Editor and reviewer comments are in *gray italics* (we provide below the full context of the reviewers 1-2 comments below, and extract the parts of reviewer 3' comments that required changes to this manuscript), while our responses are in regular text.

Response to editor's comments:

We are happy to hear that the three reviewers recognized the value of our work and that the manuscript is ready for publication. We have worked to fix the last small points raised by the reviewers, and we would like to thank them again for their careful reading. Furthermore, we also particularly appreciate the perspective on the policy implications from M. Authier, and will reach out to him to discuss further directly (as both he and we appear to agree that this is perhaps the topic for another paper).

Response to reviewers' comments:

Reviewer 1

The authors have addressed all of my concerns and I believe the MS is (nearly) ready for publication. I suggest one more pass for copy editing to fix small mistakes like inconistencies for example in "figure" "Figure" and "e.g." "e.g.,".

We thank the reviewer for their positive comments regarding our manuscript. We have worked to correct these inconsistencies.

Reviewer 2

All of my major comments were sufficiently addressed in this revision of the manuscript. I appreciate that the supplement is now referenced in the main document (I missed it on my last round of review). I have a few editorial comments for the main text and a few minor comments about the supplement, but I feel this manuscript is almost ready for publication.

We are glad that the reviewer found the manuscript has improved. Thank you for this last feedback.

Manuscript Minor comments:
Abstract: potential typo in second-to-last sentence. Should "where forecasting a natural output" be "where forecasting is a natural output"?

...

Line 39: should "ignore" be "ignores"

Thanks for catching this, it is fixed now.

Line 51: "hypotheses" or "a hypothesis" rather than "hypothesis"

Corrected.

Line 138: there appears to be an extra "in" in "in only in"

Corrected.

Line 215: comma after parameters

Added.

A couple typos in Box e.g., "calibrating submodes separately allows to avoid" missing "us"?

Changed to: "Calibrating submodels separately makes it easier to avoid the likely reality that [...]"

Supplement Minor comments:
Although the supplement works through an example far simpler than the types of research questions discussed in the main text, it serves as a reasonable (and accessible) example of how the workflow could be implemented. There is a brief discussion towards the end of the supplement about the "scalability" of the proposed workflow, but it leaves something to be desired. Have the authors applied this workflow (or pieces of it) to examples that are more complex that could be cited, or is that future work?

We disagree that if the model is not sufficiently complex it leaves 'something to be desired.' As we want this workflow to be approachable and leave room for readers to think through how they would advance upon it (to whatever degree of complexity they see as biological relevant), we have left it as is.

First green box: typo in RQ.."What is global trend..." should read "what is the global trend..."

Thank you, corrected.

Step 2: so assuming populations are exchangeable? Is this reasonable?

We have edited the text here to clarify this assumption:

> "In this first hierarchical model, we thus assume that populations are exchangeable—an assumption we might refine later."

Step 2: priors on sigma_alpha,sp and sigma_beta_sp do not seem appropriate, these parameters have positive support, so should these be half-normals (folded normals)?

We understand this confusion and now mention that sigmas are bounded to zero.

Stan model not included so could not be assessed (model1_nc2.stan), I recommend sharing for transparency.

Apologies that we forgot to share the Stan code. We have now added the code in the workflow example.

Figure 15, unclear what the posterior predictive distribution is of, this detail could be added for transparency. Could also add many posterior predictive datasets to the same plot and then overlay the observed data to facilitate easier comparisons between y-tilde and y-obs.

We now specify that the first posterior predictive distribution we show corresponds to the predicted counts across all species and populations, and show 9 draws.

Reviewer 3

First, I would like to congratulate the authors for a very clear and well written manuscript. The main thesis of the authors is to outline modelling workflows and illustrate in broad brush strikes how to bridge the current gap in ecological forecasting with simulations of synthetic data among the several steps of model building (the pre-data/post-model third corner in their Fig. 1) to improve confidence in modelling, inference and prediction. This is an important contribution, one that echoes many early calls (e.g. Thomas 1997): our community in ecology, despite its beefing up with respect to quantitative skills, remains fragmented and disjoint. I thus wholeheartedly endorse the main thesis of the authors and agree with it about research. There remains one concern that I would like to discuss with the authors, and possibly hear back from them although it is somewhat minor with respect to research but more problematic with respect to policy. It should not deflect from publication of the article in Philosophical Transactions of the Royal Society A.

We are glad that the reviewer found our manuscript well-written, and that he endorses our main messages. We are particularly grateful for his detailed comments, with much broader perspectives than just our manuscript. We believe that they deserve longer discussions (and could constitute in themselves another paper!), so we will reach out directly to M. Authier to share our detailed perspective on this. Below we address the comment that require changes to the current text.

The authors have addressed comments from the first round of reviewing. Figure 1 is central to their thesis: it is clear and well explained. However, I feel the lower left corner, corresponding to the pre-model/post-data area should be some darker grey to highlight that this corner is not overlooked but simply does not make a lot of sense in the current proposal (and in the idealized view of the current scientific process). It may sometimes correspond to mindless monitoring if one is uncharitable (Yoccoz et al. 2001, Legga & Nagy 2006). I believe it should not be greyed out the same colour as the other corners.

Thank you for this suggestion. We now use a darker gray to highlight this corner.

Moving to the upper right corner (post-model/pre-data), the authors could mention Simulation-Based Calibration (Modrák et al. 2025) to point readers to this useful tool for code checking, which is also central to workflows. I suspect this omission might be wilful not to overload readers. It might still be useful to point SBC on ecologists' map as new models keep being proposed but are insufficiently put to the test beyond one case study in many instances (see also Nichols et al. 2019).

We now mention simulation-based calibration (line 101).

With respect to terminology, the authors are mentioning "research spheres" in the abstract, which is a catchy term, and one I agree may be more appropriate than that of a research programme for what the authors are discussing. What is a research sphere is not articulated further by the authors and there must be a "policy sphere" at which the paper is alluding to. In fact, where I find the authors' thesis is less convincing is with respect to policy outreach. In policy, vague terminology such a "research sphere" is useful to get discussions started as loosely defined terms facilitate interactions. This can also lead to talking past each other but, usually, terminology gets to be clarified at some point when misunderstandings are too numerous to be ignored (and one get a common understanding document out of a round of meetings). My point is that the expectations of the "policy sphere" are usually not completely aligned with the "research sphere". And this realization bears on workflows.

...

I disagree with the framing on lines 7-9: the explanation provided from the two paths is not very convincing. The authors seem to have implicitly in mind some global policy, and hence justify a need for a global analysis. However, much of conservation is local, national or regional. Some surveys are designed and carried out at these scales, but when pooled in a global analysis, their analysis can become problematic because of sampling imbalances. However, the aim is now to make a global claim but that does not mean the local or regional trend was inaccurate in the first place. Yet the global analysis may be taken as evidence of biases (more on this term below) in policy fora. I overall think there is an unspoken assumption throughout in this opinion piece which is that all policies are somehow global. I see here a disconnect between high-profile papers on global trends which get attention precisely because they are global (and likely to get many citations and thus are a staple of high impact journals too), and a lot of policy which is much more local in scale.

We do agree that our framing is not directly adapted to the more local-scale decisions. We now mention that the policy needs we refered to are rather "global" (line 7).

My remaining points are more concise. - Terminology. If the authors are also aiming at the policy sphere, the word bias (e.g. lines 13, 134) should be wielded with extreme caution as it is easily weaponized. In policy circles, the term bias is, in my experience, the kiss of death as it is often interpreted as lopsided evidence. The authors are discussing preferential sampling with respect to space, time or taxonomy and its consequences in modelling results. Preferential sampling is more descriptive than bias, which also has a prescriptive meaning, esp. in policy. Also, a well- designed survey at some spatial scale may become an instance of preferential geographic sampling when collated with other surveys and data. This sort of data collation is often the case with global trend analyses that the authors seem to have in mind.

...

Also, taxonomic biases are bound to happen in global analyses. How do we redress those? That is, what can we say about data-poor species? How a richly-parametrized workflow calibrated on data collected on a few very well-studied species is going to help the majority of data-poor species? Taxonomic gaps have been identified without workflows and how to address those gaps point to monitoring and the need of appropriate governance to ensure long- term monitoring and inclusion of data-poor species. Research alone may not provide enough incentive or impetus for inclusion of data-poor species, especially if what is needed for policy about them is descriptive and routine science.

The workflow may allow to identify a missing phylogenetic structure that could be used to improve inference for rare species (with information from close taxa). Moreover, the workflow can also help identify the limits of our model—e.g. through data simulation, we could see that the model does not do a good job at predicting data-poor species. This could thus help to identify priorities for data collection.

- ML enthusiasm (lines 298-303): the point is granted for prediction but again evidencing a trend for policy is often required for further action. I fail to see how ML, if focused on prediction, is going to help. A trend is a useful fiction that is easily understood in policy fora where interpretability still has a huge premium. And a trend is very much interpretable. I think the authors may be going overboard in their claims if, again, their proposal is to have policy-relevance and not only research-relevance (a worthy goal of its own).

Sorry it was not clear. Our point here was not to share any enthusiasm for machine learning, but rather to highlight the need for an even more careful and coherent workflow to build interpretable models (i.e. avoiding "complexity trap" and build models with "clear mechanistic drivers"). We tried to make this point more explicit (line 301).

- Robustness is great. Policy loves the word but too rarely elaborates on it. Thus, on line 62, these methods should be robust against what more precisely?

We have made this point clearer (line 62):

> "This reality should drive researchers to use more coherent methods that remain robust for the intended inferential goals, even with imperfect data and knowledge gaps."

- typo on line 138: "These feedback loops are applied currently in only in Bayesian frameworks (to our knowledge)..."

Thank you for catching this! We fixed this typo.

Last, I would like to re-iterate that I view this manuscript as a very useful contribution for the research spheres in ecology. It can be published as is for this audience. I do not dispute any of the

claims made with respect to research, namely that disciplining model building to the rationalizing force of workflows is a good thing. I dispute what I think are shortcuts and could be perceived as a naïve outlook on the policy process. I do not assume that the authors subscribe to the linear model of science but a too quick reading of their opinion piece may suggest otherwise, especially the introduction when it refers to policy. For example, if lines 32- 41 are an accurate description of the research sphere, the policy sphere is then very rational in keeping an instrumental outlook to trend detection and favouring simple methods until the research sphere gets its workflows fully operational. The path from an improved workflow in research to better policy decisions remains to be charted: scientific workflows are likely a necessary step, yet one that may not be sufficient.

We would like to thank the reviewer once again for all these very interesting avenues for discussion—beyond the community of researchers (to whom this paper was primarily addressed).