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.

We respond to each comment by reviewer #3 below (we also include the comments from reviewer #1, though they did not request any changes). In particular, we now include the requested RMSE as an additional metric for prediction accuracy in our cross-validation analysis (with results supporting our previously submitted conclusions) and we clarify how our approach is preferable to others, emphasizing the benefits of fitting a Bayesian phylogenetic mixed model (PMM). Also, we have made a minor requested edit to Fig. S9. To review our edits, we provide a version of the manuscript with track-changes as requested, and indicate our changes with line numbers below. We thank the reviewers for their feedback, which has further improved the manuscript.

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, which we believed improved the manuscript.

Reviewer #3:

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 made the manuscript stronger. We thank the reviewer for their previous comments, which lead us to develop these additional analyses.

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?

This is a great question and we realize now that we did not well explain our rationale for this approach. Leave-one-out methods are a standard tool for statistical cross validation, presenting the advantage of being an unbiased estimator of true performance (CITES). This is because every data observation is used for validation once, ensuring that all information contributes to the evaluation. These methods are designed to replicate different potential datasets that could have been collected and compare outcomes, with the k-fold methods the reviewer mentions specifically assuming each data point is equally likely to be randomly sampled. This assumption is, however, not true in many datasets and we expect especially not true in ecological trait datasets. Our method is thus an extension of k-fold cross validation that incorporates this important non-independence. Further, by leaving-one-clade-out at a time, we can leave out a relatively large number of species (when large genera are held out they can represent nearly 10% of the dataset) while also accounting for existing phylogenetic structure in the dataset over the validation process.

To explain this we have added a paragraph to the part of the supplement where we explain the method:

> Traditional cross-validation methods generally leave out a randomly selected proportion of the data (e.g., 20% in 5-fold cross validation) assuming sampling is completely random (that is, that researchers gathering data would be equally likely to select each observation) and thus ignores potentially important structure in the data that may covary with sampling. In contrast, our method leverages phylogenetic structure, dropping out data in a way that may more accurately reflect differences across differently sampled data. Because of the presence of phylogenetic structure in ecological trait data and the tendency for such datasets to sample across taxa unevenly, we often omit sets of taxa (e.g., clades) sharing similar trait values across different natural sampling regimes. Our results show that the outcomes of this structured omission varies for each model. In HMM, excluding a well-sampled clade would tend to skew parameter estimates more (because it is likely to shift the grand mean towards which all species pool) than in PMM, where the partial pooling co-varies with phylogeny and thus species-level estimates are more stable.

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 were excited to show this also and again thank the reviewer, whose comments lead us to these new analyses. The validation results show that PMM estimates are more stable and provide more robust inferences, reinforcing 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 show these RMSE now alongside the correlation results and related figures.

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 feel the insights and improved forecasting from our novel phylogenetic model were more than worth the effort for a suite of reasons. First, we believe most researchers will want to fit the most accurate and robust model: we show estimates can change dramatically with our method and that it improves model fit, thus we expect many researchers will find it worth fitting. 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. Second, given how low R^2 values often are in ecology (CITES), we believe that the improvement we show of 0.12—or 50% (from 0.231 to 0.353)—is quite high. Improvements beyond 50% may be possible in other datasets (but without using this model, researchers will never know). Third, and perhaps more importantly, the model reshapes the current understanding of phenological cues, including the following major advances:

- The major debate on the prevalence and strength of photoperiod cues is likely highly skewed by its focus on *Fagus sylvatica*, which we show is nearly five times more sensitive to photoperiod than most other measured tree species. Our results thus caution against using it to draw inferences of photoperiod responses more widely (line **??**-line **??**).
- Addressing how correlated cue responses are we find that species sensitivity to one cue does not constrain sensitivity to another cue, suggesting selection can operate independently on responses to different cues (line **??**-line **??**.
- It also provides a suite of insights in to the evolutionary history of these cues, reviewed in the section *Phylogenetic structure of phenological cues*, which required phylogenetically informed estimates.

We outline this for systems beyond plant phenology on line **??**-line **??**:

> While we focused on spring phenology here, our new approach suggests a path forward for more general forecasting of species-level climate change responses. Our results show how including the phylogenetic relationship of species in a mechanistic model of underlying cues can overcome major limitations of most current hierarchical models—correcting biased model estimates, estimating the full variability across species and reducing uncertainty around individual species estimates—while at once providing insight into the evolutionary history of biological responses. Using this approach improved forecasts of phenological responses to climate change and could help anticipate impacts on critical ecosystem services from species-level shifts and thus aid mitigation and human adaption to warming.

Finally, we expect this method will follow other improved models that incorporate evolutionary history. Now 'classic' phylogenetic comparative methods (which correct residuals though some additional effort by researchers to fit a slightly more complicated model) were introduced into cross-species analysis over 20 years ago and are today de rigueur for most multi-species analyses, because they provide improved, less-biased and thus more robust estimates. We expect the same path for our new approach here and do not see the minor additional efforts as any reason to avoid these advances.

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 important empirical question (and related dataset) and find species-level variation that reshapes how we understand phenological cues, then show how this impacts forecasting. We mention this in detail (line **??**-line **??**) but have now made it more clear through small changes to the abstract (line **??** and line **??**) discussion (e.g., line **??**).

To facilitate use of our method, we include several versions of our code as well as an introduction to the method (see e.g., Appendix XX). We first review a simple version of the model, as a place to start for other analyses, and then we review the full model we used in the paper (alongside all empirical data). We have added careful annotation to all code in addition to text introductions to the code. We believe and hope that this will allow interested researchers to implement their own Bayesian PMMs.

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

Yes, gladly. We previously identified all changes to the manuscript through providing line numbers and some quotes in our response to reviewers. We now further include a file with track changes in our current revised manuscript.

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

Many thanks for spotting this. We have corrected it.