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 both reviewers recognized the value of our work and we appreciate the opportunity to revise our manuscript. We have worked to clarify the points raised by the editor and the two reviewers in this new version.

Response to reviewers' comments:

Reviewer 1

This opinion manuscript (MS) argues that adapting a modern statistical workflow will help resolve some key issues in forecasting models used in ecology to inform policy about biodiversity loss and climate changes impacts. Specifically, the authors identity two specific "disconnected paradigms" for making such forecasts and argue that a workflow adopted from the Bayesian literature, and that emphasizes data simulation, would help unify these paradigms and make for a better scientific understanding of these important issues. I found the MS well-written, concise and easy to read, and well-organized with clearly laid out challenges and understandable solutions. I agree with the central proposal that the adoption of a "principled" workflow would be beneficial in many areas of ecology, and these types of studies in particular. I have no experience with these specific types of ecological models, but I do have experience with the Bayesian workflow underlying their key recommendation. So, I will focus my review on the recommendation and not the challenge. Below I outline several major concerns/questions, and below that minor comments and corrections.

We thank the reviewer for their positive comments regarding our manuscript. We have worked to clarify the points raised by the reviewer in this new version.

Inadequate justification of simulation.
Simulating data and fitting models is key to the step-by-step workflow but I think the MS lacks sufficient detail and motivation. The methods described come from the Bayesian literature, where simulating parameters from priors (prior predictive/pushforward) or posterior (posterior predictive or retrodictive) and then data, are natural. But a quick peek at the cited literature suggests that Bayesian models are not used. So how would one adapt them?

This is a good point, as we agree this could be applied more widely and don't mean to limit the approach to one framework. To address this, we have clarified that some terms are specific to Bayesian inference, but that data simulation still apply for all (line 98, line 106, line 138).

On L76 they say "... by fixing parameters to some reasonable range of values (which is straightforward if the parameters are interpretable)" but I disagree with this. Do you suggest analysts develop priors (using, e.g., Betancourt's cited workflow case study), and then simulate parameters from those, but ignore them in the subsequent fitting to empirical data? Or take a range for each parameter and simulate parameters across N equal steps for each one? Posterior predictive checks (retrodiction) are known to be conservative due to the double-use of the data (see Conn et al. 2018 [https://doi.org/10.1002/ecm.1314]). How does that affect your recommendations and how the analyst should use this tool?

Thank you for highlighting this as it does capture some of the complexities of how to apply this workflow. To address this, we have worked to clarify the data simulation step (line 92, line 97, line 98), and now highlight one should develop specific posterior predictive checks for the question of

interest (line 108). As we now state, we believe the best current approach is tailored to the research questions and aims, and thus exactly how conservative posterior predictive checks are may vary in different situations which the researcher should consider.

Furthermore, I think the authors should include a simple illustration of how these retrodiction checks can flag the need for model changes. For instance, I imagine they could take a fitted model from one of the key cited papers like the LPI trend, and simulate data given spatial/phylogenetic structure, but then refit a model without that and show how the model fails to fit the data and how the analysts would identify (visually, with a new figure) and resolve that by adding model structure/complexity. I am a big proponent of prior/posterior predictive checking and do not disagree with their recommended usage, rather I think more detail is needed for the MS. I think a judicious, simple example here could really help the reader interpret the recommendations and start thinking about how to apply it in their studies. Given the importance of this step in the overall thesis of the MS I believe more detail on how an analyst would do this would be worthwhile.

Thank you for this suggestion. We have added a figure in the supplementary material to illustrate a retrodictive check with a missing phylogenetic structure. We now mention this figure several times throughout the manuscript (lines 108, 132, 165, 184).

Resolving process-based model issues.
I do not see how the proposed workflow would resolve what appears to a key issue in the process-based research, namely that individual model components are done separately. Box (B) is supposed to address this but I am left unconvinced how this would resolve that. As I understand it, their workflow would lead to improved submodels, but would not change how those were linked together. Perhaps I missed a point here (admittedly I am unfamiliar with these models), but if not, I encourage the authors to try to make this point clearer. Otherwise will the workflow "bridge the gap" between the paradigms?

Our apologies that this was not clear. Applying the workflow to a process-based model would imply applying it to the model as a whole. This would make researchers calibrate the full model jointly, rather than treating submodels separately. We now clarify this in Box (C).

Causal analysis as a framework to develop research questions and elicit expertise.
The authors repeatedly highlight (rightfully so) that more time needs to be spent developing the research question. E.g., L266 "By focusing on model development more tightly tied to ecological expertise...". However, they give no advice on how to do that. I pose that the causal analysis literature can and should be considered as a good source to bolster this argument and guide readers, as it provides a robust scientific framework for this step of the workflow. For instance, Grace and Irvine (2020) [https://doi.org/10.1002/ecy.2962] is an accessible introduction to this topic with ecological examples. I suggest the authors either incorporate this literature into their recommendation, or justify why not. As Grace and Irvine argue, it is not really fair to ask readers to develop these hypotheses without giving advice on how to do just that.

Thank you for this reference. We have added it and now mention the importance of carefully developed hypotheses to drive mathematical model development in ecology (line 80).

Improvements to Figure 1:
I like this figure but improving the caption would really help the reader take in the material more efficiently. I suggest moving the upper right panel to the upper left, as it serves as a sort of legend of the three other panels and should be read first. I would label this one A as well. What is the significance of the figure-8 in B? How is that different than a circle? This caption does not define or discuss the panels A,B,C, and I think that it should. Why are the terms "pre-data" and "pre-model"

only used in the capture and nowhere else in the MS? "Retrodictive checks" may need to be further clarified in the caption.

We have reorganized the figure based on your suggestions, and we now explain more clearly the different panels in the caption. We also now used the pre- and post-data/model in the main text (lines 65, 65, 69, 74, 80, line 94, and 104).

"..where a model is built.."
"..some feedback to.."

Thank you for spotting these mistakes. We have corrected them and apologize for the errors.

Minor
Page 1, Line 40 in abstract: what is "unified" about the workflow?

We removed this word.

Page 1, Line 45 in abstract: drop "for"

Done, thank you.

L6: "date, alongside"

Done.

L36: These workflow citations are all unpublished (arXiv). I'm fine with that, but consider citing some from other fields that have gone through peer review. The caption of Figure 2 includes more, suggest putting them here too.

Done.

L103: What does "change the model structure" mean here? Adding constraints and priors? Removing or adding complexity?

Great point, we realized we could also better reference the example workflow we provide as supplemental files, which we now do (lines 85, 185). We also illustrate this point in the retrodictive check figure in the supplementary.

L110: delete "then"

Done.

Fig 2: Caption should mention this is expanded from Fig. 1C.

Done.

L116, L168: I disagree that "reduce uncertainties" is the goal – instead it is to more accurately quantify it. See L 257 "properly accounting for their uncertainty" for instance.

We have clarified this (line 143, line 198).

L143: FYI the fisheries literature has used the terms pre-data post-model etc. for decades now and have implemented this simulation approach. E.g., https://journal.iwc.int/index.php/jcrm/article/view/239/17. I don't think it's necessary to cite this but worth being aware of. You might want to re-cite Hilborn and Mangel (1997) here as they explicitly call for more simulation testing in ecology.

Done.

L201: "..workflow, non..." This sentence partially answers my question about L103 above

Great, as mentioned above we have tried to add more pointers to what we mean and how to apply this workflow through the revised manuscript.

Box A: Please check the logic of the first sentence. More papers because of less time confronting their models? Is that what you meant to say?

Apologies that this was not clear. We have clarified this first paragraph in the revised manuscript.

"For examples" should be "For example"?

Thanks for catching this. It is fixed now.

Box B: Please clarify the last sentence if possible. I'm not sure what "inappropriate way to accommodate their complexity" means here.

We have claried this sentence:

> "The way models are currently calibrated is likely not a coincidence, but rather a workaround to avoid confronting the full complexity of the model. Calibrating submodels separately allows to avoid the fact that, if the model were fitted as a whole, many parameters would compensate for one another—revealing structural degeneracies and making the model far more difficult to use."

L225: I like this paragraph and it makes a point that I hadn't seen written out so clearly.

Thank you!

L245: I feel like section 4.2 is slightly out of place. I'd rather see more text devoted to retrodictive checks than this. Consider dropping it and focusing on just a conclusion here

We have removed this last section and have expanded our conclusion.

Reviewer 2

The paper is well written and in a style that is accessible to a wide audience. The topic fits under the theme of "workflow for applied data analysis" by proposing a workflow that if used thoughtfully could lead to less divergence in forecasts and therefore more trust in the scientific process. The proposed workflow is similar to the Bayesian workflow outlined in Gelman et al. (2020) and is translated to two different approaches to ecological forecasting. I think the proposed workflow could be applied or modified to be used with a wide range of research questions within ecological forecasting if careful attention is paid to nuances and idiosyncrasies associated with each unique research question, and I think this necessary attention to question-specific nuances could be highlighted more in the paper. There are places where I feel more specificity (or clarity) with respect to these nuances could improve the manuscript and its potential for impact with respect to building trust in the scientific process. There is also a lack of specificity in how, exactly, this type of workflow could be implemented because there is no example application to data. I have outlined these areas under "Major comments" and provided minor editorial comments in "Minor Comments."

We are glad that the reviewer found our manuscript well-written. We thank them for the feedback, which has improved the manuscript.

Major comments

1. Clarity on inferential goals of "disconnected paradigms"

When the paper discusses "harmonizing both trend estimation and forecasting" what is meant by this? The title of the paper suggests the inferential goals from both approaches to be predictive in nature (i.e., "forecasting trends"), but it seems plausible that the simpler models are developed for more of an explanatory purpose (e.g., investigating relationships between covariates and mean response) rather than predicting or forecasting future trends, which would make the questions being asked by each modeling approach fundamentally different (see, Shmueli 2010). The paper could be

improved if more clarity was provided on the inferential goals of both types of modeling paradigms it discusses, and perhaps provide some discussion why they need to be (or if they should be) harmonized if they are in pursuit of different types of inference.

Thank you for this suggestion. We added a paragraph in the introduction to discuss the inferential goals of these two approaches (line 22).

2. Can this workflow be applied to non-Bayesian models
The emphasis on simulation-based investigation of model components is highlighted as a key feature of the workflow, so it seems a generative (or at least partially generative) model is assumed. Does this model need to be Bayesian? Explicit connections to posterior predictive checks and more discussion for how both prior- and posterior-predictive assessment could be used in a non-Bayesian context would strengthen this work. (Also note that prior predictive checks should be made explicit too – see minor comments).

We clarified that some terms are specific to Bayesian inference but that this workflow can be applied more broadly (line 98, line 106, line 138).

3. Discussion of Calibration:
Lines 98-104 seem to be referring to a check on estimating "the model" using synthetic data that are generated from the model (either through optimization or McMC), but there is no mention of various computational approaches that may be taken to fit "the model." How does the chosen computational strategy fit into the workflow if at all?

This is a good point, and we could have been clearer on different frameworks that the workflow can be used with. We have worked to clarify that in several places in this revision (in particular, see Box).

Further, is the proposed workflow to conduct a full-blown simulation calibration study (i.e., simulate-fit-check for capture) many times and across all possible parameter values? Assuming yes, I'd be interested in how something like that would scale for the types of models that are the focus of the paper? Could the authors provide more guidance on how to conduct such a study in this context?

We have clarified the data simulation step. We believe a first step would be to simulate from one parameter set (line 92), and then to repeat the simulations with several parameter sets (line 97, line 98).

4. Gaps in the workflow
A big argument in the paper is that the proposed workflow could lessen discrepancies in results between studies of trends (lines 123-128). However, in the workflow there is no mention of transparent communication of study design, which if ignored in analysis, could result in discrepancies in results for studies with similar research questions. Differences in sampling designs must also be considered prior to integrating different data sources (if integration is even possible). The importance of study design/sampling design and its ties to appropriate modeling is not discussed. Is it implicitly assumed that there is alignment between design, question, and model in this workflow? Some mention of the importance of study design would strengthen the work.

This is a great point. To address it we have added an entire paragraph that discusses the methodological aspects, including study/sampling design, that should be more transparently shared along with the use of this workflow (line 284).

Additionally, the choice of prior and posterior predictive checks (PPC) is non-trivial and depends on the context of the study. There is discussion of these types of simulation-based investigations, but no mention of the importance of developing checks that assess key features of the observed data

that are tied to the research question. Acknowledging the complexity of this step, and including some ideas for how to do this in the context of the types of problems discussed in the paper could strengthen the paper.

Thank you for this suggestion. We now say that the checks should be developed specifically for the research question (line 108), and now show a simple example of retrodictive check in supplementary material.

5. Theoretical examples provided, but no data application of the proposed workflow
Sections 3.1 and 3.2 and the Box outline theoretical examples of how the workflow could be applied for trend estimation and forecasting in the context of types of data, but there are no concrete data examples. A comment on how the proposed workflow could be applied to real data and how it would scale with complexity of data and research questions would be a good addition to the paper if it is not feasible to include real-data examples with the general discussion.

Good point. We provide a worked example, but we realized upon receiving this feedback, that we could better reference the example workflow here and in other places, which we now do (lines 85, 185).

6. Adopting the workflow
Recommendations are made to train ecologists more quantitatively, but there is no suggestion or encouragement of collaboration with statisticians. These types of workflows are generally part of statistical practice, so could an alternative to more quantitative training (which would have to come at the cost of some other aspect of training) be more cross-disciplinary collaboration with statisticians?

Good point. We have adjusted our language in one place (line 263) and we have added this suggestion also (line 281).

Minor comments
Abstract, line 45: remove "for" in, "could transform for ecological modeling."

Thanks for catching this, our apologies for this typo.

Line 52: change "in term" to "in terms"

Done.

Figure 1 caption, line 34: change "where a model is build" to "where a model is built"

Done.

Lines 75-81, the paper discusses a form of prior predictive checking without explicitly mentioning the connection. The connection should be mentioned and perhaps

Good point. We now mention that this step can help check prior assumptions (line 98).

Line 80: typo in "workflow" in "(see example worklow)"

Thanks for catching this; it is fixed now.

Line 86: change "also sometimes called a posterior predictive check" to "in a Bayesian workflow this is called a posterior predictive check."

We have edited the text here to clarify.

Line 87: could be more specific about "model output" as typically a PPC investigates how the entire posterior predictive distribution for the response or some summary statistic that is a function of the response and parameters compares to the observed data (or observed summary stat computed from

observed data).

We now mention that the summary statistics should be developed specifically for the research question (line 108).

Line 117: should "hypothesis" be "hypotheses"?

Yes, and this is now corrected.

Lines 198-201: if the goal is accurate and relatively-precise forecasts, do all of the nuisance parameters need to be identifiable for a model to be useful? That is, does identifiability of all parameters matter for process-based models with a goal of forecasting?

We believe that the strength of process-based models (as compared to machine learning for example, line 299) is their interpretability. Their parameters should represent biological processes, and their values are of interest. The inferential goal of process-based models is both predictive and explanatory (line 22), and poor identifiability could reduce this (line 213).