parameters in the nonlinear mean function, $f(\boldsymbol{\alpha},\mathbf{x})$, independent of the issues discussed here related to identifiabilty of random effects (Beven and Freer 2001, Luo et al. 2009, Parslow et al. 2013, Hines et al. 2014). Other approaches to addressing non-identifiability in complex or nonlinear models, including specification of informative priors, are discussed elsewhere (e.g., Omlin and Reichert 1999, Eberly and Carlin 2000, Raue et al. 2013, Hines et al. 2014).

Multiple groups of fixed effects.—The examples that we have provided thus far focus on solutions to potential identifiability problems that arise in random or mixed effects models. These same identifiability issues can also arise when incorporating additive (or multiplicative) *fixed* effects. For example, Eq. (1) includes fixed effects for species such that the overall intercept ($\beta_{0,s}$) varies by species (s), plus a random effect for plot (ε_p). We already discussed the non-identifiability of $\beta_{0,s}$ and ε_p, and approaches to addressing this problem. What about models that involve nested or multiple groups of fixed effects? For example, $\varepsilon_{p,w}$ and γ_w in Eq. (11) could represent a random effect associated with individual p nested in species w ($\varepsilon_{p,w}$) and a fixed effect for species w (γ_w). Here, each γ_w (species fixed effect) would be assigned an independent prior following Eq. (2), but we still need to employ the sum-to-zero or post-sweeping solutions so that γ_w and the overall intercept are identifiable. Thus, nothing would change in terms of implementing solutions to the identifiability problem: the model (priors) only changes slightly to reflect our interpretation of γ_w as a fixed or random effect; for example, as a fixed effect, there is no longer a variance term associated with the γ_w effects. Similarly, the crossed effects (ε_p and λ_d) in Eq. (14) could represent fixed effects of, say, species p and drought treatment level d. Again, β_0, ε_p, and λ_d are non-identifiable. Given that species and drought level are viewed as crossed, fixed effects and are likely assigned independent (non-hierarchical) priors following Eq. (2), it is inappropriate to hierarchically center one of these effects around the global intercept (this would introduce a variance term that does not currently exist). However, we could still use either the sum-to-zero constraint or the post-sweeping of (fixed) effects solutions to overcome the identifiability problem, as discussed for the random effects examples.

However, an alternative solution for dealing with non-identifiable fixed effects that are assigned independent priors, as illustrated in Eq. (2), is to pick one of the levels to serve as the "reference" level or cell, and fix the reference level's effect at zero for additive effects or at one for multiplicative effects (Gelman and Hill 2007). The remaining effects are assigned priors according to Eq. (2), for additive effects. The reference level may be chosen to represent some nominal level (e.g., ambient conditions) or the level associated with the greatest number of observations. Thus, the intercept (e.g., β_0 in Eqs. [11 and 14]) or the prefactor (e.g., α_0 in Eq. [16]) are interpreted as the intercept or prefactor associated with the reference level, and the fixed effects associated with non-reference levels are interpreted as deviations from the reference level.

Conclusions

It is not clear if existing applications of hierarchical Bayesian models to ecological data address the aforementioned identifiability problems that arise by including additive (or multiplicative) random and/or fixed effects. The code (e.g., OpenBUGS, JAGS, or Stan) for implementing such models is rarely provided with publications, though, sharing of data and code will likely become more common (Nadrowski et al. 2013, Michener 2015, Bond-Lamberty et al. 2016, Dai et al. 2018, Powers and Hampton 2019), and thus it is difficult to evaluate if identifiability issues have been dealt with. If not, this can result in poor mixing and convergence and require a greater number of MCMC iterations, leading to longer run times and potentially unduly wide interval estimates for non-identifiable quantities (e.g., Table 1). Our personal experience, based on both informal consultation and evaluation of other's code, and formal reviews of manuscripts and code (when provided), suggests that many practitioners are not aware of these identifiability issues and are not implementing appropriate solutions.

Thus, the goal of this paper is to both bring awareness to the ecological community about these issues, especially since hierarchical Bayesian models are becoming increasingly popular (Ellison 2004, Clark and Gelfand 2006, Ogle and Barber 2008) (Fig. 1), and to provide explicit solutions. Regarding the latter, we provide examples of how to code (see Appendix S1: Sections S3–S5) the hierarchical centering, sum-to-zero, and post-sweeping solutions for the models defined in Eqs. (5, 11, and 14), representing models with a single group of random (or fixed) effects (Appendix S1: Section S3), nested effects (Appendix S1: Section S4), or crossed effects (Appendix S1: Section S5), respectively. The code can be implemented directly in JAGS or OpenBUGS, and it can be easily modified for application in Stan.

While the model and code examples are provided in the context of normally distributed data, e.g., $y_i \sim \text{Normal}(\mu_i, \sigma^2)$, we again note that the data model (likelihood) could be replaced by some other distribution that is relevant to a particular problem (e.g., Binomial, Poisson, log-normal, etc.). The mean model for μ_i in the normal-data example would thus represent the linear model for some transformation of $E(y_i)$, such as a typical link function in a generalized linear model (GLM). Thus, one would simply modify the data model and provide the link function that relates the linear model, μ_i, to $E(y_i)$; some minor modifications may be required if additional parameters are introduced or if some parameters are no longer relevant (e.g., σ^2). For example, in a logistic regression involving binomial response data, y_i, where we may specify a model like (1) data model (likelihood): $y_i \sim \text{Binomial}(p_i, N_i)$, where N_i is the known number of trials and p_i is the probability of "success," and (2) mean model with link function: $\text{logit}(p_i) = \mu_i$, with μ_i as defined for the "normal data" examples described herein. The modeling of the random and fixed effects and specification of priors in the model for μ_i are as previously outlined, regardless of the data model.

Acknowledgments

We thank the numerous graduate students and post-docs that have come through our labs for valuable discussions related to this paper, and for encouraging us to write this paper, in hopes that others will benefit from our suggestions and code. Simulated data and model code are provided in Appendix S1 (Sections S1-S6) and are archived as .R and .csv files (see Data Availability).

Literature Cited

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2004. Hierarchical modeling and analysis for spatial data. CRC Press, Boca Raton, Florida, USA.

Barker, R. J., D. M. Forsyth, and M. Wood. 2014. Modeling sighting heterogeneity and abundance in spatially replicated multiple-observer surveys. Journal of Wildlife Management 78:701–708.

Beven, K., and J. Freer. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology 249:11–29.

Bond-Lamberty, B., A. P. Smith, and V. Bailey. 2016. Running an open experiment: transparency and reproducibility in soil and ecosystem science. Environmental Research Letters 11:1–7.

Carlin, B. P., and T. A. Louis 2009. Bayesian methods for data analysis. Third edition. CRC Press, Boca Raton, Florida, USA.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, A. Riddell, J. Q. Guo, P. Li, and A. Riddell. 2017. Stan: A probabilistic programming language. Journal of Statistical Software 76:1–29.

Casella, G., and R. L. Berger 2002. Statistical inference. Second edition. Duxbury, Pacific Grove, California, USA.

Choy, S. L., R. O'Leary, and K. Mengersen. 2009. Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. Ecology 90:265–277.

Clark, J. S. 2007. Models for ecological data. Princeton University Press, Princeton, New Jersey, USA.

Clark, J. S., and A. E. Gelfand. 2006. A future for models and data in environmental science. Trends in Ecology & Evolution 21:375–380.

Coomes, D. A., E. R. Lines, and R. B. Allen. 2011. Moving on from Metabolic Scaling Theory: hierarchical models of tree growth and asymmetric competition for light. Journal of Ecology 99:748–756.

Dai, S. Q., H. Li, J. Xiong, J. Ma, H. Q. Guo, X. M. Xiao, and B. Zhao. 2018. Assessing the extent and impact of online data sharing in eddy covariance flux research. Journal of Geophysical Research—Biogeosciences 123:129–137.

Delean, S., B. W. Brook, and C. J. A. Bradshaw. 2013. Ecologically realistic estimates of maximum population growth using informed Bayesian priors. Methods in Ecology and Evolution 4:34–44.

Dorazio, R. M.. 2016. Bayesian data analysis in population ecology: motivations, methods, and benefits. Population Ecology 58:31–44.

Draper, D., J. S. Hodges, C. L. Mallows, and D. Pregibon. 1993. Exchangeability and data analysis. Journal of the Royal Statistical Society Series A 156:9–37.

Eberly, L. E., and B. P. Carlin. 2000. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. Statistics in Medicine 19:2279–2294.

Ellison, A. M. 2004. Bayesian inference in ecology. Ecology Letters 7:509–520.

Finley, A. O. 2011. Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. Methods in Ecology and Evolution 2:143–154.

Foss-Grant, A. P., E. F. Zipkin, J. T. Thorson, O. P. Jensen, and W. F. Fagan. 2016. Hierarchical analysis of taxonomic variation in intraspecific competition across fish species. Ecology 97:1724–1734.

Gamerman, D., and H. F. Lopes 2006. Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.

Gelfand, A. E., and K. Sahu. 1999. Identifiability, improper priors, and Gibbs sampling for generalized linear models. Journal of the American Statistical Association 94: 247–253.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin. 1995. Efficient parametrizations for normal linear mixed models. Biometrika 82:479–488.

Gelman, A. 2004. Parameterization and Bayesian modeling. Journal of the American Statistical Association 99:537–545.

Gelman, A. 2005. Analysis of variance—Why it is more important than ever. Annals of Statistics 33:1–31.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1:515–533.

Gelman, A., F. Bois, and J. M. Jiang. 1996. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. Journal of the American Statistical Association 91:1400–1412.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. Bayesian data analysis. Third edition. CRC Press, Boca Raton, Florida, USA.

Gelman, A., and J. Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York, New York, USA.

Gelman, A., A. Jakulin, M. G. Pittau, and Y. S. Su. 2008. A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics 2: 1360–1383.

Gilks, W. R., and G. O. Roberts. 1996. Strategies for improving MCMC. *In*W. R. Gilks, S. Richardson, and D. Spiegelhalter, editors. Markov Chain Monte Carlo in practice. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Gimenez, O., B. J. T. Morgan, and S. P. Brooks. 2009. Weak identifiability in models for mark-recapture-recovery data. Pages 1055–1067*in*D. L. Tompson, E. G. Cooch, and M. J. Conroy, editors. Modeling demographic processes in marked populations. Springer.

HilleRisLambers, J., W. S. Harpole, S. Schnitzer, D. Tilman, and P. B. Reich. 2009. CO_2, nitrogen, and diversity differentially affect seed production of prairie plants. Ecology 90:1810–1820.

Hines, K. E., T. R. Middendorf, and R. W. Aldrich. 2014. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. Journal of General Physiology 143:401–416.

Hobbs, N. T., and M. B. Hooten. 2015. Bayesian models: a statistical primer for ecologists. Princeton University Press, Princeton, New Jersey, USA.

Holand, A. M., and I. Steinsland. 2016. Is my study system good enough? A case study for identifying maternal effects. Ecology and Evolution 6:3486–3495.

Kang, E. L., and N. Cressie. 2011. Bayesian inference for the spatial random effects model. Journal of the American Statistical Association 106:972–983.

Kery, M., and J. A. Royle. 2008. Hierarchical Bayes estimation of species richness and occupancy in spatially replicated surveys. Journal of Applied Ecology 45:589–598.

Kropp, H., and K. Ogle. 2015. Seasonal stomatal behavior of a common desert shrub and the influence of plant neighbors. Oecologia 177:345–355.

Kruschke, J. 2014. Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan. Second edition. Academic Press, London, UK.

Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li 2004. Applied linear statistical models. Fifth edition. McGraw-Hill, New York, New York, USA.

Latimer, A. M., S. Banerjee, H. Sang, E. S. Mosher, and J. A. Silander. 2009. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. Ecology Letters 12:144–154.

Lemoine, N. P. 2019. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. Oikos 128:912–928.

Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best. 2009. The BUGS project: Evolution, critique and future directions (with discussion). Statistics in Medicine 28:3049–3082.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing 10:325–337.

Luo, Y. Q., E. S. Weng, X. W. Wu, C. Gao, X. H. Zhou, and L. Zhang. 2009. Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. Ecological Applications 19:571–574.

McCarthy, M. A. 2007. Bayesian methods for ecology. Cambridge University Press, Cambridge, UK.

McCulloch, C. E., and S. R. Searle. 2001. Generalized, linear, and mixed models. John Wiley & Sons Inc, New York, New York, USA.

McElreath, R. 2016. Statistical rethinking: a Bayesian course with examples in R and Stan. Chapman-Hall/CRC Press, Boca Raton, Florida, USA.

Michener, W. K. 2015. Ecological data sharing. Ecological Informatics 29:33–44.

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

Morris, W. K., P. A. Vesk, and M. A. McCarthy. 2013. Profiting from pilot studies: Analysing mortality using Bayesian models with informative priors. Basic and Applied Ecology 14: 81–89.

Morris, W. K., P. A. Vesk, M. A. McCarthy, S. Bunyavejchewin, and P. J. Baker. 2015. The neglected tool in the Bayesian ecologist's shed: a case study testing informative priors' effect on model accuracy. Ecology and Evolution 5:102–108.

Nadrowski, K., et al. 2013. Harmonizing, annotating and sharing data in biodiversityecosystem functioning research. Methods in Ecology and Evolution 4:201–205.

Neal, R. 2011. MCMC using Hamiltonian dynamics. Pages 116–162 *in* S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors. Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Ogle, K., and J. J. Barber. 2008. Bayesian data-model integration in plant physiological and ecosystem ecology. Progress in Botany 69:281–311.

Ogle, K., J. Barber, and K. Sartor. 2013. Feedback and modularization in a Bayesian meta-analysis of tree traits affecting forest dynamics. Bayesian Analysis 8:133–168.

Ogle, K., S. Pathikonda, K. Sartor, J. W. Lichstein, J. L. D. Osnas, and S. W. Pacala. 2014. A model-based meta-analysis for estimating species-specific wood density and identifying potential sources of variation. Journal of Ecology 102:194–208.

Ogle, K., D. Peltier, M. Fell, J. Guo, H. Kropp, and J. Barber. 2019. Should we be concerned about multiple comparisons in hierarchical Bayesian models? Methods in Ecology and Evolution 10:553–564.

Omlin, M., and P. Reichert. 1999. A comparison of techniques for the estimation of model prediction uncertainty. Ecological Modelling 115:45–59.

O'Neill, B. 2009. Exchangeability, correlation, and Bayes' effect. International Statistical Review 77:241–250.

Ovaskainen, O., and J. Soininen. 2011. Making more out of sparse data: hierarchical modeling of species communities. Ecology 92:289–295.

Parslow, J., N. Cressie, E. P. Campbell, E. Jones, and L. Murray. 2013. Bayesian learning and predictability in a stochastic nonlinear dynamical model. Ecological Applications 23:679–698.

Peltier, D. M. P., M. Fell, and K. Ogle. 2016. Legacy effects of drought in the southwestern United States: A multi-species synthesis. Ecological Monographs 86:312–326.

Plummer, M.2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Plummer, M.2012. JAGS Version 3.3.0 user Manual.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: convergence diagnosis and output analysis for MCMCM. R News 6:7–11.

Powers, S. M., and S. E. Hampton. 2019. Open science, reproducibility, and transparency in ecology. Ecological Applications 29:e01822.

Price, C. A., K. Ogle, E. P. White, and J. S. Weitz. 2009. Evaluating scaling models in biology using hierarchical Bayesian approaches. Ecology Letters 12:641–651.

Qian, S. S., T. F. Cuffney, I. Alameddine, G. McMahon, and K. H. Reckhow. 2010. On the application of multilevel modeling in environmental and ecological studies. Ecology 91:355–361.

Raftery, A. E., and S. M. Lewis. 1996. Implementing MCMC. *In*W. R. Gilks, S. Richardson, and D. Spiegelhalter, editors. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Ramsey, F. L., and D. W. Schafer. 2013. The Statistical Sleuth: a course in methods of data analysis. Third edition. Brooks/Cole, Boston, Massachusetts, USA.

Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Systematic Biology 51:754–760.

Raue, A., C. Kreutz, F. J. Theis, and J. Timmer. 2013. Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. Philosophical Transactions of the Royal Society A—Mathematical Physical and Engineering Sciences 371:20110544.

Reid, N. 1995. The roles of conditioning in inference. Statistical Science 10:138–199.

Rich, L. N., et al. 2017. Assessing global patterns in mammalian carnivore occupancy and richness by integrating local camera trap surveys. Global Ecology and Biogeography 26:918–929.

Rivot, E., E. Prevost, and E. Parent. 2001. How robust are Bayesian posterior inferences based on a Ricker model with regards to measurement errors and prior assumptions about parameters? Canadian Journal of Fisheries and Aquatic Sciences 58:2284–2297.

Sauer, J. R., and W. A. Link. 2002. Hierarchical modeling of population stability and species group attributes from survey data. Ecology 83:1743–1751.

Stan Development Team. 2018. Stan Modeling Language User's Guide and Reference Manual, Version 2.18. https://mc-stan.org/docs/2_18/stan-users-guide/index.html

Swartz, T., Y. Haitovsky, A. Vexler, and T. Yang. 2004. Bayesian identifiability and misclassification in multinomial data. Canadian Journal of Statistics 32:285–302.

Thomas, D. L., D. Johnson, and B. Griffith. 2006. A Bayesian random effects discrete-choice model for resource selection: Population-level selection inference. Journal of Wildlife Management 70:404–412.

Thorson, J. T., and J. M. Cope. 2017. Uniform, uninformed or misinformed?: The lingering challenge of minimally informative priors in data-limited Bayesian stock assessments. Fisheries Research 194:164–172.

Tobler, M. W., A. Z. Hartley, S. E. Carrillo-Percastegui, and G. V. N. Powell. 2015. Spatiotemporal hierarchical modelling of species richness and occupancy using camera trap data. Journal of Applied Ecology 52:413–421.

Touchon, J. C., and M. W. McCoy. 2016. The mismatch between current statistical practice and doctoral training in ecology. Ecosphere 7:e01394.

Ver Hoef, J. M., E. E. Peterson, M. B. Hooten, E. M. Hanks, and M. J. Fortin. 2018. Spatial autoregressive models for statistical inference from ecological data. Ecological Monographs 88:36–59.

Wakefield, J. 2013. Bayesian and frequentist regression methods. Springer-Verlag, New York, New York, USA.

Wikle, C. K. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84:1382–1394.

Wikle, C. K., A. Zammit-Mangion, and N. Cressie. 2019. Spatio-temporal statistics with R. CRC Press, Boca Raton, Florida, USA.

Wooliver, R. C., Z. H. Marion, C. R. Peterson, B. M. Potts, J. K. Senior, J. K. Bailey, and J. A. Schweitzer. 2017. Phylogeny is a powerful tool for predicting plant biomass responses to nitrogen enrichment. Ecology 98:2120–2132.

Zipkin, E. F., A. Dewan, and J. A. Royle. 2009. Impacts of forest fragmentation on species richness: a hierarchical approach to community modelling. Journal of Applied Ecology 46:815–822.

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/eap.2159/full

Data Availability

Data and code are available on GitHub via Zenodo: http://doi.org/10.5281/zenodo.3743847