We would like both the editor and the reviewers for continuing to work with us on this manuscript. We particularly appreciate the thoughtful comments pertaining to precision of language and terminology, which we have now worked hard to address in the revised version. With the help of the reviewers we now feel this manuscript has been even further improved as an introductory guide to the world of mixed effects modelling for biologists. We have responded to each individual comment below.

Editor's Comments

MAJOR REVISIONS

I would like to thank the reviewers for yet again providing extremely thoughtful and useful comments on this manuscript. I feel that the authors have set themselves a very ambitious task in covering so much material in a single manuscript, even with possible future manuscripts to flesh out aspects that cannot be fully addressed here, in a technically correct manner whilst preserving the necessarily readability for their audience, and responding to the reviewers' comments will help in achieving this goal.

While your preference for a single manuscript is clear, you may of course wish to revisit the option to split this manuscript into two separate manuscripts in your response at this stage. However, I also appreciate that we risk running in circles over the same ground with this, and perhaps some other topics, and so I will ask that you carefully consider all of the reviewers' points and for each, either make changes or provide a justification for not making changes. I agree with the reviewers that the revised version of the manuscript is stronger and hope that it will become stronger still with further revision.

I will draw the authors' attention in particular to the (major) points 2, 3, and 4 from the third reviewer, which need careful consideration, alongside carefully addressing reviewer two's first point. Reviewer two's third point raised an important practical issue (ID labelling) that I think is well worth addressing as I regularly encounter PhD candidates who are unclear whether they are numbering within or across clusters. Needless to say, all of the three reviewers have provided valuable feedback and all of their comments should be addressed.

I think that, in general, you have done well to avoid many overly definite statements in the manuscript, but at line 121, I wonder if a caveat along the lines that “As with all heuristics, there will be situations where these recommendations will not be optimal. If the researcher has concerns about the appropriateness of a particular strategy for a given situation, they should consult with a statistician who has experience in this area.” would be useful to remind the reader that relatively few rules, if indeed any, in statistics are universal and to encourage them to collaborate with specialists as appropriate. This would not, of course, relax requirements around technical correctness, but it would, I hope, alert the reader to proceed with due care in this challenging area.

>>>We have added an edited version of this sentence at the end of the first paragraph as requested “As with all heuristics, there may be situations where these recommendations will not be optimal, perhaps because the required analysis or data structure is particularly complex. If the researcher has concerns about the appropriateness of a particular strategy for a given situation, we recommend that they consult with a statistician who has experience in this area.”

>>>We appreciate that some of the reviewers strongly advocate splitting manuscript into two and have responded to Reviewer 3 on this issue below. We do not wish to split the manuscript and draw the editor’s attention to the overwhelmingly positive feedback we have received from scientists on the two preprint versions of the manuscript. To date the preprint has had nearly 3000 visitors and over 1700 downloads, and we expect this substantial interest in the paper to tield a large number of citations. The manuscript was always meant to be an all-inclusive roadmap to both mixed effects model fitting and methods of model selection, leaving the reader to seek out finer detail elsewhere, and we believe the manuscript functions best in this format.

We have made thorough revisions to the manuscript in line with the extensive and helpful reviewers comments.

Reviewer 1 (Anonymous)

Basic reporting

no comments

Experimental design

no comments

Validity of the findings

no comments

Comments for the Author

I have reviewed an earlier version of this MS. The authors have addressed my comments and it seems they did so for others' comments. I have noticed that one of the other referees raised a point about missing data, which I did not comment on. Actually, missing data on mixed models are still underdeveloped and an emerging research topic. There are a couple of papers that directly discuss missing data in mixed model contexts, but which are not currently cited in the MS but relevant.

Nakagawa, S. (2015) Missing data: mechanisms, methods and messages In: Ecological Statistics: contemporary theory and application (eds. Fox, G. A., Negrete-Yankelevich, S. & Sosa, V. J.). Oxford University Press, Oxford. pp. 81-105

Noble, D. W. A. & Nakagawa, S. (submitted) Planned missing data design: stronger inferences increased research efficiency and improved animal welfare in ecology and evolution. bioRxiv

https://www.biorxiv.org/content/early/2018/01/11/247064

The latter one includes an interesting example of how planned missing data design can be used in a repeated measurement design (which is modelled by mixed models). I thought the authors may be interested. No need to cite these but the

authors would like to note that missing data on mixed models are topics under development.
>>>Thank you for your continued work on the manuscript. We have added both papers to the further reading of the 'Missing Data' section

---

Reviewer 2 (Anonymous)

Basic reporting

Language and presentation are appropriate

Experimental design

Not applicable

Validity of the findings

Not applicable

Comments for the Author

In this paper, which I reviewed previously, Harrison et al. provide an overview of considerations associated with (generalised) linear (mixed) modelling in ecological datasets. Overall the paper has much improved from the previous version, and will likely make a valuable contribution to the literature. I feel that a few of my previous comments still apply to the new version, so I hope the details below clarify my concerns for the authors:

1. Regarding the simulation analysis, I accept the authors' response about choosing not to run more models. However, I am still concerned about the mixing of NHST and IT philosophies in the interpretation. Specifically, there is heavy reliance on interpreting results as "Type I error" – strictly speaking one cannot (or perhaps,

should not) falsely reject a true null hypothesis with an information theory approach, because one is not doing null hypothesis significance testing. Of course, I agree with the authors that it is important to consider how often one might draw a false conclusion about whether a given variable is an important predictor of a response. However, it is important not to assess models built under IT using NHST. A better approach would be for the authors to use the inference strategies they advocate elsewhere in the paper, such as testing model assumptions, reporting R2, and basing interpretation on the presence and size of effects in the final model (see also my comment below on RI) – would these metrics collectively lead a researcher to erroneously conclude that they had settled on an appropriately parameterised final model that suggests strong effects? Similarly, the sentence at L944-945 presents a very frequentist viewpoint on model selection, which is inappropriate in a discussion of information theoretic approaches to model selection.

>>> In response to this comment and that from Reviewer 3, we have now removed the simulations from the manuscript. We have also adjusted our terminology with respect to 'Type I error' to avoid giving the impression that we are mixing analysis paradigms of information theory and frequentist statistics. We now refer to the inclusion of uninformative parameters following the terminology in Arnold 2010 J Wildlife Management.

2. Regarding the use of delta-6 or 95% summed weights, I felt that the authors response to my previous comment missed my point somewhat, which is that delta values and summed weights represent two (albeit slightly) different approaches. If one wants to "guarantee" 95% weight, then they should use 95% summed weight. For example, in the paper cited by the authors, model "M4" required delta-6.5 to encompass the best EKLD model 95% of the time (Richards, 2008, p223). While "at least delta-6" may be a good general guideline, this is not the same as "guaranteeing 95% weight". See also Burnham and Anderson, 2002, p78, for explanation of how delta values represent relative evidence.

>>>We have removed the word 'guarantee' from the section and reworded it slightly, as we agree that it gives the impression that delta-6 will always give at least 95% summed probability/weights. We have also highlighted that using 95% summed weights is a slightly different approach to using delta-scores.

3. New material has been added on the specification of random factors: I believe

the lme4 specification “Clutch Mass ~ Foraging Rate + (1|Woodland) + (1|Female ID)” should achieve the same goal as the model on L402 (should this be M5?), if all female IDs are unique. Ambiguity can arise if the females are labelled (say) 1 – 10 in woodland 1, and 1 – 10 in woodland 2: are there 10 females, each breeding at two sites, or 20 females? Labelling females as 1 – 20 (if the latter) alleviates the ambiguity between crossed and nested random factors, and it would be worthwhile explaining this. See for example lme4.r-forge.r-project.org/book/Ch2.pdf

>>> We thank the reviewer for highlighting the opportunity to talk about uniquely labelled factor levels for random effects. We have now added a section on this topic:

“We advocate that researchers always ensure that their levels of random effect grouping variables are uniquely labelled. For example, females are labelled 1 – *n* in each woodland, the model will try and pool variance for all females with the same code. Giving all females a unique code makes the nested structure of the data implicit, and a model specified as ~ (1| Woodland) + (1|FemaleID) would be identical to the model above. “

4. Considering the revised section 5 (p34), I think the authors are overly dismissive of “relative importance” statistics. Unfortunately, the vernacular definition of “importance” differs somewhat from the technical definition; “relative importance” (sum of weights) informs which parameters are likely to occur in the best model, not their effect sizes. Indeed, IT-based model selection determines two things: the probability that variables should be present in the model (taking model selection uncertainty into account) and, if variables are included, what their effect sizes are. RI can assist inference around the former. I accept that there is discussion about the effectiveness and interpretation of RI in the literature (e.g. Giam and Olden, 2016; Galipaud et al., 2017), but I am concerned that the tone of section 5 might suggest to readers that they should ignore RI altogether. My view is that inference in an information theoretic framework should take into account (or at least, report!) all measures of quantitative evidence about the models, including RI and parameter estimates. For example, a predictor with RI of 1 and a small effect size may or may not be biologically “important”, while a low RI would likely indicate poor support, even if an associated effect is large. These statistics are additional tools for inference.

>>> We agree with the reviewer and have added a couple of sentences to section 5 to encourage readers to report all quantitative evidence pertaining to information theoretic modelling:
"However, summed Akaike weights for variables in top model sets still represent useful quantitative evidence; they should be reported in model summary tables, and ideally interpreted in tandem with model averaged effect sizes for individual parameters. "

5. A note on the suggested model specifications (p5, p6, p8): it is great these are provided, as knowing how to specify a model in R can be very useful to new users. These models are set within a discussion that considers the differences among models, so it would be helpful to see how the model outputs change depending on the model specifications. This could be easily obtained using a "built in" dataset in R, and presented as supplementary material with a brief commentary on how to interpret each output. For example, the output from M1 and M2 could be narrated in terms of the interpretation of output values associated with "group". The same approach could be applied to M3 and M4. As an aside, a more complete discussion of for the use of random slopes would be provided if the authors explain the circumstances under which one would prefer model M4 over a model specified as "glmer(successful.breed ~ 1 + (body.mass|sample.site)", which is sufficiently similar to model M2 that I think it is worth noting (see also e.g. Gelman and Hill, 2007, p259).
>>>We agree that this would be a useful resource for researchers. However, we are currently working on a more extensive version of the above suggestions as a separate paper designed to help researchers interpret random effects from all of these types of models fitted to complex ecological data. We thank the reviewer for this helpful suggestion and for their continued work on improving the manuscript.

6. A note on whether removing models from the set makes interpretation of the Akaike weights "difficult" (L1011): as pointed out by Burnham and Anderson (2002, p75), adding and removing models from a set requires weights be recalculated. From a practical perspective, it is straightforward to calculate the weights of any model set by manually making the list of models for further analysis with MuMIn, or by using 'get.models' (MuMIn) on an existing 'dredge' object – see the MuMIn

documentation for more info. In fact, wi values can even be manually recalculated relatively easily too (using the formula in Burnham and Anderson, 2002, p75). Explicitly specifying the model set is what the authors are indirectly recommending when they suggest that one should consider whether the "all-subsets selection" approach is appropriate, and so mentioning these methods for re-calculating weights would overcome the "difficulty" and provide readers the flexibility of choosing their own model set.

>>> We agree that it is simple to recalculate model weights from a user-chosen set of models. Because model weight recalculation is done automatically in the *MuMIn* package when subsetting a model list to a certain delta value, we do not think this is something users will struggle with. This statement actually arose as a result of a discussion between one of the authors (XH) and Shane Richards, who advocated that in the case of applying the nesting rule the model weights should be thrown out as they were essentially meaningless. We have no quantitative proof of this, so added this statement merely as a caution to readers about how weights are interpreted. However, this sentence has caused a lot of problems, both here and in another manuscript, so we have removed it from this version.

Minor comments

L105 "best practice", in light of the title change, best to change this phrase here too?

>>>We think the reviewer is referring to the reference to the Zuur data exploration paper, which does actually deal with best practice issues. We weren't using it to refer to our own work, so we have kept the phrase here.

L369 "checking the assumptions of the LMM or GLMM is an essential step" – given the high level of detail and instruction provided elsewhere in the paper, I felt like this statement needs a sentence or two also providing suggestions for some of the methods/packages that could be used.

>>> We have added a parenthetical statement pointing the reader to the section 'Quantifying GLMM Fit and Performance' which makes a strong case for distinguishing between model fit and model adequacy, and points to some packages that can be used.

L641 note that it is also possible for data to be under-dispersed (less variance than expected by chance, i.e. more uniform than chance).
>>>We have added a sentence to discuss underdispersion and the spaMM package that can fit underdispersion models.

L719 isn't a Gaussian GLMM a LMM?
>>> We have changed this to 'Gaussian models'

L769 a dataset cannot be "made up of missing data" – suggest rephrasing
>>>We couldn't find the phrase the reviewer refers to but have rephrased some of the sentences in the missing data section where we suspect clarity can be improved.

L839-840 "there is no 'null'" – I see the authors' point, although one occasionally sees the intercept-only model referred to as the 'null model' – suggest rephrasing.
>>>We have changed to "because NHST requires a simpler (nested) model for comparison"

L1062 "data dredging" and "fishing" are no longer major components of the paper, and could therefore be removed from the Conclusions?
>>>We have removed this sentence in response to this comment, and also because it makes reference to 'false positives' which could be considered use of frequentist language in an IT framework.

The paper contains a lot of abbreviations – some of which are widely used, and others less so. I suggest the authors consider providing a glossary of abbreviations. Furthermore, many abbreviations that are used only once or twice are probably unnecessary, and the paper would be more readable if the terms were spelt out e.g., "GLS" (L633), "IT-AIC" (L775) "ASS" (L897 and 898).
>>> We have added long-form versions of these abbreviations where appropriate. Reference to 'ASS' has been removed in the updated version. We have not provided a glossary of abbreviations and instead have tried to use long-form wherever possible.

Literature cited in this review

Burnham, K.P., Anderson, D.R. (2002) Model selection and multimodel inference: a practical information-theoretic approach. Berlin, Springer.
Galipaud, M., Gillingham, M.A.F., Dechaume-Moncharmont, F.-X. (2017) A farewell to the sum of Akaike weights: The benefits of alternative metrics for variable importance estimations in model selection. Methods in Ecology and Evolution, 8, 1668-1678.
Gelman, A., Hill, J. (2007) Data analysis using regression and multileval/hierarchical models. Cambridge, Cambridge University Press.
Giam, X., Olden, J.D. (2016) Quantifying variable importance in a multimodel inference framework. Methods in Ecology and Evolution, 7, 388-397.
Richards, S.A. (2008) Dealing with overdispersed count data in applied ecology. Journal of Applied Ecology, 45, 218-227.

Reviewer 3 (Anonymous)

Basic reporting

all okay to good.

Experimental design

doesn't really apply with the exception of the simulation which is okay, but doesn't really much and conflates mode selection with testing (see my comments to the authors)

Validity of the findings

doesn't apply

Comments for the Author

General comments

This manuscript greatly improved as compared to the last version, and I appreciate the authors' efforts to deal with my comments and recommendations. However, I still see need for improvement and refinement. More specifically I see four major points (in my view issue two and four are the most crucial ones):
>>>Thank you for your careful attention to the manuscript, and especially for identifying several instances where the clarity of our terminology/phrasing could be improved. We are especially keen to avoid confusing readers by looking like we advocate mixing analysis paradigms, and appreciate the reviewer continuing to work with us to ensure our language is precise

* First, I still feel that the manuscript's two major topics, GLMM and MMI, are not really connected, neither regarding statistical theory nor in the manuscript itself, and that the manuscript would benefit from being split into two. Moreover, I even see dangers in presenting both approaches in the same manuscript. In fact, as the authors themselves (correctly) state, the number of estimated parameters (degrees of freedom) in GLMM are generally unknown (L 883-884). This, in turn poses the question of how to even determine an information criterion such as AIC or BIC which penalizes for model complexity. The authors do not elaborate on this question, beyond briefly mentioning the issue. In their cover letter the authors state that they want to present the entire 'analysis pipeline'; but to me it seems that this particular pipeline (i.e., GLMM in combination with MMI) does not rest on a solid theoretical basis (and the fact that the two methods are frequently combined in applied statistics in ecology and behaviour doesn't alleviate the problem). So maybe better not presenting as being logically following one another, perpetuating common practice but not promoting best practice.
>>>We agree that combining both topics into a single manuscript is no small task but differ in our opinion with the reviewer on whether the topics are unrelated.

Indeed, the Bolker et al 2009 TREE paper discusses both model specification and ways of selecting among those models, including how to calculate the appropriate degrees of freedom for models. We have based our structure on this paper, but updated it for the statistical tools and packages now available

nearly 10 years later. We certainly do not feel that it is dangerous to present both topics in a single manuscript (as Bolker et al. also did). In fact, we would argue the converse – that simply dealing with model specification but leaving (novice) readers without an adequate roadmap on how to begin to select among models would itself be dangerous. We do agree the issue(s) of model selection could easily form an entire paper on their own to deal with each issue in depth, but of course doing so was never the original goal of this manuscript. We feel that we have achieved our aim of providing a guide to the entire analysis pipeline and pointing the reader to the resources needed to truly understand each step of the pipeline in more detail.

We hope that the reviewer will be understanding of this difference of opinion, and we are very grateful for their continued help and advice on this manuscript

* Second, I still see need for refinement regarding the sections about random slopes (particularly L 379-402). Although the authors have considerably improved the manuscript regarding this point, the current manuscript still cuts the issue too short. The key point is that the authors seem to mix random slopes and correlations among random intercepts and slopes. In fact, the maximal model proposed by Barr et al. (2013. J Memory Lang, 68, 255–278) represents a model comprising random intercepts, all possible random slopes and also all correlations among them. The authors of the current manuscript are correct in stating that (nearly) perfect correlations are indicative of a model being too complex for the data at hand. However, the situation is not an all-or-nothing. In fact, when the correlations cannot be reasonably estimated, it might still be possible to fit a model including random intercepts and all random slopes but not the correlations among them (see Bates et al. 2015. Journal of Statistical Software, 67, 1-48). This is also the approach somewhat proposed by Barr et al. (2013) and Bates et al. (Parsimonious Mixed Models. http://arxiv.org/abs/1506.04967v1): begin with the most complex model but simplify it, beginning with the exclusion of correlations among random intercepts and slopes when it appears too complex.

There is one additional point I feel the authors must make: it should not be a matter of taste whether random slopes are considered. Instead, based on the available

literature Schielzeth & Forstmeier (2009. Behav Ecol, 20, 416-420), Barr et al. (2013), and Aarts et al. (2015. BMC Neurosci, 16, 94) it simply seems needed to account for random slopes to achieve the nominal type I error rate (and correspondingly unbiased standard errors and confidence intervals); and when all or several random slopes are neglected because the model would get too complex to fit it, then one has to face a perhaps highly elevated type I error risk. Also at other occasions (e.g., L 172-187), the authors still leave the impression that decisions regarding whether to include random slopes are a matter of 'taste' or the aims of the study, but in my view Schielzeth & Forstmeier (2009), Barr et al. (2013), and Aarts et al. (2015) made convincing cases that these have to be included to prevent type I errors.
>>>We have removed the sentence in the indicated section that suggests that the decision to fit random slopes can depend on the goals of the analysis. We have also added a sentence in the 'Random Slopes' section to acknowledge that removing the correlation between intercepts and slopes is an option. We think this section now gives appropriate advice on how to specify random effects structures, and where to find more detailed information on these topics.

* Third, I feel concern about the section headed 'Stepwise Selection, Likelihood Ratio Tests and P values' (716-750). In my few this conflates a couple of issues in an inappropriate and confusing way. First, NHST is one of the three major statistical 'philosophies' (together with Bayesian and information theory based inference) used to draw inference about a model and/or individual predictors. Stepwise procedures, in turn, are a specialized technique which had been (and unfortunately still is) used as a means of model simplification; stepwise can be based on NHST, but also on an information criterion like AIC or other criteria (e.g., an F value). As such NHST has nothing to do with stepwise procedures, and the far, far majority of uses of NHST have nothing to do with stepwise selection. As the authors correctly state, both have come under heavy criticism, but for very different reasons. NHST is criticised since decades for the many weaknesses and pitfalls the approach has; stepwise has been criticised for a couple of particular weaknesses (e.g., Whittingham et al. 2006. Journal of Animal Ecology 75, 1182–1189) among which is an extremely inflated type I error probability (Mundry & Nunn. 2009. Am. Nat., 173, 120-123).

>>> We agree with the reviewer that stepwise inference and NHST are separate entities. We made sure to state this clearly in the original manuscript, by first defining NHST and then defining that stepwise selection can use the NHST framework. We realise that the final paragraph of the section could be interpreted as attributing the flaws of stepwise procedures to the use of NHST, which is misleading for (especially novice) readers, and we thank the reviewer for pointing this out. We have edited this section to remove this issue.

Use stepwise AIC but is far less common and so we did not give it any treatment in the previous version, which was an oversight. We have now edited this section to separate NHST from stepwise procedures and acknowledge that some people use stepwise AIC.
Of course, discussing stepwise procedures in general is difficult because one must then avoid language like 'significant' predictors, as this phrase doesn't apply to stepwise AIC. So we have added a caveat that we focus on stepwise NHST procedures so that we can be consistent with language.

Here are a couple of clarifications which seem needed:
- NHST, when not coupled with stepwise and used appropriately in the sense of Forstmeier & Schielzeth (2011. Behav. Ecol. Sociobiol., 65, 47–55) does not lead to overestimated effect sizes or an inflated type I error rate.
- the null model sensu Forstmeier & Schielzeth (2011) is not one that differs from the full model by a single predictor, but by a set of predictors and the comparison between the two reveals a global test of their combined effect (appropriately accounting for multiple testing which otherwise would be an issue in case one would base inference on the significance of the individual terms in the model without conducting a full-null model comparison).
>>>We agree with the reviewer here, and have mentioned this in the 'Global Model Reporting' section. We do not think we gave the impression in the 'Stepwise selection' section that the F&S 2011 null model was a single-predictor-difference model. We have now moved the 'global model' section up to be included in the 'stepwise' section as they are complementary.

- Mundry (2011) did not argue in favour or model simplification.

>>>We have removed the phrase ‘model simplification’.

- Murthaugh (2009) found that stepwise and all subsets approaches revealed largely comparable results (wrt their predictive performance and number variables selected) and regardless of whether stepwise was based on an F-test or information criteria. As such the article does not really compare NHST with other philosophies but compares model selection techniques.
>>>Thank you for drawing our attention to this. We have removed this sentence.
“
* My last major concern is about the fact that the authors lack clarity and rigour regarding the issue of mixing model selection with significance testing. In fact, Burnham & Anderson (2002) have repeatedly pointed out that mixing the two philosophies is a no-go (and Mundry 2011 showed that it leads to drastically inflated type I error rates). For instance, when applied appropriately, there is no such thing as a type I error when using model selection based on an information criterion (such as AIC) because in the context of such an analysis one simply must not use any tests (see Burnham & Anderson (2002), e.g., P 202-203). In the view of the authors of the current manuscript, however, such an option seems to exist (see, e.g., L 805-807; 850-852; 855-859; caption of Fig. 4). Its worth emphasizing here that the term 'test', as I used it here, also encompasses checks of whether confidence intervals comprise the zero (and that bias in P-values gets along with parallel biases in standard errors (being too small) and effect sizes (being too large)).
>>> We thank the reviewer for identifying this inconsistency in our language. We have now changed the phrasing used here to refer to inclusion of uninformative parameters, rather than presence of Type I errors. We state again for emphasis that in no way were we intending to advocate mixing of analysis paradigms.

Apart from that, I am still not very convinced by the simulation (L 813-836), since it conflates a couple of issues, namely (i) combining model selection with significance testing (see also my previous comment), (ii) conducting pair-wise comparisons without a global test of the effect of a factor, and (iii) neglecting the full-null model model comparison). Each of these have been addressed more thoroughly and

clearly in other papers (i: Burnham & Anderson 2002; Mundry 2011; ii: all stats books covering ANOVA; iii: Forstmeier & Schielzeth 2011), and I feel this section does not contribute much to the existing literature nor to the clarity of the manuscript. On the other hand, the authors should make clearer statements about not combining model selection and significance testing (in a wider sense, i.e., encompassing also inspection of whether confidence intervals encompass zero; see also my previous comment).

>>> We agree with the reviewer that these simulations conflate several sources of error in a single outcome, and cannot partition the observed rates of uninformative parameter inclusion among them. As such we have decided to remove the simulations from the updated draft of the manuscript to avoid causing unnecessary confusion for the reader.

Specific comments

L 33: Pretty abrupt transition from one topic to another.

>>>We have edited the abstract to reflect all the edits made in this version, and so have changed this sentence.

L 67-68: inference 'about'.

>>>Changed

L 69-72: I felt these examples are confusing: if the fixed effect varies or is manipulated at the clutch level, how can pseudo-replication happen on the chick level (it happens on the clutch level if each comprises several chicks, each measured once). And 'if fixed effects vary at the level of the chick': between or within chick? In case of the former the authors are correct that non-independence occurs among clutches or mothers'; but in case of the latter it also happens on the chick level.

>>>We have clarified the wording of this section:

“In our example, if the fixed effect varies or is manipulated at the level of the clutch, then treating multiple chicks from a single clutch as independent would represent pseudoreplication, which can be controlled carefully by using random effects.”

L 145: 'hierarchically' implies that certain effects get priority when being estimated; but to my knowledge all effects are modelled/estimated simultaneously.
>>>We agree that all effects are estimated simultaneously. Our point was to emphasise the hierarchical structure of variance components in such models. We do not think that this section implies temporal ordering of component estimation, and in fact removed any misleading references to this in the first revision of the manuscript. We have not made any changes here.

L 165-167: in my view random slopes need to be included to keep type I error rate at the nominal level of 0.05 and obtain unbiased standard errors and confidence intervals.
>>>We have removed this sentence now to prevent readers thinking it is a choice. We discuss the random slopes paper a few lines below this and have updated the other section in the manuscript in line with the comments above.

L 179-182: it seems worth mentioning that this model comprises the random intercept, the random slope and also the correlation between the two.
>>>We have now added this as requested

L 183: Schielzeth & Forstmeier (2009) *show* that...
>>>Changed

L 274-279: I don't get the point here. Certainly, the random effects structure needs to be set up appropriately; but as Barr et al. (2013. J Memory Lang, 68, 255–278) showed, P-values for individual effects should best be based on likelihood ratio tests comparing the full with respective reduced models. This, in turn, doesn't rely on determination of residual degrees of freedom.
>>>If one were performing F tests for models fitted to a Gaussian trait, then determination of the residual degrees of freedom is central to the test to derive a p value. We have noted this particular misuse of statistical tests between full and reduced models when such pseudoreplication is present.

L 285-287: *general* linear models make the assumptions of normality and homogeneity of residuals (in my view, 'linear models' is a generic term

encompassing also Generalized Linear Models such a Poisson or logistic models).
>>>We respect that this may be the view of the reviewer, but we don't believe it is universally shared. However, we have changed the syntax as requested.

L 287: lower case 'normality'.
>>>Changed

L 308-318: my intuition tells me that those not being already familiar with the link function wouldn't get this section.
>>>We agree that link functions are a tricky topic for scientists to get to grips with initially. We deliberately point the reader to appropriate texts

L 372-376: but this applies in general (not only to GLMM).
>>>We have changed this to '(G)LMMs'

L 374: 'Crossed factors *allow to* accurately estimate...'.
>>>We have changed to "Crossed factors allow the model to accurately estimate…"

L 566: Zuur et al *give*.
>>>Changed. Thank you for pointing this out.

L 586: 'estimates *of* …'.
>>>Changed

L 658-659: to my knowledge, Wald-/t-tests are not applicable for random effects (and maybe not even F-tests).
>>>We have changed this section to simply say 'for tests of random effects'.

L 694: 'often' or 'always'.
>>>We have changed this sentence to "When collecting ecological data it is not possible to measure all of the predictors…"

L 779: 'reference*s*'.
>>>Changed

L 788: '...details on *how* AIC...'.
>>>Changed

L 792-795: comparing the full with a null model is not an alternative to NHST, but simply NHST applied appropriately (avoiding type I errors due to multiple testing)..
>>>We have changed this sentence to "Performing 'full model tests' (comparing the global model to an intercept only model) before investigating single-predictor effects controls the Type I error rate…"

L 805-807: but this applies only when one selects the best model and then tests it, something which should never be done as repeatedly and strongly stated by Burnham and Anderson (2002).
>>>We have changed the wording here to say that this approach increases the risk of including uninformative parameters, following the language used by Arnold (2010). We certainly weren't advocating testing the top model, but can see that the language we used may have made it seem that way.

L 817: 'ASS' not formally introduced.
>>>We have now removed this section pertaining to the simulations that mentioned ASS

L 818: I guess you the authors mean 'different' rather than 'separate'.
>>>This section has now been removed.

L 817-819: I don't see any difference: what the function dredge of the package MuMIn does is exactly dredging in the sense of fitting all possible subsets of models.
>>>This section has now been removed...

L 849-852: as Burnham and Anderson (2002) have pointed out repeatedly one should not use the term 'significant' in the context of an AIC-based analysis (e.g., P 84, 203), and I feel the authors should adhere to this rule.
>>>We apologise again for imprecise language. The authors all believe firmly

in not mixing analysis paradigms and do not wish to give the impression that we do. We have changed this phrase to 'uninformative parameters'

L 855-860: another example of the authors mixing model selection and significance testing. According to Burnham and Anderson (2002) such exercise must not be done (see previous comment), and Mundry (2011) clearly showed how such practice leads to drastically inflated type I error rate.
>>>As above, we apologise for imprecise language and have removed this sentence, especially as it pertains to the removed simulations

L 859: which 'conditions'?
>>>We have removed this sentence.

L 930: 'included' in what? The AIC-best model?
>>>We have changed this to "included in the top model set"

L 936: I'd use 'similar' rather than 'equal'.
>>>Changed

Figure 1, caption: reference to (B) is missing. The fact that the overall intercept is 0 is irrelevant here.
>>>We have now added the missing figure caption. Thank you for pointing this out. We have also removed the reference to the intercept being zero.

Figure 2, caption: I'm not sure if I would speak of a biased estimate when its average seems pretty much perfectly matching the simulated value (I'd speak of 'bias' then the average deviates from the simulated value).
>>>We have changed the phrasing to "With moderate collinearity, estimation of β_{x1} is precise, but certainty of the sign of β_{x2} is low. When collinearity is strong, estimation of β_{x1} is far less precise, with 14% of simulations estimating a negative coefficient for the effect of *x1*"