Editor and reviewer comments (we provide below the full context of the two reviewers' comments) are in *italics*, 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 really interesting perspective on the policy implications from M. Authier, and...

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 possible to avoid the fact 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?

...

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 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.

Sorry we forgot to share the Stan code. We 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.

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.

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.

...

In policy, a trend in abundance is needed as evidence for decision and further actions. Whether that trend comes from a phenomenological or mechanistic approach is often of secondary interest to policy-makers: what matters is that the trend passes some checks (including statistical significance). Much has been written on the topic, especially with respect to statistical power. So "outlin[ing] the steps of a universal workflow" (line 48-49) is both novel and worthy. Yet I am not convinced that it will solve some of issues that are pregnant in the policy sphere. For example, timeliness is important in policy. As the authors are right to point, workflows require time to be developed. Sometimes a lot more time is also needed just to agree on the workflow, and then there is time and man-power needed for documenting, maintaining and updating the workflow. In short, one need a governance structure for the workflow. If the goal is to evidence a negative trend because this is the minimum requirement for taking corrective actions, then the authors' proposal, while clearly an improvement for the research sphere, will not impress the policy sphere as it is also likely to deliver too late; nor will it necessarily lead to better decisions. Awareness, design and publication of workflows will also have to happen in relevant policy circles, not just in scientific journals. That we should start with scientific journals in the right first steps, but reaching out to the policy sphere will not happen on its own either.

...

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).

To be clearer, I will provide an example (see Authier 2025). Using citizen science data on 724 species, Callaghan et al. (2021) estimated global population sizes for 9,700 extant bird species with modelling. These model-based estimates are not gauged against independent estimates for species well-monitored at the global scale. Owing to their large-size and conspicuousness several seabirds species can be censused accurately. Although imperfect, these estimates represent the best available knowledge for policy (ACAP 2021). Model-based abundances were ¿ 20 times higher/smaller than IUCN ones for 45/5 (out of 96 seabird species) and up to 100 times higher/smaller for 18/1 species (Figure 1). This is particularly concerning for endemic species, in which the local and global geographical scales are completely confounded, and which can comprise a single small population. A startling example is the Mascarene petrel (Pseudobulweria aterrima), a marine megafauna species endemic to Reunion Island in the Indian Ocean. Data on Mascarene petrels were used in the model training set and the estimated abundance by Callaghan et al. (2021) was several thousands of individuals. The estimated population size from local scientists is less than a few hundred individuals (Lopez et al. 2021). The global analysis is not helpful in my opinion, and may even hinder local policy which have taken a long time to gain attention and momentum.

This is a very interesting example. As highlighted in your next comment, maybe Callaghan et al. would have identified some flaws in their modeling approach if they had spent more time in the pre-data/post-model quadrat? If they predicted abundance estimations for 9000 new species outside their data, it seems like their inferential goal was above all prediction—and thus they should have focused on testing this goal in their workflow and should have developed some data simulation steps coherent with this inferential goal. It would be a very interesting example applied to conservation to illustrate how step-by-step workflows could help identify these local-to-global discrepancies, and maybe *in fine* reconcile local observations and actions with global trends and policies?

Callaghan et al. (2021) are concerned about forking paths and aware of reproducibility issues. They described a validation workflow which does not take stock of the pre-data/post-model corner but claimed nevertheless accuracy and success in redressing some biases (e.g. geographic and taxonomic). More generally, this example illustrates the following science- policy workflow: policy-makers read about a new global analysis published in a scientific journal. The analysis has a model workflow. Policy-makers look for the modelled-based abundance of their pet species and compare those to their knowledge and data. Local, national or regional scientists involved in policy are then contacted to explain any discrepancies. If there are indeed discrepancies, the scientists have now to explain their origins, looking into the model workflow and how it espoused good practices or not. With a growing familiarity of what are workflows, this should become more and more seamless in the long run. In the short run however, this science-policy workflow (i.e. the linear model of science, see below) is not scalable with respect to scientist's time in my opinion and it will not necessarily help policy if scientists (especially non-tenured ones) think that this is time spent not doing field work, writing grants and supervising students to further their research.

First, we would like to answer about the time it may take for this type of workflow to become

more common in research spheres. A large portion of scientific results today are investigated and written by non-tenured scientists. We have examples around us of graduate students who are likely more inclined than their supervisors to spend time developing carefully checked models and following workflow steps that give them more confidence in their results. While their supervisors may not push them to spend more time doing so, they will not prevent them from doing it either. Therefore, we believe that this paper—among many others—could actually help spread the use of robust modeling workflows (faster than it seems). Although the road between being a graduate student and obtaining a tenure-track position is long and uncertain, those graduate students are soon to become reviewers and submit grants (if they decide to pursue in academia), and interact with decision-makers, at least at a local scale (and even for those who will quit academia).

The above scenario (which can be applied to global analyses with or without workflows) is not to deflect from the authors' main point being made that "that an improved workflow that required data simulation and retrodictive checks would lead to larger model advances and a greater recognition of uncertainty—thus highlighting likely consistency in estimates across models—that could better aid policy". The "could better aid policy" part is simply too vague and could betray, in my opinion, that the proposed "universal framework" is implicitly embracing the linear model of science (Fernándes 2016) of 'truth speaks to power'. In other words, bad policy decisions are assumed to mainly result from a deficit of knowledge so that all that is needed would be better knowledge to naturally lead to good policy decision. I find this less and less convincing given that we've never had so much data or knowledge than now. We know what are the main drivers of biodiversity loss (e.g. Maxwell et al. 2016). We are literally awash with models (just think about the number of Species Distribution Models being published nowadays). Yet biodiversity, on average, does not fare any better.

We agree that the main obstacles—at least at the global scale—for effective climate and biodiversity policies are not a lack of knowledge. However, coming back to your previous point ("*Local, national or regional scientists involved in policy are then contacted to explain any discrepancies*"), we believe a different interpretation of the usefulness of the workflow for policy is possible. You noted above that scientists and decision-maskers can spend time discussing discrepancies with a high-profile global study—rather than on better decision-making or further research. We believe that a more systematic use of this kind of workflow would reduce such discrepancies (between different studies, estimates) and methodological controversies (a point we develop in the manuscript). So in this way, an improved workflow could aid policy.

While I agree with the authors that the workflow will help research spheres in disciplining modelling, I disagree with the policy implications. As highlighted in the box by the authors, "[m]odel development is the central step of the process-based workflow, typically requiring several years, yet it often remains opaque for anyone who has not worked the model itself." This is not a small issue.

We wanted to highlight that process-based model development can sometimes be very long (with or without this workflow), and thus that it would be better to follow coherent steps in model development and checking—given the amount of time process-based model development takes anyway. But to be clear, for most applications, applying the workflow do not require years.

For example, in the International Whaling Commission, established in 1946, many scientists are now retiring without new ones stepping in their shoes at the Scientific Committee (SC). The IWC SC pioneered some work on simulations (e.g. De la Mare 1986) and devised complex population models that were heavily tested to assess whether they would be fit-for-purposes. It is ironic that the simulation-savvy and model-based procedures encapsulated in the Revised Management Procedure that was developed in the late 1980s and early 1990s was never used in practice (for policy reasons,

not scientific ones, to make a long story short).

We are not familiar with this example. Even without considering policy reasons, we do feel however that science sometimes goes in circles—and that great improvements in computational tools since 90s have not necessarily led to more robust models.

The IWC example underlines the long-term commitment that is needed and the bigger question is how to square such a commitment in a competitive academic landscape where projects, when they are funded, are meant for a couple of years and the research proletariat is expected to publish regularly in prestigious journals and be highly mobile with respect to institutions. I contend that the incentives for the research and policy spheres are simply not aligned enough at the moment. How does one ensure the sort of long-term commitment needed in policy where stability of procedures is important (as any management procedures or indicators are going to have to jump many hoops to get agreed by all parties) and a steady stream of publications in an academic ecosystem where novelty and innovations are a premium?

...

The authors are alluding to the current publishing culture as a barrier but do not elaborate much. That is fair to omit in a short paper, and again I agree with their main point with respect to the research spheres.

...

I do think, however, that evidencing a trend serves a different goal in policy, and that having a workflow to bridge trend estimation and forecasting is not as urgent as the authors imply if the ultimate aim is to enable timely decisions to address biodiversity loss. In fact, the insistence on workflow may backfire simply because it can easily lead to either distraction or displacement in the sense of Rainer (2012). Distraction may happen because it could be easy to ask for more testing of some model features (a new twist to some mechanistic parts in the model) and ask that no decision be taken before that an updated workflow is completed. This could easily align with the common policy requirement of using the 'best available evidence', especially if some cutting-edge new results are brought forward at the policy table and one is being tasked with updating the current model. Displacement is more subtle but can also happen: managing the workflow becomes the policy and simply maintaining it absorb most of the available resources. As Rayner (2012) puts it "displacement occurs when an organization engages with an issue, but substitutes management of a representation of a problem (such as a computer model) for management of the represented object or activity." I think these two issues are real.

This is a very interesting question. However, we believe that the previous example of Callaghan et al. illustrates how bridging the differential goals of understanding and forecasting (or extrapolating to poorly observed species) can actually have important implications for policy. Similarly, it does not seem entirely accurate to think that this type of workflow should have been applied earlier (Revised Management Procedure of IWC) while also considering that it might be now 'too late' to highlight its importance (because there is no more time to lose to adress biodiversity loss?). We think this issue is linked to other challenges we encouter as scientists today: is there a risk in better quantifying and communicating the uncertainties of our results, if these uncertainties lead decision-makers to avoid making decisions? Or, by being the most transparent about our uncertain results (and spending more time applying consistent workflows), do we rather serve not only research but also the policy sphere?

I would love to actually read the authors' thoughts here and ideally be proven wrong that an improved workflow would be enough for enabling policy decisions whose beneficial impacts would be commensurate with the current magnitude of biodiversity loss7. In particular, I would be interested in a better articulation of how they see global studies feed-backing on policy which is usually more local, national or regional.

...

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 policy8. 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.

...

- 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).