Editor and reviewer comments (we provide below the full context of each review) are in *italics*, while our responses are in regular text. Comments by Lizzie (that need to be deleted) are in blue.

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 perhaps not clear enough. We also now provide additional tests to compare the phylogenetic and non-phylogenetic models using a leave-one-out approach (see Appendix SXX), provide more details on data sources and their distribution (see Supporting Table SXX, and Figures SXX-XX), and re-analyze data using a different metric of chilling (see Figures SXX-XX). We thank the reviewers for their constructive feedback.

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. We certainly will make the full dataset available on KNB, as well as the specific subsets requested by reviewers. Further, we have posted a GitHub repo 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 review makes a fair point, we should have included a more graphic description of our dataset. We now include both an appendix showing the location of studies in the dataset (see Supporting Fig. SXX), a summary of the data that also accommodates a request by reviewer3 (see Supporting Table SXX), and have made the following edits to the methods in the main text (see lines XXx-XX). We would like to highlight that the geographical bias present in our dataset mirrors the existing bias in the literature. We have incorporated all studies meeting our standardized search criteria (refer to Methods, lines XX-XX), resulting in a limited yet valuable dataset that accurately represents species from temperate biomes in the Northern Hemisphere, including 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 Supp)."

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 undersampled as species to be included (see lines XX-XX). 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 other species. 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 discussion about the limitations of current data on phenology coming from experiments and how adressing these limitations by future research is urged and raises new exciting research questions (e.g., is variability in cue sensitivity across species larger in temperate than tropical latitudes?; see lines XX-XX).

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 completely agree. As outlined above, we believe there is little data to address this gap currently but we now highlight in the discussion 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 utilise 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 will update 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 made a GitHub repo – see LINKwhere 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.
I see the point of the reviewer, but then it would not be the knowledge but the little or unsufficient knowledge we have. Any ideas on edits here are welcome, Lizzie, Jonathan.

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
Should we join the two paragraphs or remove the last sentence from the second?. It does not feel so repetitive to me, but ok editing to show our good intent.

- 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.

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 a surprising finding ... but our methods are already addressing issues with sample size and our variability in sampled photoperiods is very high, given our focus on experimental cutting studies ....

I would (1) edit this section a TINY amount to mention the range of photoperiods in the treatments and (2) edit the methods a bit to say that these methods are especially robust to uneven sample sizes and variances.

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

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.
- Figure 2: place lambda and sigma in the figure itself and explain in the caption – will make the figure slightly more intuitive
- Figure 3: state the definition of the acronym PMM in the figure caption

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.

Add some text here we completely agree about the complexity of understanding forcing versus chilling ... tweak some bits of the discussion and reference those lines in a response here. We also need to clarify we calculated chilling across all studies using start dates and we did it for field chilling and experimentally-applied chilling ...

I am not really sure what the 'model also work on phylogenetic trees originated from genome datasets' thing means ... we could clarify the methods behind the off-the-shelf tree in the methods section with a half-sentence .... Jonathan suggested we could also say you can put in whatever kind of tree you want; so if you had a tree based more on the underlying genes you could add that ... but need to check with Will if tree always will need to be ultrametric.

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 15.9 °C affect dormancy release differently. The assumption of the Utah model was against the recent findings that a wide range of temperatures (-2 to 10) 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?

Best way to deal with this if not too hard: We also could do chill portions instead of Utah chill and report those in the supp, that would be a STRONG reply and a good way to say to the editor we ran all new tests, got new data (chill portions) etc.

(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 (indeed our efforts led us to write an entire paper on the complexity and importance of this problem, see https://besjournals.onlinelibrary.wiley 2435.14329). We apologize this was not clear enough before and have now clarified our methods in lines XX-XX.... To address different day/night or ramped temperatures we take a weighted hourly average over the full period of forcing. We also adjusted for differences caused by co-varying thermo- and photo-periodicity for both the forcing temperatures and photoperiod values.

Tweak methods to make sure we're clear on this, then line reference here.

Besides, there are some minor points:

A lot of the non-method points below are a touch weird to me ... for example, 'suggest add some evolutionary prospectives' makes me think this person wants us to go crazy with what the phylogeny tells us; I advise against this. I suggest really minor edits for most of these points and then reference lines where we already do what they're asking ... for example the whole paragraph 'Weak phylogenetic signal in photoperiod sensitivity...' is about evolutionary perspectives.

Abstract:
Add one or two lines for introduction of the dataset you applied for the analysis and construction of ecological models.

Introduction:
Line 10. "High variability observed in responses….". Please tell what Kind of variability?
Line 13- 15. "Much of it, however,… few well-studied species". Not logical. Please check and revise.
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.
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.

Results and Discussion:
I suggest add some evolutionary prospectives and their relationship with phenological sensitivity to three environmental cues, chilling, forcing, and photoperiod. Figure 4 is not precise and should be revised to reflect the frequency distribution and future forecasting.
What are the limitations of this model? The study for further research should be added in the discussion section also.

Could reference a new paragraph about non-temperate and Southern Hemisphere species here.

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

We review all citations, then included those where we could calculate chilling, forcing and photoperiod treatments, see line XX.

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.
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.

Maybe cite a paper that explain polytomies?

The authors should describe the importance and significance of the Bayesian hierarchical phylogenetic model in this section.

The authors should add a detailed description table of all studies for this research.

Yes! Deirdre made an awesome table for one of her papers, she might be able to help ... though we should try to evenly pass out tasks.

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 phenol-

ogy. 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.*

Some nice visualizations of the raw data in the supp would work.

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.

It seems like the reviewer specifically does not want us to do simulations ... Not clear what they want though.... Perhaps subset the data and do CV using the current model and lambda=0 models? TO DISCUSS.

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.

Get Will to write a response to this (start asking NOW and ask OFTEN so he does this by deadline). We could also add something to the supp explaining the model more.....

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

This is a great question and something we are actively working on. 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 (CHECK?) making it very difficult to separate out species versus study effects. This is not a new problem (see cite Kharouba et al. PNAS 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 species (in a phylogenetic framework) from study 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 XX-XX).

We can add a half or full sentence to methods along the lines of 'Many species occurred in only one study, making it difficult to separate the effects of study and species, thus we do not include study as a separate parameter here and average over it in our model estimates.'

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 approches too.

Hmm, we could just remove the two refs and leave Freckleton. We never got these models to work in other packages.. My last email from Tony Ives is: "It should take (phylogenetically) random slopes pretty easily. Having multiple observations for the same species shouldn't be a problem; you can just include a species-level random effect. (Okay, to do this in phyr you have to code the matrices yourself, but you've already done that for stan).' but I think adding a species level random effect is not what we want... I suggest we pull the refs but reply in saying we did reach out to Tony to try to code these in phyr and never could. However we'd like to see these methods used more, so if the reviewer can provide code to implement the models we have here in these packages then we'll happily test and include it ... TO DISCUSS though a minor point.

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

L174-188: Good discussion here.

Thank you! This actually is a lovely paragraph .... including our sigma simulations.

L263: What is the total number of data points?
L275: With a branch length of 0? Or do you mean with a branch length of the congeneric basal node age?
*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?*
*L308: Trait *i*? *i* was used previously for observations, can you use a different letter to be less confusion?*