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:

The referees' reports seem to be quite clear. Naturally, we will need you to address all of the points raised. We draw your attention specifically to the need to provide more methodological details, and justifications of the choices made in the analyses, as well as the need to compare the modelling done to 'true' parameters.

We review below, point-by-point, all the changes made in responses to the reviewers' comments. In particular, we have provided extensive additional methodological details, and rephrased details that were present before, but not clear enough. We also now provide additional tests to compare the phylogenetic and non-phylogenetic models using a leave-one-out approach (see new section, "Leave-One-Clade-Out model cross validation," in Supporting Information), provide more details on data sources and their distribution (see new section "Details on data: sources and structure" in Supporting Information, including Table S1, and Figures S1-S2), and re-analyze data using a different metric of chilling (see Tables S4-S5). We thank the reviewers for their constructive feedback, and believe the manuscript is much improved.

Please note that the reviewers have also highlighted that sharing of the code and data will be an important part in ensuring that this paper has broad impact and use for the community; therefore please do let me know if you think at this point that you will not share the data/code on publication of the work.

Our labs are committed to open and reproducible science and share all possible data and code. The full dataset is already available on the Knowledge Network for Biocomplexity (KNB, a federated data repository with high metadata standards), as well as the specific subsets requested by reviewers. Further, we have posted a GitHub repo (link below in response to Reviewer 1) with annotated code for both Stan models, and analyses in R of models outcomes.

Reviewer comments:

Reviewer #1:

The manuscript 'Phylogenetic estimates of species-level phenology improve ecological forecasting' presents a novel methodology to model and estimate species-level phenological responses to temperature and day length. The authors have filled a large knowledge and ecological forecasting gap that has been previously overlooked in other research by using Bayesian hierarchical phylogenetic modelling – allowing the evolutionary history of the species studied to be a factor that can shape species responses to temperature and day length, rather than just using it as a corrective factor in modelling. The output of this new method will allow better ecological forecasting by considering a range of cues within a species, and the evolutionary history of the species which then allows

for better predictions especially where data are sparse for species' phenologies.

Interestingly, the authors found limited responses across all species to phylogeny compared to temperature in phenological shifts. They also found that average shifts do not accurately explain species' level responses. Finally, in applying their model, the authors found that there was an increase in variability across species phenological responses when a phylogenetic structure was employed compared to no phylogeny and this then led to decreases in estimate uncertainty for individual species' responses to temperature. This finding has broad implications for ecological forecasting.

We thank the reviewer for their positive comments and are excited they see the broad value in our approach for ecological forecasting.

Major comments
I have a concern about the generality of the data and therefore the generality of the conclusions drawn from the data. It is not clear in the methods section where the phenological data are located but of course temperate Europe and North America would be highly overrepresented in these data. It would be good to see the countries, regions or locations of the 44 studies from 33 papers which this papers data originate from, that are a subset of the OSPREE data. This location data should be made available at least in the supplementary material, but ideally in the materials and methods. Only when going further into the methods and cross-referencing Ettinger et al. 2020 is it evident where the broader OSPREE sites/papers are located and it is evident all studies are Europe/North America except ∼4 in South America/Africa.

The reviewer makes a fair point, we should have included more graphical descriptions of our dataset. We now include both an appendix showing the location of studies in the dataset (see Supporting Fig. S1), and a summary of the data that also accommodates a request by Reviewer 3 (see Supporting Table S1). We agree with the reviewer about the value of Southern hemisphere data: the data we present here mirrors the existing bias in the literature. We have incorporated all studies meeting our standardized search criteria (refer to Methods, lines 283-289), resulting in a limited yet valuable dataset that accurately represents species from temperate biomes in the Northern Hemisphere. While our dataset is biased to temperate biomes, it includes a number of studies (and species) from Mediterranean regions.

> In our dataset most studies come from Europe (n=37) and a few from North America (n=7). The same bias towards Europe is found across the full OSPREE dataset with less North American (n=19) than European (n=60) studies and only 3 studies located in the Southern Hemisphere. Given our need of daily gridded data for chilling we only include studies from Europe and North America, with most of these sites in temperate areas and a few in European Mediterranean areas (see Map in Supporting Fig. S1).

We recognize that there are limitations to the available data and this was, in part, the motivation for developing the model we present here. Our approach allows us to better quantify uncertainty and leverage information on plant evolutionary relationships (which covary with biogeography) to generate more robust predictions.

Interestingly, phenological studies in the southern hemisphere are showing differing responses to climate than northern hemisphere studies (Chambers et al. 2013, Everingham et al 2021). And some studies have shown precipitation in the season prior to the phenological event (e.g. flowering; Everingham et al. 2021) is more correlated with shifts in flowering time than temperature especially in regions where inter-annual temperature varies much less and so the 'chilling' period prior to flowering is not as pronounced as the northern hemisphere and may not be a driving factor of within-species cues. Likewise, 'leaf-out' is a less important or not-existent phenological event in regions outside of north America, Europe and Eastern Asia as many species outside of these regions are evergreen. You suggest all species in your data "respond to all three primary cues – forcing, chilling, and photoperiod" however, how general is this response across other species outside of your study regions?

We completely agree with the reviewer about the need for more data beyond the temperate zone. We used standardized meta-analytic methods to review the literature, but the outcome is data highly skewed towards angiosperms in Europe and North American temperate zones. Our analysed data does include some Mediterranean species (n=3) and we did not exclude any species based on whether they were evergreen or deciduous, though cutting experiments are strongly biased towards deciduous species. Indeed our full dataset includes a number of evergreen conifers, but they were too under sampled as species to be included (see lines 281-282). We reviewed the references provided by the reviewer, but they are all observational and so cannot be included in our approach, which relies on experimentally manipulated cues. We seem thus to be limited by data and unfortunately, we could only speculate about how transferable the responses we found are to species from very different biomes. To answer the reviewer, until more experimental data is published we cannot know how general the relative importance of forcing, chilling and photoperiod would be for species in other regions. Yet, it is safe to assume that, one of our major findings—i.e., the large cross-species variability—would only be reinforced if additional species from other regions strongly differed in their relative cue importance.

We are now more explicit in the methods about the limitations of current data on phenology coming from experiments and how addressing these limitations by future research is critical (e.g., whether variability in cue sensitivity across species is larger in temperate than tropical latitudes; see lines 289-292).

I believe the paper would significantly improve if this gap in the data and/or methods was addressed, ideally through the inclusion of more Southern Hemispheric or evergreen species data and the inclusion of other important phenological drivers and cues where relevant. Otherwise, a strong discussion of the transferability of this phylogenetic estimation of species-level phenology to other regions/phenological events/climatic drivers (e.g. precipitation), is required, especially in order to make the concluding impact statement the authors have drawn on such a broad scale "While we focused on spring phenology here, our approach suggests a path forward for more general forecasting of species-level climate change 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."

We agree. As outlined above, we believe there is little data to address this gap currently but we now highlight in the methods how much could be learned—both important to forecasting and to fundamentally understand the underlying evolutionary history of phenological cues—through more efforts to increase data beyond the temperate zone, especially in the Southern Hemisphere.

I also am concerned about the accessibility of the source code and data used to test and present the Bayesian hierarchical phylogenetic model in this study. Transparency of code does not equate to reproducibility of methods/analyses. If the authors believe that this robust new method will improve phenological forecasting predictions, their code and data should be reproducible to be implemented in future studies. As with my point above, please indicate the subset of data used in this study (locations, studies, references etc) rather than citing the original OSPREE dataset so that others can reproduce your Stan code. Likewise, making the Stan code more accessible through more comments (comments of the models - names or short description) rather than just the parts of the models ("slope" "intercept" "phylogeny") may allow other researchers to utilize your methods in Stan. Perhaps even making the source code open on a shared repository (e.g. GitHub) would also improve user accessibility in the future.

These are great points and we are happy to address them. As the reviewer noted, some of the data are already available on a federated data repository (KNB) with no limits on using the data. We have updated this entry with the full raw data and—given the reviewer's request—a subset specific to this paper is now available, together with the phylogenetic tree used to conduct analyses. Additionally, we have posted a GitHub repo – see https://github.com/MoralesCastilla/PhenoPhyloMM, where fully annotated code to conduct models in Stan and to analyse model outcomes in R, is available.

Minor comments:

Abstract
- First sentence to open the paper is important but here is unclear – is 'adaptation' referring to humans or societal adaptation to climate change? Or plant adaptation to climate change and therefore it should read "Knowledge of plants abilities to adapt to climate change hinges on accurate ecological forecasting to predict shifts in key ecosystem services such as carbon storage and biodiversity maintenance" as it isn't the ability of plants to adapt that hinges on the ecological forecasting itself but the knowledge or quantification of this adaptation hinges on the ecological forecasting.

We apologise for this lack of clarity, we were referring to societal adaptation; we have rephrased this sentence as follows: "Our ability to adapt to climate change requires accurate ecological forecasting to predict shifts in key ecosystem services, such as carbon storage and biodiversity maintenance."

Introduction
- Line 22: replace 'confounded' with 'limited' or 'restricted'
Done.

- The second and third paragraph of the introduction end on very similar concluding sentences making the introduction feel repetitive on the point that species-specific models are important. Although this is an important implication of the study, it would be good to see the concepts merged together in one paragraph or less repetition of this topic in these two paragraphs
Done, we have now deleted the last sentence from the first of these two paragraphs.

- Line 31: remove 'at once'
Done.

- Line 35: remove 'whereas,'
Edited, the sentence still needed an adversative conjunction.

- Line 39: remove 'especially'
Done.

- Line 44: elaborate on conflicting results as this statement comes across as vague. Provide some detail on the conflicting results.
Done, we now clarify how previous results differed, see lines 50-60.

Results and discussion
- Lines 140-156: Your results suggest that there is limited variability in species-level responses to photoperiod which is an interesting and unique result compared to previously published data. However there are limited species (and mainly those that are well studied and therefore have a better sampling of their populations across a broad geographic range) that have high variability. Could you test if variability in photoperiod is related to sampling effort or coverage? Perhaps you don't see a signal at the species-level here as the day-length in the local environments does not vary very much between populations, especially compared to variation in chilling and forcing.

We agree that this is surprising, which is why we devote a substantial portion of the discussion to interpret this finding. Our method already addresses issues with sample size (all models we use estimate effects based on the full dataset, adjusting for sample size and variance for each species), which we now mention in the main text (lines 236-240), and again in the methods (see lines 384-387). In addition, the variability in sampled photoperiods is very high–i.e., from 6h to 24h; see new Fig. S2 in the Appendix (with differing treatments within an experiment varying often by 4 or more hours)—given our focus on experimental cutting studies. We do detect a signal of photoperiod at the species-level, but its magnitude and variability is lower for this cue than for forcing and chilling for most species. Further, as we discuss (lines 154-165) the effect of photoperiod is very strong for one very well studied species, *Fagus sylvatica*, showing that our model can recover strong responses to photoperiod. The data, however, suggests that most species do not have a strong response, especially compared to other cues.

- Line 201: place a full stop at the end of the paragraph
Done.

- Line 213: 'forecasts for Acer campestre*, which has only 6 observations, shift by up to 35% in*

our phylogenetically informed 214 model' – what is the direction and unit of measurement of the forecast shift, is it the accuracy improving by up to 35% - clarify the result here
We had provided more details in the Fig. 4 caption, but now also clarify in the main text (see lines 221-224).

Figures
- Figure 1 requires some editing on the axes. Both the titles and labels need increasing in size on the x-axis (perhaps less breaks with greater font size on the axis ticks) and the list of species down the tree could be removed as it is nearly impossible to read. If it remains, it would require a larger font size and a removal of the underscore between genus_species for each species and an italicising of the species names. It would also be good to increase the font of the family labels on the right side or utilise different colours to show the family groupings more clearly.
We have done our best to accommodate the reviewer's suggestions. We would rather keep species names in the figure to facilitate tracking down species responses by zooming in. We have enlarged fonts where appropriate, italicized species names, removed underscores, and moved family labels closer to species names.

- Figure 2: place lambda and sigma in the figure itself and explain in the caption – will make the figure slightly more intuitive
Done.

- Figure 3: state the definition of the acronym PMM in the figure caption
Done.

Reviewer #2:

In this manuscript, the authors estimated species-level responses to two major environmental cues of spring phenology, temperature and daylength by using Bayesian hierarchical phylogenetic models. They found that predicting how each species responds to a combination of cues is more important than the focus on identifying which cue is the strongest. The findings are very interesting, with the methods very novel, and the conclusions are very crucial for improving current ecosystem models for predicting future shifts in ecosystem functions under background of climate change. However, the robustness of results depended on the accuracy of quantifying chilling, forcing and daylength from an meta-analyzed experimental dataset. The research lacks discussion to overcome the limited dataset problem, and how to fit multi-species model in different geographic sites is also a problem. Besides, is this model also work on phylogenetic trees originated from genome datasets? There are multiple spelling and grammar mistakes in this manuscript. I suggest a major revision decision for this manuscript.

We thank the reviewer for acknowledging the interest and value in our work, and for pointing out issues in need for attention. We agree that understanding the intricate relationships between forcing and chilling is complicated and have now expanded on this point (see lines 304-312), as well as on data limitations.

The reviewer is of course correct that our model can only be as good as the data that informs it, and is thus no different to any model fit to data. However, an advantage of our Bayesian approach is that we are better able to accommodate imprecision in the data that informs our model, which might arise from multiple sources, including measurement or experimental error, and the general stochasticity associated with limited sample size and unbalanced species representation. Critically, by partially pooling across species and weighting by phylogeny, we gain strength from species estimates that are informed by more data, such as within *Betula* and Fagaceae, but avoid skewing estimates for phylogenetically distant clades that may have been exposed to different selective regimes. As we mention in comments to Reviewer #1, above, part of the motivation for developing the model we present here was to help improve estimates when the available data are limited (see lines 231-246).

The reviewer raises another interesting point on the fitting of our model to species in different geographical areas, which can be extremely challenging when data are observational as disentangling cues, which often covary, presents numerous difficulties. The data we include here are from experimental manipulations, allowing us to compare treatments more easily across datasets. Nonetheless, species with different geographic origins will likely have adapted in response to local climates, and thus we might expect them to respond to cues differently. Our model allows species to have individualistic responses, which we argue is one major advance of our approach, and also to have responses that can covary by phylogenetic relatedness, which might capture shared historical selective regimes.

Regarding whether or not phylogenomic trees could be utilized with our method, yes, our method would potentially work with any phylogenetic tree. In our current implementation, we assume the phylogeny has branch lengths proportional to time. However, different assumptions could be made, and if the genes underlying plant responses to particular cues are known, it could be possible to estimate branch lengths directly from mutational changes along these sequences. In the absence of such detailed gene specific data, evolutionary time provides a useful proxy for species differences. We are now explicit about this in our methods (see lines 332-336).

Finally, we have conducted an exhaustive revision and corrected typos and grammar. Please see our detailed response to reviewer's comments below.

My two major concerns are as the following:
(1) Regarding chilling, the authors stated that they estimated field chilling by Utah units. However, the Utah model was first used for agricultural crops and assumes that temperatures between 1.4 and 15.9affect dormancy release differently. The assumption of the Utah model was against the recent findings that a wide range of temperatures (-2 to 10°C) have very similar effects on dormancy releases (Baumgarten et al 2021, doi: 10.1111/nph.17270). How do you prove the Utah model is suitable for species investigated in this study? To strength the findings, more reliable and various algorithms should be included (e.g., <5°C model). Furthermore, the start date of chilling accumulation is unclear, and 33 papers where the data of this study is from often used different dates. How to deal with this problem? Another problem is why just estimated field chilling? Did all papers use natural chilling treatments rather than artificial chilling treatments?

The reviewer raises an important point here, which we address in a two-fold manner. First, we have clarified our rationale for using Utah units (lines 304-312), which is common metric to estimate chilling. We choose this metric because several papers reported chilling as Utah units (with no additional temperature or duration information), thus the choice of Utah units allowed us to use the largest amount of data. However, because chilling is effectively a latent process for which we have no true measure, there are many models for it and current evidence is still equivocal for at what temperatures and how 'chilling' accumulates (e.g., Baumgarten's results suggest chilling could occur at negative temperatures, or at temperatures $> 10°C$).

Second, we fit an additional model of chilling, per the reviewer's suggestion to include additional algorithms. We ran both our models (PMM and HMM) using another metric of chilling, chill portions instead of Utah units. Model outcomes show similar results, strengthening our findings—we have these results added as two new tables in Supporting Information (new Tables S4 and S5).

(2) In some papers, forcing temperatures differed between day and night. Sometimes, the forcing temperature is always changing (e.g., gradually increased), how to determine forcing temperature in these cases? Similarly, some experimental papers used a changing photoperiod (e.g., gradually increased daylength), how to determine the daylength in this case? All these details could not be found in Methods & Materials.

This is a great point and something we thought and worked deeply on, but we agree that we should have explained it in more depth. We apologize this was not clear enough before and have now clarified our methods in lines 313-318:

> Forcing and photoperiod treatments occurred after chilling treatments; we report photoperiod as the length of light and weighted these treatments by the reported photo- and thermo-periodicity (Buonaiuto et al., 2023). Most studies reported two temperatures per day across the whole experiment, one for day and night, but some had ramped temperatures and/or photoperiods (or other complexities). In these cases we built an hourly model of the full treatment period until budburst and took the mean value.

Besides, there are some minor points:

Abstract:
Add one or two lines for introduction of the dataset you applied for the analysis and construction of ecological models.
Done, see abstract lines 5-7.

Introduction:
Line 10. "High variability observed in responses....". Please tell what Kind of variability?
Done, see lines 10-11.

Line 13- 15. "Much of it, however,... few well-studied species". Not logical. Please check and

revise.
Done, see lines 14-16.

Line 24-27. check and improve the expressions. Besides, in this section there is a lack of state of the art about the phylogenetic signals in plant phenophases and their relationship with climatic responses. The author should discuss it.
We now provide a bit more background on phylogenetic signals of plant phenophases below, see lines 40-60.

What is the mechanism of shared evolutionary traits in plant phenophases, and how to predict these with climatic variations should be discussed in the introduction part.
We now expand on putative mechanisms underlying phylogenetic conservatism in phenological traits (see lines 42-50). Specifically, we clarify what phylogenetic conservatism represents—that more closely related species share more similar phenologies, likely reflecting evolutionary conservatism in responses to common cues. We also discuss two broad mechanisms that could give rise to patterns of evolutionary conservatism in phenology. First, close relatives will tend to share similar ecologies and physiologies, and thus be sensitive to similar environmental pressures. Second, close relatives derive from common geographic centers of origin, and thus their ancestors will have been exposed to—and have adapted to—similar environmental cues. Importantly, underlying both mechanisms is evolved plant sensitivity to environmental cues.

Results and Discussion:
I suggest add some evolutionary prospectives and their relationship with phenological sensitivity to three environmental cues, chilling, forcing, and photoperiod.
We provide an evolutionary perspective in the section "Phylogenetic structure of phenological cues" and especially its third and fourth paragraphs discuss in depth the evolutionary implications and interpretation of our results. We purposely avoided over-interpreting our results but would be happy to add additional prospectives if the reviewer wishes to provide more specific guidelines of what additional points should be covered.

Figure 4 is not precise and should be revised to reflect the frequency distribution and future forecasting.
The figure shows shifts in the forecasts and frequency distributions of these shifts. In addition, we provide a complementary figure (Fig. S8 in Supporting information) that specifically shows differences in forecasts. The aim of the figure(s) is to compare forecasts among methods and show when their results are discrepant—i.e., underrepresented species. We believe that Figure(s) 4 and S8 serve this purpose.

What are the limitations of this model? The study for further research should be added in the discussion section also.
We now discuss some limitations of, and future research from, our work. This includes data limitations (see lines 289-292, and lines 231-246), and we suggest future research venues in a different part of the manuscript (lines 256-264, and lines 290-292).

Methods section:
What are the criteria for choosing the citations after yielding the search from ISI Web of Science and Google Scholar?

We reviewed all citations, then included those where we could calculate chilling, forcing and photoperiod treatments, see lines 270-283.

Line 260-263. "For our analysis here, ...resulting in 44 studies from 33 papers". These lines are unclear, and no parameter was discussed to extract the results. The author should revise these lines.
We have clarified the paper selection criteria, which required either raw experimental data or included a figure from where data could be extracted, see lines 281-283.

Line 275. How do the Polytomies work and affect only 46 out of 191 species? Please discuss in detail. Line 287-295. These are very common and should be discussed further in one or two lines.
Polytomies represent parts of the phylogenetic tree lacking resolution. In our dataset, polytomies were introduced when we added species missing from the mega-phylogeny of plants. We now explain this in more detail and provide an example of *Acer*, for which several species where included in the OSPREE dataset but the megatree lacked species-level resolution within the genus. We now make clear that, while errors in phylogenetic topology and branching times could impact model estimates, if errors were large the contribution of phylogeny would effectively be scaled to zero by the λ transformation that is simultaneously fit in our model. To assess whether the inclusion of polytomies in our data biased model estimates, we ran sensitivity analyses excluding these species from models (see Table. S8).

The authors should describe the importance and significance of the Bayesian hierarchical phylogenetic model in this section.
We stress the importance of our phylogenetic model and how it differs from previous approaches in lines 87-97.

The authors should add a detailed description table of all studies for this research.
We thank all reviewers for this suggestion. We have added a map, a table and a figure in the Supporting Information with a description of studies, their geographic distribution and the underlying data (see Table S1, Figs. S1, S2).

Reviewer #3:

This study, titled "Phylogenetic estimates of species-level phenology improve ecological forecasting", incorporated evolutionary history to study the impacts of different environmental cues (temperature and daylength) and concluded that doing so improved forecasting of plant phenology. I welcome this message yet there are some critical flaws in the methods.

Thank you!

*First, the conclusion of this study was largely based on the comparisons between the results of the phylogenetic and the non-phylogenetic (traditional) mixed models. However, doing so won't allow us to concluded that the phylogenetic models are *better / improved* than the traditional models. To prove this, one needs to compare the phylogenetic models against the "true" parameters and to compare the non-phylogenetic models against the "true" parameters, then compare which models' results are closer to the "true" parameters. This is easy for simulation but no*

so for empirical studies as the "true" parameters are unknown in general. However, this study used data collected from controlled experiments, which should allow the users to calculate an approximate of the "true" parameter. I did not see any visualization of the raw data, which should be the baseline of the comparisons instead of the results of the non-phylogenetic models.

We thank all reviewers for highlighting the need for a better description of our data and data sources. We have added a map, a table, and an additional figure in the Supporting Information to provide detail of the studies included, illustrate their geographical distribution, and to meaningfully present the raw data (see Table S1, Figs. S1, S2). We would be happy to expand on this further if the reviewer has additional suggestions. The reviewer also queries whether our phylogenetic model is 'better' than more traditional models. This is an important point. 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. A separate, but related question is whether the extra effort of fitting the phylogenetic model is worthwhile. We now evaluate this directly by performing formal cross-validation analyses, and comparing predictive accuracy on hold-out data between the phylogenetically informed model versus the model without phylogeny. We describe this new analysis in more detail in comments below.

Second, similar as the first point, to compare the performance of the phylogenetic and the non-phylogenetic models in forecasting, one needs to compare with the "true" patterns. Given that the authors used some historical time windows, this should be doable. Again, the fact that having different results when fitting the phylogenetic and the non-phylogenetic models do not necessarily mean that the phylogenetic models are better. One needs to compare with the "true" parameters.

As we describe above, our approach does not inherently assume that a phylogenetic model will perform better than a non-phylogenetic model, but rather allows for the possibility that phylogeny might be important. However, we appreciate that it would be useful to know whether estimates are more accurate when we account for phylogeny. It would, of course, be ideal to compare modelled parameter estimates to known 'true' values, unfortunately, even though the data we fit to our model is derived from experimental treatments, we do not have information on true parameter values (even if comparing historical time windows, we would lack accurate response data for most of the species in our dataset). We have attempted the next best thing, and perform cross-validation to compare model fits. In this new analysis, we iteratively drop a genus from the data set (a sensible approach given the structure of the data), and then fit models to the reduced taxon set - in the man text we refer to this as Leave-One-Clade-Out cross validation. We examine two measures of model performance. First, we evaluated stability of model estimates to subsetting the data. Second, we used fitted models to predict response values for the left out (out of sample) species - a traditional cross-validation approach. Both analyses confirm that the phylogenetic models outperform the non-phylogenetic models: model coefficients were more similar between subset models and full models, and subset models predicted observed responses of left out species better in the phylogenetically informed models. We have included a new section in the Supporting information (see "Leave-One-Clade-Out cross validation") with full details of the new methods and results.

Third, the authors set up the phylogenetic model with the phylogenetic var-covar matrix of the form $\sigma\lambda\Sigma$, which is equal to the traditional model when $\lambda = 0$, and when $\lambda > 0$, it will be a

phylogenetic model. This method, however, indicate that it is an 'either / or' logic here. In real world, it is most likely that we have both phylogenetic and non-phylogenetic components of variations. A better way would to set an additional non-phylogenetic term along with the phylogenetic term explicitly (e.g., $\lambda\Sigma + (1-\lambda)I$). In this way, both the phylogenetic and the non-phylogenetic terms can be estimated simultaneously, and one can test whether the phylogenetic component is necessary.

We agree with the reviewer about the importance of including both phylogenetic variation and variation that is independent of phylogeny in the model. In fact, our model already includes both components of variation. In Equation (1), σ_e^2 is a variance term that is completely independent of phylogeny. Furthermore, (in contrast to the reviewer's statement) it is not the case that our phylogenetic variance-covariance matrix has the form $\lambda\Sigma$ (we omit the leading σ term in the reviewer's first equation under the assumption that Σ is already a variance-covariance matrix). Rather, the phyogenetic variance-covariance matrix in Equation (4) represents a λ transformation of a variance-covariance matrix, which is exactly equal to reviewer's suggested equation, $\lambda\Sigma + (1-\lambda)\sigma^2 I$ (note we have added σ^2 before I as we believe this was the reviewer's intention, given that σ^2 will generally be much larger than $1-\lambda$). To illustrate:

$$\begin{aligned}
\lambda\Sigma + (1-\lambda)\sigma^2\mathbf{I} &= \begin{bmatrix} \lambda\sigma^2 & \lambda\sigma_{12} & \lambda\sigma_{1,3} \\ \lambda\sigma_{2,1} & \lambda\sigma^2 & \lambda\sigma_{2,3} \\ \lambda\sigma_{3,1} & \lambda\sigma_{3,2} & \lambda\sigma^2 \end{bmatrix} + \begin{bmatrix} \sigma^2-\lambda\sigma^2 & 0 & 0 \\ 0 & \sigma^2-\lambda\sigma^2 & 0 \\ 0 & 0 & \sigma^2-\lambda\sigma^2 \end{bmatrix} \\
&= \begin{bmatrix} \lambda\sigma_1^2+\sigma^2-\lambda\sigma_1^2 & \lambda\sigma_{12}+0 & \lambda\sigma_{1,3}+0 \\ \lambda\sigma_{2,1}+0 & \lambda\sigma_2^2+\sigma^2-\lambda\sigma_2^2 & \lambda\sigma_{2,3}+0 \\ \lambda\sigma_{3,1}+0 & \lambda\sigma_{3,2}+0 & \lambda\sigma_3^2+\sigma_3^2-\lambda\sigma^2 \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & \lambda\sigma_{12} & \lambda\sigma_{1,3} \\ \lambda\sigma_{2,1} & \sigma^2 & \lambda\sigma_{2,3} \\ \lambda\sigma_{3,1} & \lambda\sigma_{3,2} & \sigma^2 \end{bmatrix}
\end{aligned} \tag{1}$$

Assuming that the covariance terms above, $\sigma_{x,y} = \sigma_i\rho_{xy}$, where ρ_{xy} is the phylogenetic correlation, this matches Equation (4) in our paper.

Also, why did not include "study" as a random term here? I expect multiple measurements from the same study, right?

Some of the major differences between studies are accounted for in how we have carefully calculated the treatments of chilling, forcing and photoperiod. Beyond that, some species in our data occur across many studies, but most occur in only one study (162 out of 191 species) making it very difficult to separate out species versus study effects. This is not a new problem (see Kharouba et al. 2018:PNAS, 20, 5211-5216, for more discussion), but a difficult one. We're currently working on a version of this model with statistical expert Michael Betancourt to try to tease out study from species (in a phylogenetic framework) from a project using far more data and still struggling. This goal, unfortunately, is not possible for this dataset currently, which we

now acknowledge in the methods (lines 387-391).

Minor comments:
L80-84: Disagree about the claim here. The models described in Ives and Helmus 2011, as well as Hadfield 2010 also allowed different estimations of the phylogenetic components of different predictors. In other words, the model described here can be fitted with those approaches too.

Based on the reviewer's feedback, we can see how this may have been confusing. We have now removed references to the models by Ives and Helmus and by Hadfield and made it explicit that we meant how traditional phylogenetic models were concerned with phylogenetic correction.

We did reach out to Anthony Ives to try to code our phylogenetic model in "phyr" and never could, though he suggested it should be possible. However, more ways to implement these methods would be valuable in our opinion, so if the reviewer can provide code to implement similar phylogenetic models in these packages, then we would happily test and include this.

L100-108: In theory, models without considering auto-correlations should still give unbiased estimations of the mean.

We did not intend to imply that HMM estimates were biased. We simply meant to report how much coefficients shift from one model to the other, showing that shifts are not too large. This issue has been discussed in papers comparing coefficients among models with and without considering auto-correlations (e.g., Bini et al. 2009:Ecography, 32, 193-204). To address this, we now go further comparing the PMM and HMM outcomes in the new Supporting section on model cross-validation. If the reviewer wishes, we could include some discussion of the topic (we have not yet, as we fear doing so would divert attention).

L174-188: Good discussion here.
Thank you!

L263: What is the total number of data points?
Done, this is now included.

L275: With a branch length of 0? Or do you mean with a branch length of the congeneric basal node age?
Thanks for spotting this, corrected.

*Eqn 3: what is n?? Number of species? If so, you already used *j* to represent species in eqn. 1, why use a different letter here?*
Yes, n in equations 3 and 4 referred to the total number of species, as made explicit in the lines between equations. We believe we need a j^{th} index to refer to each individual species. We will gladly change this if the reviewer has any specific suggestion. We could also switch to sp, in lieu of j, if the reviewer thinks that would help clarity.

*L308: Trait *i*? *i* was used previously for observations, can you use a different letter to be less confusion?*
Thanks for catching this. We have now removed the *ith* index there as it did not seem necessary.