Editor and reviewer comments (we provide below the full context of each review) are in *italics*, while our responses are in regular text.

Editor's comments:

At this stage, I would like to ask you to make minor revisions to the work in order to align your paper with the remaining requests raised by R2, and ensure that the paper fits in with our formatting guidelines. We are unlikely to return the paper to reviewers at this point but will assess editorially whether the work is ready to move forward. Please do address the remaining reviewer concerns in a 'response to referees' letter, with changes made tracked within the text.

Add RESPONSE.

Reviewer comments:

Reviewer #1:

The manuscript 'Phylogenetic estimates of species-level phenology improve ecological forecasting' although already novel and of high-impact in the phenological forecasting ecology field; has now been significantly improved through the review process and the written and analytical elements are now of an exceptionally high standard that have broad and important implications in this scientific field.

I appreciate the time the authors took to address my comments and the comments of the other reviewers and use this feedback to improve the manuscript. I believe the manuscript now addresses my major concerns by including discussion of how their method generally applies across the globe and also presenting the raw data and raw methods used to model the phenological forecasts. Likewise I appreciate the authors prioritisation and assurance of making their code and data reproducible.

Likewise, the authors have taken on board my minor edit suggestions to improve the text and the figures in the manuscript.

We thank the reviewer for their positive feedback as we have done all in our hands to address raised concerns.

Reviewer #3:
Lizzie/JD to edit this whole section...

Thanks for the additional analyses that the authors have conducted to support their argument that phylogenetic mixed models performed better than traditional hierarchical mixed models. These additional analyses improved the manuscript though I still have some major comments.

We agree that the new analyses showing how and when the phylogenetic mixed model outperforms traditional mixed models are both reassuring and have contributed making the ms. stronger.

First, what is the reason to leave one genus out instead of the more traditional way of randomly selecting say 20% of species to be the testing data?

The leave-one-genus-out is one (out of many) posible approach to cross-validate our results. Leave-one-out methods are a standard tool for statistical cross validation (see e.g.,) presenting the advantage of being an unbiased estimator of true performance. This is because every data observation (i.e., genus in our case) is used for validation once, ensuring that all information contributes to the evaluation. A k-fold validation as suggested by the reviewer would have a major limitation in our case given that it only uses part of the data for validation (e.g., 20%), which can lead to biased results. To avoid bias, k-fold validation could be repeated a high number of times (e.g., 100, 1000), which would consume an excessive amount of time for our dataset and Bayesian modelling approach.

By leaving-one-clade-out at a time, we find a compromise between leaving out a large-enough number of species (when large genera are held out they can represent nearly 10% of the dataset) and, accounting for existing phylogenetic structure in the dataset over the validation process.

Second, it is cool to see that the predicted values based on the phylogenetic models are more correlated with the observed values than those based on traditional mixed models.

Thanks, we agree that the validation results showing how PMM estimates do a better job at inferring phylogenetically structured held out values reinforce the main message of the paper.

What are the RMSEs look like for both models? I think RMSEs are more typical when one tries to evaluate predictions.

We agree that RMSE is a standard metric to evaluate model predictions and have now calculated RMSEs and added results to Fig. S9. The prediction error is again lower in PMM (RMSEpmm = 21.6) than in HMM (RMSEhmm = 22.4) confirming previous results. We still keep correlations results for two reasons. First it is a straight-forward, easy to interpret measure. Second, and more importantly, because visual inspection of Fig. S9 makes it evident that beyond accuracy metrics, HMM constrain predicted values to be below 50, while PMM allows for higher estimates, a pattern better captured by correlation.

Also, given the marginally improved r values here, when should one take the effort and time to fit phylogenetic mixed models? Was the slightly improved model prediction worth the effort?

We would argue that increasing r by 0.12 (or 50% from 0.231 to 0.353) is beyond marginal. We explained in depth when it may be appropriate to fit PMMs both in the ms. (see lines XXX) and in our previous response letter to reviewers. While we believe that allowing for greater interspecific variation is inherently preferable, our model fitting allows the phylogeny to be scaled such that predictions will simply converge on a non-phylogenetic model if phylogeny is not important.

The reviewer raises again the question of whether the extra effort of fitting the phylogenetic model is worthwhile. Given that our analyses show that estimates can change with our method, then it should be worth for any researcher seeking the best estimates. This is particularly true

for ecological forecasting contexts, where it is easy to see how a geographically structured shifts of up to 8 days in phenological estimates for species underrepresented in a dataset (see Fig. S8), can have profound ecological implications. In addition, it is possible that we might observe larger effects in other datasets, but determining so would imply to first comparing the models. We now emphasize more (see lines XX-XX) what is to be gained from making the extra effort of fitting a Bayesian PMM such as ours.

Third, I am not completely sure about the main purpose of this study. Is it an empirical study to investigate different drivers of budburst and predict their future impacts or a method paper to proposal a new model to analyze similar data? If it is an empirical study, then the authors should highlight and discuss the results more. If it is a method paper, how do you facilitate others to use your method? The current manuscript seems to try to do both, which may not be the best choice here.

We would agree with the reviewer that our paper does both: we present a novel statistical approach applied to an empirical dataset that shows some interesting patterns which we interpret. We have adjusted the abstract to make this more clear on lines line **??** and line **??**. We are not exactly sure what additional aspects of our empirical results the reviewer would like us to discuss more but would happily do so if any guidance was provided.

Regarding how we facilitate to use our method, we provide all code and data hoping that interested researchers will start using Bayesian PMMs (see e.g., Appendix XX). We ignore why the referee states that doing both would not be the best choice as such opinion is not justified. We disagree on the basis of several papers published in NCLIM where methodological developments accompany empirical study cases (e.g., Ettinger et al. 2020; XXX). However, we now sign-post the different contributions of the paper more clearly (see lines XX-XX, XX-XX, XX-XX).

Minor comments:
Can the authors provide a file with track changes? It is so hard to see what have been changed without it.

We identified all changes to the ms. providing line numbers in our response to reviewers. We realize that a file with track changes would be useful and include it in our current revised manuscript.

Fig. S9 the panel labels are different from the caption.

Thanks for spotting this mistake! We have now corrected it.