Perils and pitfalls of mixed-effects regression models in biology

Matthew J. Silk[1,2], Xavier A. Harrison[1] and David J. Hodgson[1]

[1] Centre for Ecology and Conservation, University of Exeter, Penryn, Cornwall, UK
[2] Environment and Sustainability Institute, University of Exeter, Penryn, Cornwall, UK

ABSTRACT

Biological systems, at all scales of organisation from nucleic acids to ecosystems, are inherently complex and variable. Biologists therefore use statistical analyses to detect signal among this systemic noise. Statistical models infer trends, find functional relationships and detect differences that exist among groups or are caused by experimental manipulations. They also use statistical relationships to help predict uncertain futures. All branches of the biological sciences now embrace the possibilities of mixed-effects modelling and its flexible toolkit for partitioning noise and signal. The mixed-effects model is not, however, a panacea for poor experimental design, and should be used with caution when inferring or deducing the importance of both fixed and random effects. Here we describe a selection of the perils and pitfalls that are widespread in the biological literature, but can be avoided by careful reflection, modelling and model-checking. We focus on situations where incautious modelling risks exposure to these pitfalls and the drawing of incorrect conclusions. Our stance is that statements of significance, information content or credibility all have their place in biological research, as long as these statements are cautious and well-informed by checks on the validity of assumptions. Our intention is to reveal potential perils and pitfalls in mixed model estimation so that researchers can use these powerful approaches with greater awareness and confidence. Our examples are ecological, but translate easily to all branches of biology.

INTRODUCTION

Linear mixed-effects models (LMMs) and generalised linear mixed effects models (GLMMs) have, in recent decades, gained prevalence as statistical tools in biological research. They have replaced blocked, split-plot and hierarchical analysis of variance as the tool of choice for experimental biologists, and have helped countless surveyors of natural systems to measure hierarchical partitions of variation, and to deal with non-independence among their subjects. (G)LMMs are flexible, powerful, well-served by popular statistical software, and relatively easy to build and interpret. However, inevitably, they are also easy to get wrong. As peer reviewers, too often we see (G)LMMs being used naively and without validation.

(G)LMMs differ from simple linear models (LMs) and generalised linear models (GLMs) by incorporating random effects among the explanatory variables alongside fixed effects (*Gelman & Hill, 2006*; *Bolker et al., 2009*; *Zuur, Hilbe & Ieno, 2013*; *Harrison et al., 2018*). Fixed effects represent variables with intercepts, means or slopes to be estimated. Random effects describe experimental or survey units as members of groups, with these groups drawn from a larger population of other, unmeasured groups. For example, we might measure features of 10 populations of seabirds, but many hundreds of other populations of seabirds will not be measured in our study. Fitting 'population ID' acknowledges that each population represents a different group, that observations within a group are likely non-independent (i.e. seabirds in the same population will likely be more similar to each other than to seabirds in other populations), and that we have only measured a subsample of possible groups. The random effect components of mixed effects models infer the variance associated with group membership (*Gelman & Hill, 2006*; *Schielzeth & Nakagawa, 2013*). Consequently, the use of random effects is effective when measurements of a response variable within 'groups' or random effect levels tend to be more similar to each other than to those in the wider population. By inferring variance among groups rather than group-specific parameters, random effects absorb fewer degrees of freedom than fixed effects and this is exploited in many analyses to absorb real but 'unwanted' variation or covariance in the data (*Bolker et al., 2009*; *Harrison et al., 2018*). In other analyses, such as those used in quantitative genetics, the relative magnitudes of the variance components are of primary interest, because they provide estimates for the variation in traits explained at different levels of a model's hierarchy.

The inclusion of random effects in biological regression models is often seen as a panacea for dealing with difficult or structured datasets. One will often read statements in papers that a grouping variable was 'fitted as a random effect to control for non-independence among groups.' However, if not used appropriately, (G)LMMs can fail to achieve their purpose, or can yield false positive tests of importance of fixed effects (*Schielzeth & Forstmeier, 2008*; *Forstmeier, Wagenmakers & Parker, 2017*; *Arnqvist, 2020*). The correct specification of random effects is essential to successful analysis, but guidance is often couched in specialist statistical language, code or algebra and can be hard to find. Here we use non-specialist language to highlight the risks of coupling fixed and random effects naively in models. We explain seven common perils and pitfalls (Table 1), use simple data simulations to demonstrate the risk of inappropriate mixed-effect modelling, illustrate the problems caused by these pitfalls and show the reader to diagnose them in their own data and models.

Mixed model basics

The fundamental assumption shared by simple linear models and mixed effects models is that residual error for each measurement of the response variable is an independent draw from a Normal distribution with mean zero and fixed variance. (G)LMMs are mixed because they include both fixed and random effects, and linear because they describe the relationship between the observed data and the explanatory variables as the additive

Table 1 The perils of mixed modelling highlighted in this paper together with their potential consequences and solutions to avoid them.

Peril	Example	Consequences	Potential solutions
#1. Anticonservative significance tests at low sample size	Comparing crop yields in split-plot experiments with few replicates	p-value of Wald-like Chi-square test of significance is too low, causing high rates of Type I error	If better replication is not possible, use corrections for small sample size and accept that answers are approximations Move to statements of credibility based on Bayesian analyses
#2. Pseudoreplicated with group-level predictors	Infer the effect of maternal traits on the performance of several offspring per mother	Increased Type I error if pseudoreplication is not recognised Increased Type I error if true replication is small even if pseudoreplication is recognised	Have a firm grasp of the design level at which true replication occurs, and of the correct mixed-model specification If better replication of experimental units is not possible, use corrections for small sample size Move to statements of credibility based on Bayesian analyses with correct specification of design hierarchy
#3. Too few levels of a random effect	Fitting sex as a random effect	Model degeneracy Biased estimation of random effect variance Inaccurate estimation of random effect variance High error potential for questions related to random effects (less substantial for questions related to fixed effects)	Fit the variable as a fixed effect rather than random effect Use a Bayesian model with strong priors for the size of the random effect variance component. Such an approach requires caution and an ability to justify the inclusion of prior knowledge
#4. Random intercepts when groups vary in their response to a treatment	Testing a relationship between body size and competitive advantage in multiple populations of the same species	Increased Type I error rate when there is variation in slopes between different random effect levels. This is regardless of any correlation between the fixed and random effects	Use of random slopes model instead of random intercepts only
#5a. Confounding by cluster	Multiple observation of foraging behaviour are made at a succession of sites to test the hypothesis that foraging rates are associated with disturbance levels. However, sites have different mean levels of disturbance. Both disturbance (fixed effect) and site (random effect) are used as explanatory variables	Biased estimates of fixed and random effect parameters	Use within-group mean centring of variables alongside a group-level covariate
#5b. Informative cluster sizes	A model with offspring weight as a response and maternal pathogen load as an explanatory variable and maternal ID as a random effect, if high pathogen loads also cause reduced litter sizes	Possibility of biased estimates for fixed effect parameters if they are correlated with the random effect especially if the model only includes random intercepts	Fit cluster size as a covariate where appropriate (see main text) Joint modelling of the response variable of interest and cluster size in a multivariate model
#6. Group means are not normally distributed	An unmeasured variable causes differences between sub-populations Skewed differences in a trait between sub-populations that are unexplained by fixed effects	Fixed effect estimates robust unless (a) mis-specification of random effects is extreme or (b) fixed effects are correlated with random effects Random effect estimates can become less accurate and systematically biased	Use non-Gaussian random effect distributions (challenging, but made available by use of Bayesian models) Fit the variable as a fixed effect instead. Take extra care to check for other violations of mixed model assumptions
#7. Use of categorical random effects for autocorrelated data	Region is used as a random effect to control for spatial autocorrelation when modelling abundance of a species in response to a range of habitat variables	Poorly fitting model (with inaccurate predictions) and potential of statistical errors unless scale of random effect is correct	Use correlograms to check for covariance in the residuals of the model to ensure that categorical random effect is effective

$$\mathbf{y} = \beta_0 + \sum \beta_i \mathbf{x}_i + \gamma + \varepsilon$$

| Response variable | Global intercept | Fixed effect parameters | Fixed effect variables | Random effect variance | Residual variance |

Figure 1 **A mathematical and verbal representation of a simple mixed effects model.** y is the response variable, β_0 is the global intercept (the expectation of y when all fixed effects are zero, and for members of an average group in the random effect), x_i is the measured value of the ith fixed explanatory variable, β_i is the additive expected change caused by the value of each of the fixed explanatory variables, γ is a draw from the distribution of category means for a normally distributed random effect (with mean of zero and variance equal to the random effect variance), and ε is a draw from the normal distribution of residuals (with mean of zero and variance equal to the residual variance).
Full-size 🖼 DOI: 10.7717/peerj.9522/fig-1

contribution of intercepts and linear slopes (of the fixed effects), and deviations from these expectations caused by random effects and residual noise. The adjective Generalised in GLMMs refers to the use of canonical link functions that help deal with non-normal response variables. For both LMMs and GLMMs random effect group means are typically assumed to be Normally distributed on the link scale, although other distributions can be considered (described in more detail below). In general, the choice of whether a variable should be included as a random effect or a fixed effect is best decided by the researcher working on a problem based on a combination of question-driven and practical considerations (*Schielzeth & Nakagawa, 2013*; *Harrison et al., 2018*). We expand on other practical considerations in subsequent sections.

As a verbal model, any LMM can be described as a response variable (y) being an additive function of a global intercept, change caused by fixed effects, deviation caused by membership of a random effect category, and any residual variation not explained by the fixed or random effects. This is most elegantly described using statistical algebra (Fig. 1; this will be the only mathematical equation we will present). A simple example of LMM inference is provided in Fig. 2. While it is possible to fit mixed effect models in a variety of software (e.g. SAS (SAS Institute Inc.); SPSS (IBM Corp.); STATA (StataCorp. 2017)), we focus on the fitting of LMMs and GLMMs in the R statistical environment (*R Core Team, 2019*) using the packages lme4 (*Bates et al., 2015*) and nlme (*Pinheiro et al., 2014*). We, along with many of our biologist colleagues, enjoy the open source philosophy and community spirit associated with R. It is also free and extremely powerful.

Perils and pitfalls

In this guide, we identify seven major risks of misuse of (G)LMMs. We explain each of them in simple terms, discuss when they are most likely to cause problems and provide solutions in the form of better-formulated models. Some of the problems stem from an assumption that simple (G)LMMs are a panacea for non-independent data and can be used universally to model difficult datasets. However, complicated hierarchical designs must be treated with appropriate caution, as (G)LMMs are not infallible if specified
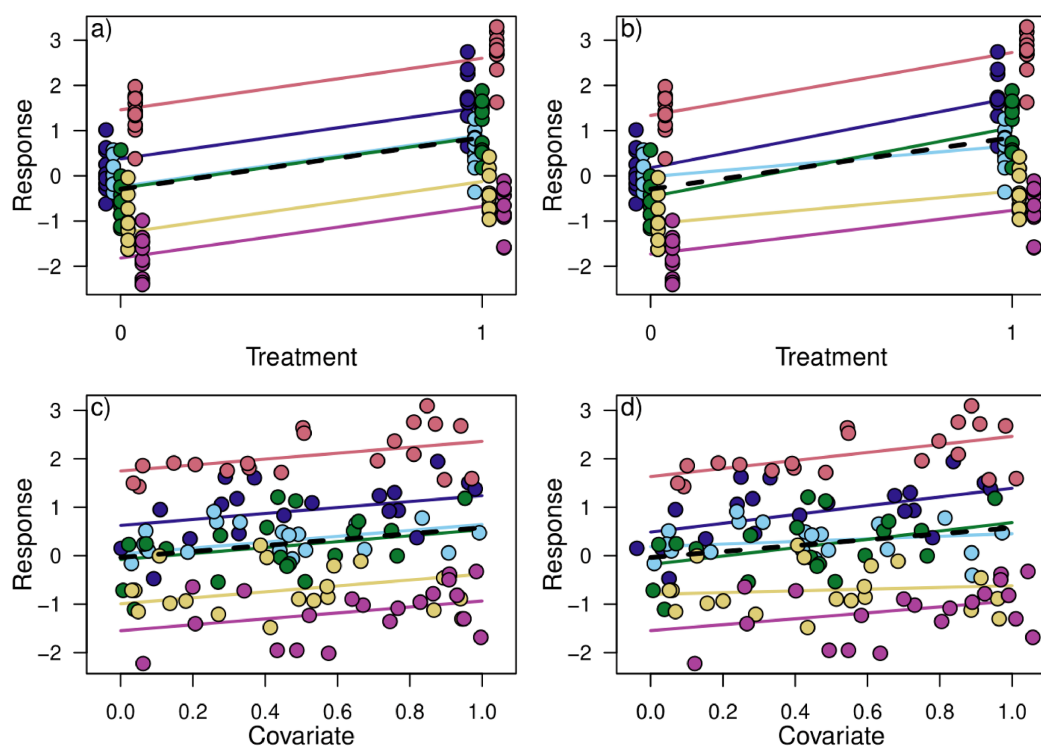
Silk et al. (2020), *PeerJ*, DOI 10.7717/peerj.9522

4/20

Figure 2 Simple examples of linear mixed models. In (A and B), offspring from six families (different colours) are allocated to one of two treatments (a categorical fixed effect). In (A), a *random intercepts model*, the treatment (fixed) effect is the slope of the line connecting each family cluster. The different elevation of the lines for each family indicates their difference from the overall intercept as drawn from a Normal distribution that represents a wider population of families. The random effect absorbs variation in intercepts among the families, helping to reveal the independent, global influence of treatment (black dashed line). In (B), a *random slopes model* reveals a global slope having absorbed variation among families in both intercept and effect of treatment. In (C and D), the response variable is instead regressed against a covariate (continuous fixed effect) measured, for each offspring, between zero and one. The random intercepts model (C) reveals the global slope (black dashed line) of the relationship between response and explanatory variables, having absorbed variation in the intercept among families. The random slopes model in (D) reveals a global slope having absorbed variation among families in both intercept and slope. The random intercepts model infers parallel changes caused by the fixed effect among groups, while random-slopes allows the model to infer variation in slopes as well as intercepts.

Full-size 🖼 DOI: 10.7717/peerj.9522/fig-2

incorrectly (*Arnqvist, 2020*). Further issues can arise because these hierarchical models are amenable to the standard toolkit of significance tests, model simplification or multi-model inference taught in undergraduate statistics modules. Analysts, taught by frequentists to infer significance from *p*-values, commonly struggle to understand how to infer the importance of fixed and random effects reported by (G)LMMs. There is no simple answer here, but simple mistakes can be avoided. The perils and pitfalls we describe are focused on attempts to infer the importance of fixed-effect explanatory variables, but most remain relevant to attempts to understand the importance of hierarchical variance components.

Peril #1: (G)LMM *p*-values can be anti-conservative

A common worry for the frequentist statistician is the risk that the *p*-value of a test (i.e. the probability of finding signal as extreme, or even more extreme, than the observed signal,

supposing that the null hypothesis is true) might exceed 5% even when the null hypothesis is true. This situation, also known as an inflated type I error rate or being anti-conservative, risks the false conclusion of the importance of a fixed effect explanatory variable (false positive) and has contributed to a replication crisis in some research fields (*Forstmeier, Wagenmakers & Parker, 2017*). Anti-conservative inference from (G)LMMs can arise when sample sizes are small because maximum likelihood estimates of variance components become biased. This problem is fixed by Restricted Maximum Likelihood (REML) algorithms, which are the default settings for (G)LMMs in most software. However, REML models are not amenable to standard model simplification or tests of significance. It is common practice to convert REML models to Maximum Likelihood and simplify these using likelihood ratio tests. These tests follow chi-square distributions only approximately. As sample size increases, the approximation improves, but for small sample sizes the outcome tends to be anti-conservative. Neither can analysts safely fall back on information theoretic approaches such as the well-known Akaike Information Criterion to compare REML models. AIC is a function of the model's likelihood and the number of parameters estimated. However, REML likelihoods contain a nuisance factor that is dealt with differently by different software, meanwhile the number of parameters estimated is linked to the number of degrees of freedom associated with the random effects, and that number is hard to define (*Bolker et al., 2009*). Adjustments for AIC (such as AICc) that cope with small sample sizes are approximations and do not completely deal with this problem.

Guidance on what qualifies as a small sample size is lacking (*Gelman & Hill, 2006*), but it is safe to assume that some datasets in biology will suffer from these issues. Small datasets are also at risk from over-fitting of models, which can also lead to high rates of false positives (*Forstmeier, Wagenmakers & Parker, 2017*), and with more complex model structures it can be more challenging to identify situations where overfitting is an issue. These potential issues can arise regardless of the approach taken to statistical inference, although they can be exacerbated by model simplification (*Forstmeier & Schielzeth, 2011*; *Forstmeier, Wagenmakers & Parker, 2017*).

A simple simulation illustrates how this issue arises in small datasets. Imagine a simple treatment-control experiment that applies a treatment to half of the members of each of six families. Here, family ID is the random intercept term, and the mean value for the response variable for each family is drawn from a Normal distribution with mean = 0 and standard deviation = 1. We then simulate values for the response variable so that the experimental treatment itself has no effect on values of the response variable (mean = family mean and standard deviation = 1). Here, family ID is the dominant source of variation in the response variable, and the null hypothesis is true and should be accepted. We conducted a poorly replicated version of this experiment, in which only one individual from each family is measured in each experimental group (treatment and control). We then tested for statistical significance of the treatment effect using mixed-effects models and: (a) by determining if the 95% confidence interval of the treatment effect in the full model overlapped zero; (b) a likelihood ratio test that compared Maximum Likelihood models with and without the fixed effect of treatment; and

(c) a comparison of the AIC of the models with and without the treatment effect. We used a sigma threshold of 0.05 for approach (b) and considered the effect of treatment to be important for approach (c) when the AIC of the model including the fixed effect of treatment was more than two AIC units less than the null model. All three approaches produced high rates of false positives: in 122 out of 1,000 (12.2%) simulations, 105 out of 1,000 (10.5%) and 100 out of 1,000 (10%) respectively (with identical results from nlme and lme4). This is around double the 5% threshold desired by frequentists and indicates a high false positive rate. The analysis is anti-conservative. The problem of anti-conservatism fades with increasing sample sizes: when each of six families contribute ten control and ten treated individuals, the error rates recorded were 5.9%, 5.7% and 3.3% (5.1% for nlme) respectively. Consequently, while this is a pitfall to be aware of when using (G)LMMs in small datasets, it is likely to be less important with adequate replication.

Peril #2: (G)LMMs do not cure all types of pseudoreplication

Pseudoreplicated experimental designs are those in which fixed effects vary at a hierarchical scale higher than the measured subject. If, for example whole families are exposed to experimental treatment, and family members resemble each other beyond the impact of treatment, then individual members of families do not represent independent measures of the effect of treatment and should not be considered true biological replicates. In an extreme situation where there are many non-independent observations within each of a small number of random effect levels this could cause the number of residual degrees of freedom (those used to infer residual variance) to greatly exceed the number of experimental units to which treatments were applied (number of families). While, traditionally, F-ratio tests would reveal this problem, likelihood ratio tests hide it from the analyst because they use only the degrees of freedom associated with the explanatory variables (*Arnqvist, 2020*). AIC values also hide the problem by providing no information on degrees of freedom at all. When (G)LMMs are fitted in R this problem is managed by fitting fixed effects at the correct level of hierarchy. This deals with the pseudoreplication pitfall but can result in poor replication of the fixed effects if the number of random effect groups is low (Fig. 3). The model may struggle to partition variance explained by the random effects from unexplained (residual) variance. This results in a marked decline in the accuracy of model inference and increased risk of encountering anti-conservative p-values from likelihood ratio tests (see "Peril #1") with no way for the inexperienced analyst to check.

To demonstrate, we repeated our simulation from the previous section, but this time applied experimental treatments to whole families (so that every individual in each family was in the same experimental group). We simulated 600 observations in total, then in separate analyses we divided observations among 6, 8, 10, 12, 15, 20 or 24 families (with concomitant reduction in sample size per family, ranging from 100 to 25). Each family was allocated either to the treatment or the control group. In this design, the individuals in each family do not provide independent measures of the effect of treatment. As before, we simulated treatment to have zero effect so the null hypothesis is true and should be accepted. When the number of families was small we observed a high false
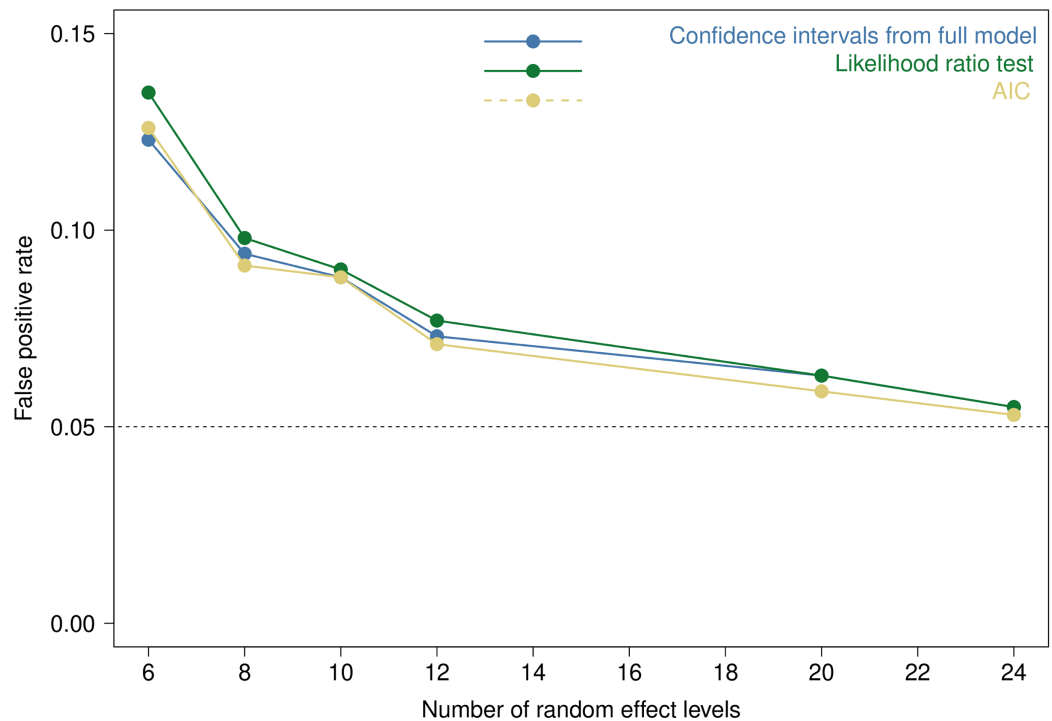
Figure 3 The false positive rate for group-level variables in linear mixed-effect models with small number of random effects using common forms of statistical inference. For each number of random effect levels (x axis) we simulated 1,000 datasets containing no effect of an experimental treatment applied at the group-level and recorded the number of times that a statistically significant effect of treatment was recorded using model estimates and 95% confidence intervals from the full model (blue); a Likelihood ratio test between the model with an effect of treatment and a null model (green); and ΔAIC > 2 between the two models (yellow). All code for this example is provided in the Supplemental Material. Full-size ◩ DOI: 10.7717/peerj.9522/fig-3

positive rate, even though the per-family sample size was large: the effect of treatment was statistically significant in >10% of simulations, except when a t-test using the mean and standard error of the treatment effect from the nlme model was used (Fig. 3). The false positive rate then declined to approximately 5% in cases where there more families were sampled (e.g. our simulation runs with 20 or 24 families). Our simulations therefore demonstrate that caution is required in interpreting the statistical significance of group-level explanatory variables when the number of random effect levels is small.

There are additional circumstances where using random effects to attempt to control for non-independence may not be straightforward. Imagine an experiment where we grow two genotypes of wheat in replicate planters containing varying, finite amounts of phosphate to test a hypothesis about the performance of each genotype in different conditions. You would be correct in assuming that measurements from within the same feeding trial are not independent. One potential option here would be to fit planter ID as a random effect to control for this. However, higher uptake by one genotype within an experimental block means less phosphate available for the other, causing the performance of each genotype within a planter to be negatively correlated. A standard 'blocked' (G)LMM with planter ID as a random effect would not control for this issue.

The risk of anti-conservatism does not mean that (G)LMMs have no value in pseudoreplicated designs. However, it does mean that the analyst must have a clear understanding of the hierarchical design of the model, in particular whether fixed effects are observation-level (apply uniquely to each observation in the dataset) or group-level (apply uniquely to each level of the random effect) (*Schielzeth & Nakagawa, 2013*). It is also important to be aware of the number of independent experimental units, the extent and nature of pseudoreplication and the ability of the statistical software to test the importance of signal at the correct level of the design hierarchy. Experimental units, that is those which can be claimed to be independent replicates of the experiment or survey design, must be replicated sufficiently to provide confidence in their inference *and* to avoid Type I errors (false positives).

Peril #3: too few random effect categories

The standard definition of a random effect is that it should describe membership of a group that is part of a random sample of a larger population of groups. Random effect hyperparameters infer the variance of means and slopes among groups. Inference of any variance parameter requires several independent replicates and improves with increasing sample size of the number of groups. Hence, accurate estimation of random effects can only occur when several groups are represented in a dataset (*Harrison, 2014, 2015*; *Harrison et al., 2018*). In practice, however, mixed effects models are commonly used to capture and absorb any non-independence due to group membership. When only few groups have been measured, inference of random effect hyperparameters will be poor; models might be degenerate; and there will be consequences for the correct inference of fixed effect parameters and tests of their importance. For example fitting sex as a random effect in a model implies the choice of calculating a variance using only two values: we wouldn't infer a simple standard deviation from a sample of two individuals, so we shouldn't use the (G)LMM algorithm to make the same crude inference. In general, it is suggested to only fit a categorical variable as a random effect if it has 5–6 levels or more (*Bolker et al., 2009*; *Harrison, 2015*).

We illustrate this problem in Fig. 4 using simple data simulations. A sample of 200 individuals is subdivided into between two and 40 groups. A single covariate has a positive effect on our response variable. Full code for the example is provided in the Supplemental Material. Estimates of random effect variance are biased small when there are fewer than six or seven groups (Fig. 4A), with the greatest problems caused by having only two or three groups. The inaccuracy of estimates of the random effect variance shows an exponential decline. This means that models tend to underestimate the standard error of the fixed-effect covariate's slope when there are too few random effect levels (Fig. 4B), indicating increased risk of Type I (false positive) errors when there are fewer random effect levels.

Peril #4: effect of predictor varies among random effect categories

Many studies that use (G)LMMs use a random intercepts model, in which the random effect measures or controls for variation among group means. However, there is no
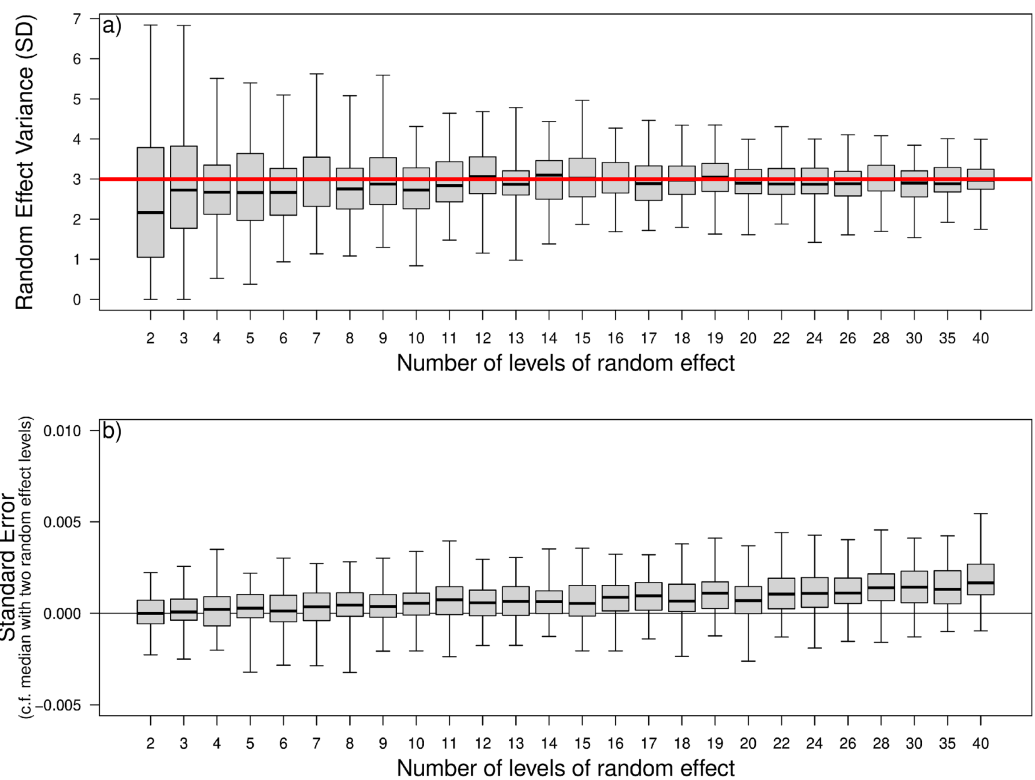
Figure 4 The effect of the number of levels of a random effect on (A) the random effect variance (displayed here as a standard deviation) in a random intercepts model, and (B) the standard error around the estimated effect size of a fixed effect covariate in the model. The red line in (A) indicates the true value of the between-group standard deviation. The true effect size for the covariate effect is one and the residual deviance had a standard deviation of three (see Supplemental Material).

Full-size ☒ DOI: 10.7717/peerj.9522/fig-4

guarantee that this is a sufficient control for variation among groups, especially when modelling a covariate whose effect may itself vary among groups (*Schielzeth & Forstmeier, 2008*; *Barr et al., 2013*; *Bell, Fairbrother & Jones, 2018*). The use of random slopes models is advisable in these contexts (*Barr et al., 2013*). Random intercepts models risk anti-conservative tests of importance, and typically predict responses of unsampled groups worse than models that include random slopes (*Bell, Fairbrother & Jones, 2018*). To provide an ecological example, we simulated a relationship between oxidative stress and parasite burden in 20 populations of wild amphibian ($n = 20$ frogs measured per population), where parasite burden is a non-integer measure of genomic equivalents of a parasite. We allowed both mean parasite burden (random intercepts) and the strength of relationship between parasites and oxidative stress (random slopes) to vary by population. We also randomly assigned populations to a two-level factor as either 'treatment' or 'control', so having no effect on parasite burden. We fitted a 'maximal' model with Gaussian error structure allowing for an interaction between oxidative stress and treatment, and containing either only random intercepts for population ID, or both random intercepts and random slopes for oxidative stress given population. After 10,000 iterations, when fitting only random intercepts, the model erroneously estimated the
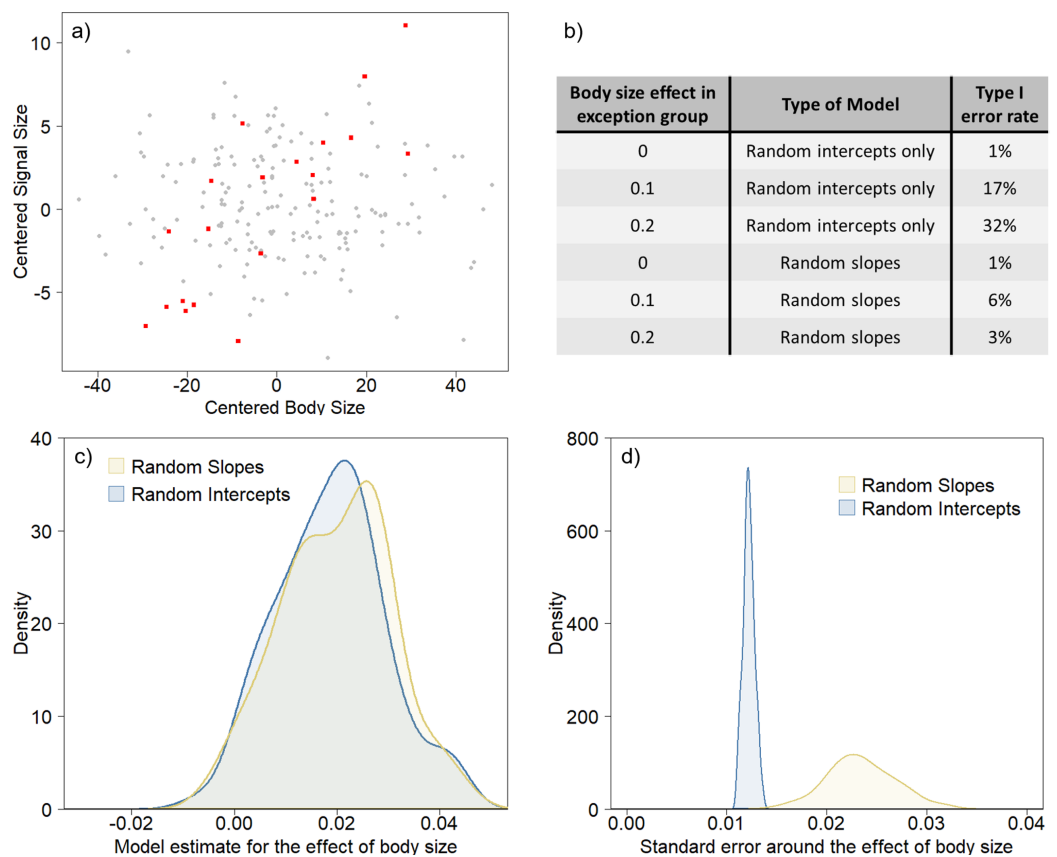
Figure 5 Random slopes models avoid the exaggerated influence of anomalous groups. Simulated datasets contain 10 groups, only one of which harboured a relationship between body size and status signal size (A). Grey points represent data from the nine typical groups, and red points data from the anomalous group. For a variety of anomalous-group slopes (0, 0.1, 0.2) we measured the proportion of models, out of 100 simulations, that inferred a significant global relationship between body size and signal size (B). The difference in type I error rate was not caused by differences in estimate for the overall relationship between body size and signal size (C), but instead an underestimate of the error around this estimate in the random intercepts model (D). In these plots the polygons are density plots of inferences of global slope from the random intercepts models (blue) and the random slopes models (yellow). All R code for this example is provided in the Supplemental Material.

Full-size ☑ DOI: 10.7717/peerj.9522/fig-5

treatment * oxidative stress interaction to be significant (95% confidence intervals not crossing zero) in 30% of cases. Fitting a random slopes model reduced the incidence of Type I (false positive) errors to 6.7%, far closer to the desired 5% rate. For further details on this phenomenon, and a more detailed set of simulations, see *Schielzeth & Forstmeier (2008)*. Clearly, not including random slopes results in elevated type I (false positive) error rates.

An important risk of random intercepts models is that unusual groups can cause false inference of whole-population effects. Imagine a test for a linear relationship between body size and the size of a status signal in a hypothetical bird species consisting of 10 different subspecies. There is a positive body size-signal size relationship for only one of these subspecies (Fig. 5A). A random intercepts model may conclude that there is a

positive effect of body size on status signal size across *all* subspecies if the relationship in that single subspecies is sufficiently strong (Fig. 5B). In contrast a random slopes model that controls for differences in this relationship among subspecies is likely to detect no overall effect, but also show that the slope for one subspecies was unusual. The difference between the random slopes and random intercepts models does not occur because of differences in their estimates of the effect of body size (Fig. 5C) but instead because the random intercepts model consistently underestimates the standard error around this estimate (Fig. 5D). Note that often random slopes models can return low estimates of variance for the among-group slopes, but this is to be expected if there is only one group with a strong relationship with the response. This is especially the case if the random effects sample size is large. Measuring the 'importance' of random effects based on these variance components is a contentious issue (*Harrison et al., 2018*), but random effects shouldn't *automatically* be immediately discounted/removed just because a variance component appears negligible.

Peril #5: correlations between fixed and random effects, and informative group sizes

Mixed models assume that values of fixed effect variables are independent of the groups (clusters) used as levels of a random effect. When this assumption is violated, the models suffer 'confounding by cluster' (*Seaman, Pavlou & Copas, 2014a*). Confounding by cluster will therefore be a potential problem when there is both within- and between-group variation in the values of an explanatory variable. It can lead to misleading estimates for fixed effect parameters and biased estimations of variance components (*Neuhaus & McCulloch, 2006*). A number of potential solutions have been suggested (*Seaman, Pavlou & Copas, 2014a*), although one of the most accessible and intuitive is to decompose the effect of a fixed effect variable into between-group (and therefore group-level) and within-group (observation-level) covariates (*Neuhaus & McCulloch, 2006*). For example a study determining the relationship between fitness and body size in a population made up of multiple social groups could contain a group-level effect of mean body size, and an individual-level effect that is group-mean centred, so that it describes the size of that individual relative to other individuals in the group.

A similar problem can arise if the cluster/grouping size (number of samples for each level of the random effect) is correlated with one or more fixed effect variables (*Seaman, Pavlou & Copas, 2014b*). For example if modelling the association between maternal stress (response variable) and offspring weight at fledging (explanatory variable) in a passerine bird, clutch ID would typically be included to control for variation in offspring weight among different breeding attempts (as might be caused by genetic differences, parental effects or the shared environment). However, if females that are more stressed tend to lay smaller clutches then the size of the cluster is informative. Informative cluster sizes can result in biased inference, especially when fixed and random effects are correlated (*Neuhaus & McCulloch, 2011*). This is a particular problem for models that only include random intercepts, where failing to account for informative cluster sizes can lead to biased estimates for the intercept. However, for most covariates

model estimates are unaffected by the presence of informative cluster sizes (*Neuhaus & McCulloch, 2011*). One potential solution for dealing with informative cluster sizes is by including cluster size as a covariate, however this is not a sensible choice when cluster size lies on the causal pathway linking a particular explanatory variable to a response variable (*Seaman, Pavlou & Copas, 2014a*): inclusion of the cluster size as a fixed covariate risks masking the real influence of the main explanatory variable. Another more complex approach is to jointly model the response variable of interest and cluster size (*Dunson, Chen & Harry, 2003*; *Chen, Zhang & Albert, 2011*; *Seaman, Pavlou & Copas, 2014a*).

Peril #6: random effect category means are not normally distributed

In many situations random effect means or slopes deviate from a Normal distribution. They might be clustered along the *y*-axis, or might be skewed or heavy-tailed. Skew and kurtosis are rarely considered by biologists, but are important when comparing variances among samples (*Hosken, Buss & Hodgson, 2018*). Typically, the random effect hyperparameter used in (G)LMMs is the standard deviation of a Normal distribution. It is possible to specify alternative hyperparameters and allow random effects to have non-Normal distributions (*Zhang et al., 2008*; *Molenberghs et al., 2010*, *2012*; *Fabio, Paula & de Castro, 2012*) but this requires non-standard modelling algorithms. In general, both LMMs and GLMMs have been found to be impressively robust against misspecification of the random effects distribution (*McCulloch & Neuhaus, 2011*; *Neuhaus, McCulloch & Boylan, 2013*; *Schielzeth et al., 2020*). The estimates of fixed effect parameters for individual-level variables are particularly robust (*McCulloch & Neuhaus, 2011*), while estimation of random effects and variances is more susceptible to misspecification (*McCulloch & Neuhaus, 2011*). However, when fixed effects are correlated with random effects, and if other assumptions of mixed models are not met, then misspecification of the random effects distribution can influence parameter estimates (*Neuhaus & McCulloch, 2011*). Diagnostic tools such as those found in the R package 'sjPlot' (*Ludecke, 2019*) allow users to diagnose the validity of their assumption of normality of random effects, and should be used where appropriate. While LMMs and GLMMs are typically robust to misspecification, other approaches are available when these assumptions are violated, for example the application of mixture models (*Hamel, Yoccoz & Gaillard, 2017*). Direct comparison of choices of random effect distributions can be made using Bayesian mixed effects models coded in standard software such as JAGS (*Plummer, 2003*) or Stan (*Carpenter et al., 2017*).

Peril #7: random effect categories used as a proxy for covariance structure

It is not unusual for non-independence among residuals, caused by spatial, temporal or phylogenetic autocorrelation, to be modelled using categorical random effects (e.g. region, year or genus). Despite increased prevalence of regressions that explicitly model covariance structures (*Housworth, Martins & Lynch, 2004*; *Dormann et al., 2007*; *Miller, Franklin & Aspinall, 2007*; *Hadfield & Nakagawa, 2010*; *Ives & Helmus, 2011*; *Rousset & Ferdy, 2014*), many researchers elect to control for these dependencies with
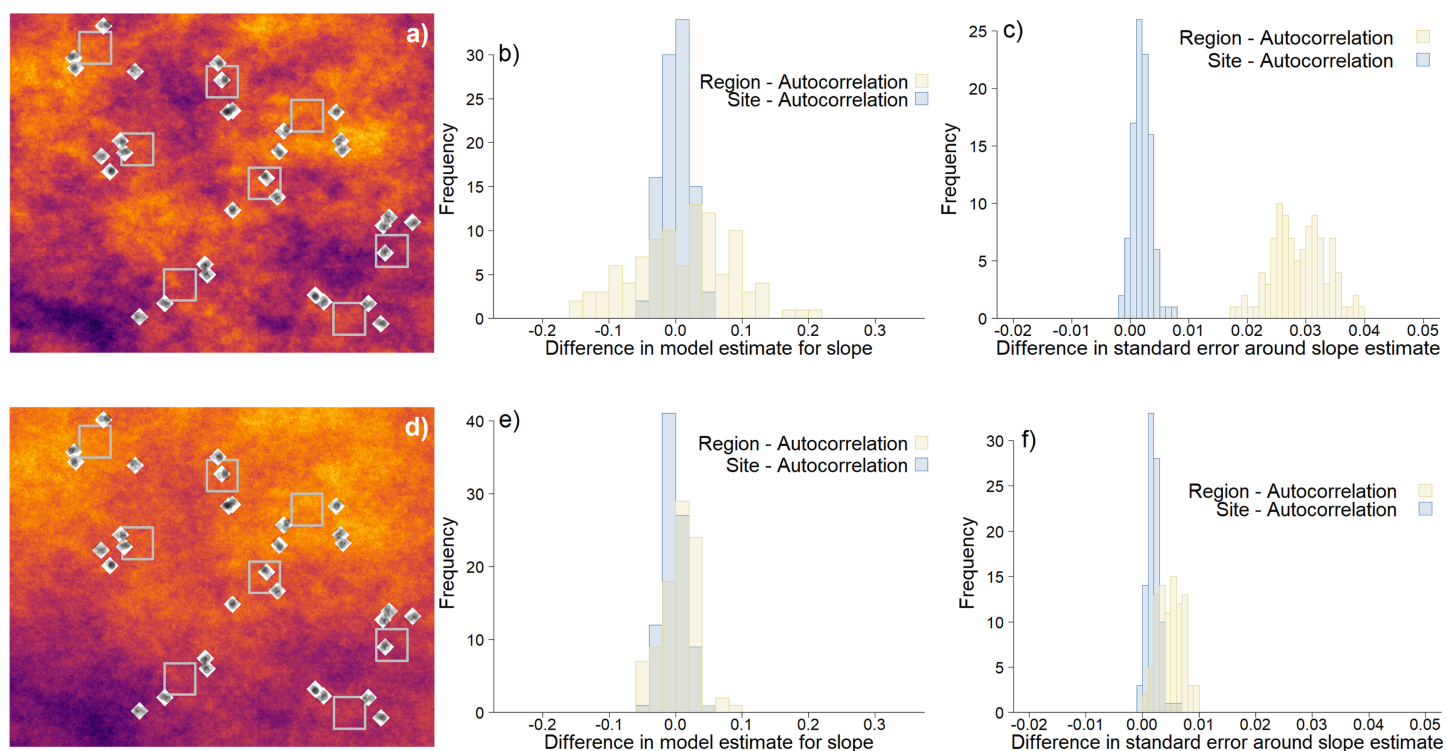
Figure 6 An illustration of the appropriateness of categorical effects to control for autocorrelated data. In this example dataset, a response variable measuring height varies according to temperature and there is spatial autocorrelation in the residual error at either a fine spatial scale (A) or a broader spatial scale (D). Sampled individuals (small, semi-transparent black points) are associated with particular sites (white diamonds), with 10 individuals per site. Each site is associated with a particular region (centre marked with light grey square), with four sites sampled in each region (all depicted in (A and D)). We fitted three models, each with height as the response variable and temperature as the explanatory variable: a generalised least squares model with a spatial correlation structure (for comparison), an LMM with site as a random effect (comparison in blue) and an LMM with region as a random effect (comparison in yellow). (B and E) show the difference between model estimates for the effect of temperature between the LMMs and the model with spatial autocorrelation structure for the case with fine scale spatial autocorrelation and broader scale autocorrelation respectively. (C and F) show the same comparison for standard error around those model estimates. All R code for this example is provided in the Supplemental Material. Full-size ◩ DOI: 10.7717/peerj.9522/fig-6

categorical variables as an alternative. While some coarse control for spatial (or temporal) structure is better than none at all, the success of using a categorical random effect can vary depending on the nature of the true covariance structure in the data. First, random effects do not control for all types of non-independence and are useful only when there is positive autocorrelation (i.e. measurements closer together in time or space tend to be more similar). Second, the success of using a categorical proxy for a covariance structure will depend on how the granularity of the effect used (e.g. site ID vs. region ID) compares with the scale of any autocorrelation.

We demonstrate here that the success of categorical proxies for continuous covariation depends on how the granularity of the random effect compares with the scale of any autocorrelation. We simulated two Gaussian fields that influence the residual error in the relationship between a single explanatory variable (Temperature) and single response variable (Height). The first (Fig. 6A) varied at a much finer scale than the second (Fig. 6D). We varied the slope of the relationship, so that we could examine the impact on model performance at a various effect sizes. There were 320 measurements in total, occurring

in eight regions that each contained four sites, with 10 individuals measured per site. We fitted three models: one directly modelled spatial covariance in the residuals (using Generalised Least Squares); one using site as a categorical random effect; and one using region as a categorical random effect. For the model in which the spatial field varied at a finer scale, model estimates from the models that used site as a random effect were in much closer agreement with the explicitly spatial model (Fig. 6B) and much more precise (Fig. 6C) than those that used region as a random effect. In contrast, when the scale of the spatial field was broader, the accuracy and precision of model estimates was much more similar (Figs. 6E and 6F). Biases in the accuracy and precision of model predictions will affect Type I (false positive) and Type II (false negative) error rates. This helps illustrate the benefit of selecting a categorical effect to control for spatial or temporal structure in the data that is as well matched to the autocorrelation in the data as possible.

Detecting and resolving problems with mixed model estimation

The issues identified here can all be identified with adequate data checks and checks of model fit (*Zuur & Ieno, 2016*; *Harrison et al., 2018*).

First, we recommend having a clear idea of the structure and hierarchy of the model. Identifying categorical variables with too few levels for random effect estimation can help avoid Peril 3, while identifying the level of the hierarchy that fixed effect variables apply at (observation-level vs group-level) can help identify situations where there is a risk of Peril 2 having an impact. Similarly, a clear idea of the hypothesis and model structure can help identify scenarios where a random slopes model might be more appropriate than a random intercepts model (Peril 4).

Second, we emphasise the importance of data exploration (*Zuur, Ieno & Elphick, 2010*). Plots of the raw data can help identify unusual random effect categories (helping to avoid Peril 6), small sample sizes (Peril 1) or confounding by cluster and informative cluster sizes (both part of Peril 5). Similarly, raw data plots might help provide an initial idea of the scale of any spatial or temporal structure to the data and so be informative in deciding an appropriate random effect (Peril 7).

Third, we highlight the value of conducting full model checks (*Zuur & Ieno, 2016*; *Harrison et al., 2018*). Plots of residuals for each level of the random effect are important in identifying heteroscedasticity in the residuals between levels of the random effect, or relationships between residuals and fitted values that necessitate random slopes models (Peril 4). Semi-variograms check the need for explicit modelling of covariance structures to avoid Peril 7 (*Fletcher & Fortin, 2018*). The simplest method to assess the extent to which the random effects distribution differs from normal (to avoid Peril 6) are quantile–quantile plots. However, various other diagnostic tools are available (*Drikvandi, Verbeke & Molenberghs, 2017*; *Efendi et al., 2017*). We demonstrate a comprehensive but simple approach to model checking in the Supplemental Material. We provide code to allow users to be confident in generating these checks themselves as well as using functions provided by existing R packages. New R packages such as DHARMa (*Hartig, 2020*) are making these model checks more available and accessible even for more complex GLMMs.

There are also now R packages that make more sophisticated modelling approaches available to a wider audience. For example brms (*Bürkner, 2017*) provides an interface to the Bayesian statistical model fitting software Stan (*Carpenter et al., 2017*). Users can exploit coding syntax similar to lme4, and an array of model specifications are available that deal with various pitfalls described here. Similarly, glmmTMB (*Brooks et al., 2017*) provides software to facilitate the fitting of more complex error distributions and correlation structures using similar syntax to lme4.

(G)LMMs are reassuringly robust to some of the individual pitfalls we have described here *Schielzeth et al. (2020)*. However, there is very little understanding of the combined impact of multiple violations of the core assumptions: there is plenty of scope for simulation studies of the power and anticonservatism of mixed effects models. The pitfalls described here exist and are prevalent in the biological literature, with little opportunity for readers to check the validity of published models and results. Hence, raising awareness of these potential pitfalls and considering them together can help researchers use mixed models more assuredly and avoid pitfalls that impact on their statistical inference.

CONCLUSIONS

Mixed modelling approaches are firmly embedded as state-of-the-art for the analysis of biological data. As these methods become available to a growing user base it is necessary to reveal the perils and pitfalls associated with them (*Arnqvist, 2020*). Often with biological datasets it is not possible to meet the statistical assumptions of these models perfectly and it may be necessary to make compromises in model design. We have provided guidance on a set of typical perils and pitfalls by providing insights from the statistical literature and illustrating key points using simulated case studies and examples. While (G)LMMs can be robust to some of the perils we describe (*Schielzeth et al., 2020*), there remains little understanding of their combined impacts. Readers can only assess the quality of inference from mixed effects models if their hierarchical structures, levels of true replication, and checks of validity, are described clearly and honestly. This review will help readers be more aware of some of the key perils in mixed model design and so improve statistical inference when testing biological hypotheses. We hope that by providing an overview of these perils, we can help researchers feel better informed and more confident that they are using mixed modelling approaches to draw correct conclusions from their data.

ADDITIONAL INFORMATION AND DECLARATIONS

Grant Disclosures

Competing Interests

Xavier Harrison is an Academic Editor for PeerJ.

Author Contributions

- Matthew J. Silk conceived and designed the experiments, performed the experiments, analysed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Xavier A. Harrison conceived and designed the experiments, performed the experiments, analysed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- David J. Hodgson conceived and designed the experiments, analysed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The R code used to conduct all simulations in the paper is available in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.9522#supplemental-information.

REFERENCES

Arnqvist G. 2020. Mixed models offer no freedom from degrees of freedom. *Trends in Ecology & Evolution* **35(4)**:329–335 DOI 10.1016/j.tree.2019.12.004.

Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language* **68(3)**:255–278 DOI 10.1016/j.jml.2012.11.001.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67(1)**:1–48 DOI 10.18637/jss.v067.i01.

Bell A, Fairbrother M, Jones K. 2018. Fixed and random effects models: making an informed choice. *Quality & Quantity* **53(2)**:1051–1074.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* **24(3)**:127–135 DOI 10.1016/j.tree.2008.10.008.

Brooks ME, Kristensen K, Van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Machler M, Bolker BM. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal* **9(2)**:378–400 DOI 10.32614/RJ-2017-066.

Bürkner P-C. 2017. brms: an R package for Bayesian multilevel models using stan. *Journal of Statistical Software* **80(1)**:1–28.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. 2017. Stan: a probabilistic programming language. *Journal of Statistical Software* **76(1)**:1–32.

Chen Z, Zhang B, Albert PS. 2011. A joint modeling approach to data with informative cluster size: robustness to the cluster size model. *Statistics in Medicine* **30(15)**:1825–1836 DOI 10.1002/sim.4239.

Dormann CF, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, Davies RG, Hirzel A, Jetz W, Daniel Kissling W. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30(5)**:609–628 DOI 10.1111/j.2007.0906-7590.05171.x.

Drikvandi R, Verbeke G, Molenberghs G. 2017. Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics* **73(1)**:63–71 DOI 10.1111/biom.12551.

Dunson DB, Chen Z, Harry J. 2003. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **59(3)**:521–530 DOI 10.1111/1541-0420.00062.

Efendi A, Drikvandi R, Verbeke G, Molenberghs G. 2017. A goodness-of-fit test for the random-effects distribution in mixed models. *Statistical Methods in Medical Research* **26(2)**:970–983 DOI 10.1177/0962280214564721.

Fabio LC, Paula GA, De Castro M. 2012. A Poisson mixed model with nonnormal random effect distribution. *Computational Statistics & Data Analysis* **56(6)**:1499–1510 DOI 10.1016/j.csda.2011.12.002.

Fletcher R, Fortin M-J. 2018. Spatial dependence and autocorrelation. In: Fletcher R, Fortin M-J, eds. *Spatial Ecology and Conservation Modeling*. New York: Springer, 133–168.

Forstmeier W, Schielzeth H. 2011. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology* **65(1)**:47–55 DOI 10.1007/s00265-010-1038-5.

Forstmeier W, Wagenmakers E-J, Parker TH. 2017. Detecting and avoiding likely false-positive findings: a practical guide. *Biological Reviews* **92(4)**:1941–1968 DOI 10.1111/brv.12315.

Gelman A, Hill J. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

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

Hamel S, Yoccoz NG, Gaillard J-M. 2017. Assessing variation in life-history tactics within a population using mixture regression models: a practical guide for evolutionary ecologists. *Biological Reviews* **92(2)**:754–775 DOI 10.1111/brv.12254.

Harrison XA. 2014. Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ* **2**:e616 DOI 10.7717/peerj.616.

Harrison XA. 2015. A comparison of observation-level random effect and beta-binomial models for modelling overdispersion in binomial data in ecology & evolution. *PeerJ* **3**:e1114 DOI 10.7717/peerj.1114.

Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CED, Robinson BS, Hodgson DJ, Inger R. 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **6(1)**:e4794 DOI 10.7717/peerj.4794.

Hartig F. 2020. *DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models*. R package version 0.3.1. *Available at http://florianhartig.github.io/DHARMa/*.

Hosken DJ, Buss DL, Hodgson DJ. 2018. Beware the *F* test (or, how to compare variances). *Animal Behaviour* **136**:119–126 DOI 10.1016/j.anbehav.2017.12.014.

Housworth EA, Martins EP, Lynch M. 2004. The phylogenetic mixed model. *American Naturalist* **163(1)**:84–96 DOI 10.1086/380570.

Ives AR, Helmus MR. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs* **81(3)**:511–525 DOI 10.1890/10-1264.1.

Ludecke D. 2019. *siPlot: data visualization for statistics in social science: R package version 2.8.1.* R package version 2.8.3. *Available at https://CRAN.R-project.org/package=sjPlot.*

McCulloch CE, Neuhaus JM. 2011. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science* **26(3)**:388–402 DOI 10.1214/11-STS361.

Miller J, Franklin J, Aspinall R. 2007. Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling* **202(3–4)**:225–242 DOI 10.1016/j.ecolmodel.2006.12.012.

Molenberghs G, Verbeke G, Demétrio CGB, Vieira AMC. 2010. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science* **25(3)**:325–347.

Molenberghs G, Verbeke G, Iddi S, Demétrio CGB. 2012. A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate Analysis* **111**:94–109 DOI 10.1016/j.jmva.2012.05.005.

Neuhaus JM, McCulloch CE. 2006. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society: Series B* **68(5)**:859–872 DOI 10.1111/j.1467-9868.2006.00570.x.

Neuhaus JM, McCulloch CE. 2011. Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika* **98(1)**:147–162 DOI 10.1093/biomet/asq066.

Neuhaus JM, McCulloch CE, Boylan R. 2013. Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. *Statistics in Medicine* **32(14)**:2419–2429 DOI 10.1002/sim.5682.

Pinheiro J, Bates D, DebRoy S, Sarkar D. 2014. *R Core Team, 2014 nlme: linear and nonlinear mixed effects models.* R package version 3.1-117. *Available at http://CRAN.R-project.org/package=nlme.*

Plummer M. 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.* 10.

R Core Team. 2019. *R: a language and environment for statistical computing.* Vienna: The R Foundation for Statistical Computing. *Available at http://www.R-project.org/.*

Rousset F, Ferdy J-B. 2014. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography* **37(8)**:781–790.

Schielzeth H, Dingemanse NJ, Nakagawa S, Westneat DF, Allegue H, Teplitsky C, Reale D, Dochtermann NA, Garamszegi LZ, Araya-Ajoy YG. 2020. Robustness of linear mixed-effects models to violations of distributional assumptions. Epub ahead of print 12 June 2020. *Methods in Ecology and Evolution* DOI 10.1111/2041-210X.13434.

Schielzeth H, Forstmeier W. 2008. Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology* **20(2)**:416–420 DOI 10.1093/beheco/arn145.

Schielzeth H, Nakagawa S. 2013. Nested by design: model fitting and interpretation in a mixed model era. *Methods in Ecology and Evolution* **4(1)**:14–24 DOI 10.1111/j.2041-210x.2012.00251.x.

Seaman S, Pavlou M, Copas A. 2014a. Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in Medicine* **33(30)**:5371–5387 DOI 10.1002/sim.6277.

Seaman SR, Pavlou M, Copas AJ. 2014b. Methods for observed-cluster inference when cluster size is informative: a review and clarifications. *Biometrics* **70(2)**:449–456 DOI 10.1111/biom.12151.

Zhang P, Song PX-K, Qu A, Greene T. 2008. Efficient estimation for patient-specific rates of disease progression using nonnormal linear mixed models. *Biometrics* **64(1)**:29–38 DOI 10.1111/j.1541-0420.2007.00824.x.

Zuur AF, Hilbe JM, Ieno EN. 2013. *A beginner's guide to GLM and GLMM with R: a frequentist and Bayesian perspective for ecologists.* Newburgh: Highland Statistics Limited.

Zuur AF, Ieno EN. 2016. A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution* **7(6)**:636–645 DOI 10.1111/2041-210X.12577.

Zuur AF, Ieno EN, Elphick CS. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* **1(1)**:3–14 DOI 10.1111/j.2041-210X.2009.00001.x.