Reviewer comments are in italics. Our replies are in plain font. All line numbers refer to the marked-up file showing changes (`bayesianflows_diff.pdf`).

Reviewer 1 – comments:

The manuscript promotes the use of Bayesian methods, and describes a workflow for Bayesian analyses. Neither is especially novel, which is not to say there wouldn't be value in a more specialist journal publishing a workflow description, if it was well done. But I am afraid this manuscript is a long way from a good standard. At the moment I think it is a long way from being publishable, and even if it could be improved sufficiently I have difficulty seeing that it would be appropriate for Nature Ecology & Evolution: the central message is not novel or important enough.

My reason for doubting this could reach a suitable standard for NEE is that Bayesian methods have been around in ecology for a long time (The Ecological Detective was published in 1997, for example). Model building and comparison is also not novel, and different people will do it in different ways. In that sense, this is only one approach amongst several, and much could be debated (e.g. in the example given I wouldn't bother with simulating before I fit to the data, because I know LMMs well enough that looking at the data first is more helpful in spotting potential problems. This is not to say that doing this is wrong, it is one way of building the intuition about models to know when you don't need to simulate).

We appreciate the reviewer's view, but disagree that our paper is not novel. We agree that Bayesian approaches have been around for some time, but our paper is specific to new workflow methods that help bridge the gap that often separates statisticians and applied quantitative ecologists. This is one part of the novelty of our manuscript, but the other is the specific workflow steps we outline. Combining simple but powerful and organized steps centered around data simulation into a scientific and statistical workflow is considered new according to the broader literature, with *Nature Reviews* publishing a major paper on the novelty and importance of this type of workflow recently (van de Schoot et al., 2021, though the workflow we outline is distinct from this one), that has been recently followed by extensions to other fields (for example in epidimeology, veterinary and cognitive sciences, Grinsztajn et al., 2021; Schad et al., 2021; Mielke et al., 2023; Hess et al., 2024) but it is not well integrated into ecology. Indeed, *Philosophical Transactions of the Royal Society A* has just commissioned a special issue on the value of the type of workflow we outline across different fields.

We have worked to highlight these recent advances (see lines 48-51 and 58-59, for example) and how we build upon them in the current version of the manuscript, and we have also shortened and clarified the aims of our manuscript substantially, as we detail below (response to next comment).

I have a lot of critical comments about the current version. The bottom line is that it does not provide a knowledgable overview of Bayesian methods, and the suggestions for the workflow are often too vague to be convincing. There should really be an example in the main text, but the one there is doesn't do anything that can't be done using frequentist methods, and does not even show that the workflow leads to a better model

We appreciate the reviewer's concerns and believe they are driven by the length of the previous version and a need to clarify our aims. We have revised the manuscript to reduce the text by 30%, which we think better addresses our aim of reaching a large audience interested

in bridging the gap between statistical and scientific workflows with Bayesian methods, but also who may not be familiar with these methods (as it seems the reviewer is). We also have clarified our aim throughout. We now state: "Our aim is to provide an approachable rubric for those new to fitting complex models, and is not intended to be a comprehensive overview (see 'Next steps' in the Supplement)" on lines 62-65.

- the arguments for the Bayesian approach are little different to the arguments that were being put forward a couple of decades ago, and even he aspects that are more recent ignore a lot of modern Bayesian statistical ecology.

We apologize that the aim of manuscript was not previously clear and have worked to address this in the current version. Our aim was not so much to promote Bayesian statistics through new arguments but to show a workflow that bridges statistical and scientific workflows, and to make this workflow we outline as approachable as possible to non-expert statisticians. As detailed above, we have clarified this in the revised version, and cut a substantial amount of text specific to Bayesian methods (e.g. deleted lines 68-85, 244-257, 278-321).

- there are too many general statements with little to back them up: no discussion or relevant citations. This is especially a problem when the statements are difficult to understand. But even when they are understandable, one has to fill in a lot of gaps. For example, I thought the authors were thinking about bespoke ecological models (e.g. mark recapture, movement, population dynamics), but the example is a simple LMM that can be fitted with less hassle in the framework the authors are criticising. - the authors do not present classical statistics correctly (their version is the horrible one that has been traditionally taught to ecologists: for a better approach see Hector's The New Statistics with R, for example).

- the authors' knowledge of Bayesian methods also seems narrow, and does not stray far beyond Stan. They make no mention of methods that use numerical integration (e.g. in INLA and nimble), or (for example) say much about prior elicitation, despite the width of opinion and the practical advantages of different regularisation approaches.

Our goal was not to criticize other inference methods and we have worked to fix this in the revised version, deleting a number of lines from the introduction and other places that could have been mis-interpreted. As we state on lines 65-66, this workflow can be applied with other statistical inference methods, but we find it easiest to apply—and thus explain—in a Bayesian approach. We can see that we had more text focused on Bayesian approaches than was necessary and we have now deleted most of the text introducing Bayesian approaches (deleted lines 68-85, 257-273, 278-321 and others) and other areas throughout, including much of what we believe the reviewer is referring to regarding prior elicitation (deleted lines 120-125).

Our co-author, Michael Betancourt, is an expert in statistical inference methods, and has reviewed these concerns and confirmed our presentation is in line with the current best understanding of these methods from a statistical perspective. We can provide citations and additional explanations for specific concerns as needed.

In aiming to reach a broad audience, we have worked to de-emphasize particular languages or package (e.g., INLA etc.) and instead emphasize the approach. We have thus revised with this aim in mind (e.g., deleted text on lines 138-139).

- the actual example is a train-wreck. I have specific comments below, but I would expect to

see a model that could not be fitted with standard methods, and where the readers can see the workflow leading to a better model.

We should have clarified that we chose a simple example on purpose and we have now updated the text related to the workflow to clarify this. We argue that part of the barrier to taking up the approach we outline is that related Bayesian approaches and statistical workflows are often complex and dense, and this has provided a barrier to their uptake in ecology. We chose an example that is simple, but shows the power of the workflow we outline (and the analysis shown was published in *PNAS*, so we do not see it as 'overly' simple). Further, we have had the manuscript read by a number of quantitative ecologists who are new to simulating data and/or Bayesian analyses and all have commended how simple and approachable the paper is.

As the reviewer mentions in their line-by-line comments, this model cannot be fit using standard methods.

We believe our extensive edits to the manuscript (see `bayesianflows_diff.pdf`) and replies above have addressed the line-by-line comments, which we retain here for completeness. We have also updated the workflow example, as outlined in the response to reviewer 2.

L37-38: You cite a paper whose analysis stops in 2010. I suspect we'd see that Bayesian methods plateaued a few years ago: they are now solidly embedded in ecology.
L38-41: Yeah. This argument was being made 20 years ago about MCMC. HMC is harder to use than most of the algorithms that were in BUGS 30 years ago (fewer tuning parameters!). Albeit, HMC is faster, and converges better. I wonder, though, why there is no mention of numerical integration. Are the author not aware of it (e.g. in INLA and nimble)?
L60-61: Really? First, it's not clear what you mean by "robust". It certainly doesn't seem to be statistical robustness, which has a specific meaning. Also, do you have a citation for this (there should one something in response to Ioannidis' "Why Most Published Research Findings Are False" paper)?
L62: You have just undermined your previous sentence!
L63: What do you mean by "fragile"?
L68-71: How can you select models without comparing across models? This is weird, and the statement that they may not generalize to provide useful forecasts is difficult to understand. Are you thinking of Brewer et al. (2016: `https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12541`)? Anyway, you need to provide evidence for this.
L75-77: This argument has always seemed weird to me: the reason why it's true isn't anything intrinsic to Bayesian methods, I suspect it's mainly because hypothesis testing was so difficult to do.
L77-79: They can also be fitted without Bayes' theorem, by maximising the log-likelihood. This is generally easier than marginalising over it, because you don't have a normalising content to worry about. L80-81: I think the key thing about MCMC is that it is a simulation of the posterior, not that it is iterative. GLMs are fitted with iterative algorithms. And other numerical integration schemes are also iterative, but don't sample the posterior.
L88-92: This undermines the workflow: it's something should follow, except we don't have to, and we've ignored some critical steps. On the first point, you should be clear when some steps can be omitted. On the second, at least indicate where these steps occur in the workflow.
l99-100: This is a huge step, especially for people who are not so strong mathematically.
L112: So what if you don' expect your data to go to infinity or be non-zero? The poor reader doesn't get any help on what to do in this situation. And often these issues are not important: a lot of models assume the data can be negative to infinite, even if in reality this is impossible.

L114-120: This is another important step where the poor reader gets little help, other than some arm-waving. There isn't even a citation to any work on how to design priors!
L125: Good news - you don't need to write to out the full likelihood in Stan! I'm guessing what you actually mean is that you write the model out explicitly in the BUGS language, rather than using a simpler, less explicit, formulation.
L132-4: This is another statement that undermines the idea. You might get the right results, or it might be totally wrong. And if it's totally wrong, then something is wrong! So we get a vague statement with no evidence to back it up, and no follow-up to explain it, or what to do in this situation.
L134-5: This is a squishy statement. For a frequentist, it makes sense because the coverage (for example) should be correct. But what does it mean for a Bayesian? L141-2: This quote needs a citation (Wikipedia has a nice discussion of its history).
L143-6: This seems like an over-generalisation. I think most ecologists will be able to interpret a simple linear regression without simulating it.
L155: Which common concern?
L156-7: Nice that you give some citations, but this does feel a bit dismissive. You are meant to be advising how to do prior predictive modelling, so outsourcing all of the advice is selling the reader short. You don't need to say a huge amount, but some general comments to help the reader, with the citations as further reading, would help.
L161-2: What sort of diagnostics? I guess you are thinking of convergence diagnostics for MCMC, but this could also mean doing things like residual plots. L171-3: Wait, so you're not suggesting fitting simpler models first?
L177-8: Huh? How does this workflow give you this? I have absolutely no idea how it does this. In the example, the estimates from your Bayesian model are the asme as from the frequentist model, with exactly the same units.
L184: I wouldn't call R^2 a diagnostic per se, and it does not compare predictions to data. There are tests that fit this description, but diagnostics like residual plots are also useful (and easier to work with than full posteriors!).
L187-9: That is a big promise, but you need to do it sensibly. This is an important step, so I don't think it's enough to offload all of the advice onto the citations: the reader won't have any idea how to do it, and implication is that they have to read 3 more papers to get any idea.
L199-201: There is an irony in this statement: in your example lme4 fails, and you simply ignore the failure, other than to go on to do something else. L205-6: This may not be distinct, depending on your definition of "best"!
L206-7: How are the feedbacks focused on what is biologically reasonable? This isn't clear, and suggestions like the one on L164-6 and in Step 4 imply that statistical and computational reasonableness are what are important, with the expectation that this is correlated with biological utility (and yes, it often is, but you need to show this).
217-221: To what extent is it this workflow, and to what extent is it that you have a workflow?
L244-5: I don't think this is a useful explanation: the reader will have difficulty working out what you mean by "identified".
L246-7: This implies that models can be non-identifiable with finite data, which contradicts your definition. It might be better to be explicit about models being weakly unidentifiable. Incidentally, this book might be worth looking at: the author works in ecological statistics: `https://www.routledge.com/Parameter-Redundancy-and-Identifiability/Cole/p/book/9781498720878` *(and no, that is not me!)*
L248-9: What is degeneracy in this context? Without an explanation of what you are on about, this is useless.
L272-3: The usual advice in model building is to start simple (even if it is too simple!), and build up. I think perhaps a target model might be a better concept: the model you think might

be the model you use.
L282-3: Simple linear models can do this! Using the relevant units is trivial: it's just scaling. So I'm not sure why it is being emphasised.
L283-6: None of these are issues that can't be handled in classical statistics. Spatial and time series models have been around for a long time (Kriging, ARIMA, Kalman filters etc.), non-linear models have been around for a long time, and methods for non-Gaussian data have existed at least since Fisher, and the big advance there was published in 1972 (Nelder & Wedderburn). L287-8: But if you can fit a Bayesian model, I can't see why you can't fit a non-Bayesian version. The only practical difference is to decide what you are maximising rather than marginalising over. You can even use MCMC! I agree that in practice, a Bayesian formulation is often easier to work with, because it's a graphical model, and you don't have to worry about what to marginalise over. But. Suspect that's largely a quirk of history. L290-292: So why can't you make the same argument for a maximum likelihood approach? There you just need to "fit" the model by maximising the likelihood, rather than finding the full distribution. Even if you can't write down your likelihood, you can use indirect inference methods. L296-299: First, "low power" as a technical concept is purely one associated with hypothesis testing. Send, if you are using the term less formally, you can see low power in wide confidence intervals. Something which most standard tools will give you. L303-4: I think it might be better to say it's a part of the best practice workflow: there are several elements missing (e.g. actually looking at your data). L320-322: "no longer"! They have no longer been the purview of the few since the 1990s, thanks to BUGS. L328-9: This is dangerous advice. Because non-informative priors can be informative, and sometimes in unexpected ways, you can't dismiss the problem. Andrew Gelman's 2005 paper about precision priors was a bit of a shock to many Bayesians, because we hadn't realised the problems. There are also issues to do with regularisation where they can be really useful (e.g. PC priors). L330-2: If it wasn't for the mention of HMC I would assume this is a relic of an early draft from the early 1990s. The Metropolis algorithm doesn't need conjugacy, and that's hardly modern. Early versions of BUGS had several other algorithms that didn't require it either (e.g. ARS, slice sampling). Conjugacy was important in pre- MCMC days because that was the best way to derive a posterior, and in the early MCMC days it helped with coding some Gibbs samplers, but even then wasn't necessary. L336-8: Random effects is terminology that makes sense in the frequentist world, because they are marginalised over. Even in the early Bayesian MCMC days, they were called hierarchical models, or graphical models. But for the frequentist if you ban "random effects" you'll have to replace it with something else. L341-3: I don't see why training in retroactive checks is key here. Understanding the models in more detail is the issue, and I learned it mainly by drawing the DAGs (if you haven't done that in Win/Open/MultiBUGS, have a go: it's a lot of fun). L354-6: I agree that the replication crisis is lurking in ecology, but I'm not sure a different workflow is the answer. The issues of parameter interpretation are about understanding the model, and nothing to do with whether one is abusing frequentist or Bayesian methods. Of course, this can (and probably should) be embedded in a proper statistical workflow, but I think your suggestion could be mis-read.

Reviewer 2 (signed Sylvain Schmitt) – comments:

Personally, I really enjoyed the Wolkovich et al. 2024 perspective entitled "A four-step Bayesian workflow for improving ecological science". As I have been fortunate enough to adopt some of the same workflows in some of my previous research that led to the same observations, but less clearly investigated and stated, I really enjoyed reading their manuscript. I have a number of suggestions listed below, but they are minor and do not call into question the quality of the manuscript, which is scientifically correct. However, I am not sure that the

manuscript provides "new insights". Furthermore, I have not yet read such a well-written argument in favor of the approach, but I have the feeling that the approach is not new, as the reference to Betancourt 2020 highlights. Personally, I don't feel comfortable enough to judge whether the manuscript should be accepted or not, and will leave it to the editor to decide.

We thank the reviewer for their positive comments about the quality and topic of the manuscript. As we discuss in response to reviewer 1, we believe our paper is novel in providing a specific new workflow (this is not previously published) that helps bridge the gap that often separates statisticians and applied quantitative ecologists in an approachable format. We believe some of the previous text—and its length—detracted somewhat from this aim, and have revised to address this (discussed above in response to reviewer 1).

The first comment concerns the availability and reproducibility of the code. The evaluator's zip contained a txt file and not an Rmd file with a "rawlong.tot2.csv" file missing for reproducibility. I found the attempt at a reproducible example very interesting for the reader and even for teaching. But to be complete, it should be fully available and reproducible. I suggest authors use a permanent GitHub repository with a fixed DOI using Zenodo for this where they could even host a GitHub page to view the resulting html file. I also suggest authors use the renv package to ensure reproducibility without the problems associated with the R and package environment. I've taken the liberty of building a very quick example here: `https://github.com/sylvainschmitt/bayesian_workflow`*. I'm sorry if you find this a bit cavalier, I'll destroy it as soon as you notice me. But I thought an example was worth a thousand words. In addition, you could consider using lintr and or styler for the writing style to further improve usability (for example* `https://github.com/sylvainschmitt/sdmverse`*), but these are details.*

We completely agree with the reviewer's concerns that our workflow example be accessible and discussed this with him over email after receiving the review. While we appreciate the benefits of the renv package, it adds a number of miscellaneous files to any repo using it, making it hard for many to find the main files of interest and, without the addition of something similar to docker (or such) does not seem to make it easier to execute the code. In light of this we have moved the workflow example to a separate repo with an improved README, which we plan to publish on Zenodo. This repo is not given currently (due to journal restrictions, as the owner of the repo is obvious) but we can share it as requested. If reviewers or editors feel strongly that we should use renv then we could provide it via an additional repository.

The second comment is uncertain. I appreciated that the perspective was concise, so perhaps my suggestion is not appropriate. However, I sometimes wondered at each stage whether illustrating your assertions with an example in the main text using the supplementary material might help to make your assertions more concrete.

We understand this concern, but worry that this will make the manuscript less approachable and canalize how people perform each step—neither of which we want to do. Thus, we have elected to keep example code separate.

L.90 and in the example. I would have liked a bit of text and visualization on model diagnostic, besides I acknowledge you pointed to the literature and to the fact that stan was throwing warnings. But diagnostics are so important to me in Bayesian modeling that I would have liked a bit more.

We appreciate this concern, but have worked to shorten the manuscript and tried to keep it more agnostic in regards to coding langauges etc. (see also response to reviewer 1). Thus we have not expanded upon this point. We link to several references with extensive information on diagnostics.

L. 103. I very much agree on thinking of the model outside of the data. On the other hand sometimes data exploration can also help build intuition on the expectation. L. 121. I also think that simulating data can even support data sampling at an early stage. I think this is worth mentioning even if it's not fondamental. This is something that we've done for instance for `https://academic.oup.com/aob/article/131/5/801/7075765` .

In revising our manuscript, we have deleted the adjoining text mentioned here.

L. 157. The simulation step can also use a grid of parameters values for extensive exploration of the model behavior. This is unclear to me in the current text. However, it also seems to me that on the other hand we should reckon that the approach besides being very insightful can be time and power consuming and raise the question of where to stop.

In revising our manuscript, we have shortened this section such that we think expanding it to discuss where to stop may no longer be useful.

L. 160 "the model - you've now validated". I prefer evaluated to validated, and beware as the validation was done for a specific set of parameters values which can depart from the true data. However, considering the fourth step this is indeed not an issue. So maybe my comment is irrelevant and due to the linear reading.

Done (line 173).

L. 162. Again I understand the brief perspective, however I would have liked a bit more on diagnostics.

See our reply above.

L. 168 "ourselves included". Congrats for the honesty!

Thanks! We kept this line in the revised manuscript (line 181).

L. 244. Simple examples of non identifiability would have been nice for readers not used to the concept.

To shorten our manuscript and make it more approachable we have substantially shortened this section and removed the sub-header about non-identifiability.

L. 287. I agree with all the advantages of Bayesian models. However, I'm always uncomfortable when it comes to listing only the advantages. I think we should be honest about the disadvantages (I'm also speaking for myself as I've taught a bit of Bayesian modeling), and recognise that Bayesian models have a time, computing and technical cost that I think is worthwhile, but needs to be mentioned.

Agreed, we have substantially shortened this section (see deleted lines 279-321).

Table 1. I think the definition of "calibration" is very narrow to this specific work, this is okay, but should be stated I think. I've only listed my suggestions for improvement, not all the '+'s I've marked throughout the manuscript for what I really liked. So once again congrats, because it reads very well and I completely agree with your point of view.

We reviewed a number of definitions across the ecological literature and found them conflicting. Thus we kept our definition, and adjusted the caption to the table to clarify the many definitions present.

References

Grinsztajn, L., E. Semenova, C. C. Margossian, and J. Riou. 2021. Bayesian workflow for disease transmission modeling in stan. Statistics in Medicine 40:6209–6234.

Hess, A. J., S. Iglesias, L. Köchli, S. Marino, M. Mueller-Schrader, L. Rigoux, C. Mathys, O. K. Harrison, J. Heinzle, S. Frässle, et al. 2024. Bayesian workflow for generative modeling in computational psychiatry. bioRxiv pages 2024–02.

Mielke, F., C. Van Ginneken, and P. Aerts. 2023. A workflow for automatic, high precision livestock diagnostic screening of locomotor kinematics. Frontiers in Veterinary Science 10:1111140.

Schad, D. J., M. Betancourt, and S. Vasishth. 2021. Toward a principled Bayesian workflow in cognitive science. Psychological Methods 26:103–126.

van de Schoot, R., S. Depaoli, R. King, B. Kramer, K. Maertens, M. C. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, and C. Yau. 2021. Bayesian statistics and modelling. Nature Reviews Methods Primers 1.