*Reviewer Comments are in italics.* Author responses are in plain text.

**Reviewer #1 (Remarks to the Author)**

*To Authors*
*As I said in my last review, this study addresses an important question using a really impressive dataset and sophisticated analyses. It is well written and the authors have done a good job of addressing my previous concerns ? certainly the methods are easier to follow now and I appreciate the addition of the PRISMA checklist.*

We thank the reviewer for this positive feedback and helping us improve the manuscript. Based on this most recent review we have found many places where additional clarity was needed, especially in differentiating between our meta-analysis of short-term experiments in controlled (generally lab) environments and how we applied it to long-term data. We hope our current manuscript is much easier to follow and understand.

*However, I'm afraid I have some major concerns about the analyses that have become clearer to me now that I understand the methods better. The finding that trees are more sensitive to chilling than forcing is surprising based on the existing literature, and for the reasons I lay out below I am concerned that this finding may not be robust.*

We completely agree with the reviewer that the existing literature is inconsistent on the relative strength of chilling versus forcing, but we believe an important distrinction here is between estimates of chilling and forcing based on long-term observational data versus data from short-term controlled experiments. As the reviewer suggests, there are many studies, mainly observational ones to our knowledge, that suggest higher sensitivity to forcing (e.g., **??**), however, there are also many studies that find higher sensitivity to chilling (e.g., **???**) and/or variation in the strength of forcing versus chilling varies across species (e.g., **?????**)—and these are all from short-term experiments that manipulate chilling and forcing in more controlled settings. Indeed, these inconsistencies, in part, motivated our meta-analysis of experiments. Further, we say in the abstract (lines line 11-line 13):

> Our results unify both sides of the debate over phenological cues: while all species may respond to all cues strongly in experimental conditions, in current environmental conditions the dominant signal of climate change is from increased forcing.

We have been in contact with the reviewer over email (and attach our email correspondence), which highlights that a major cause of these concerns is due to lack of clarity of our methods and the methods of studies that we synthesized. In this revision, we have clarified the methods and also worked to address all the concerns, as described in-detail below.

*1. Non-separation of temporal variation in drivers from spatial variation. As I understand it the focus of this study is on the effect of the drivers on temporal variation in budburst. However, the drivers (forcing/chilling/photoperiod) vary across space as well as time and I think the model does not take this into account. This means that the effects estimated are an average of the spatial and temporal effects and given that much of the variance in drivers will be spatial rather than temporal the bias this introduces could be very substantial. This issue is explained very clearly by Van de Pol and Wright 2009 and a simple remedy is to use within subject (ie. within study) mean centering for the drivers. In order to get standardized effects the z transformation could then be applied after within subject centering.*

We understand the reviewer's concern that spurious correlations with space can drive results (e.g., days to budburst may vary geographically as well as across years). Separating temporal versus spatial variation might be an especially critical component of long-term observational studies in the natural world, and within subject centering may be an effective approach in these

cases. Our study, however, uses experiments where temperature and photoperiod were generally highly controlled, thus we did not expect within-group centering should affect our results strongly.

We applied within-group centering to our data and found virtually no effect on our estimates of chilling, forcing and photoperiod (see Table below and compare to Tables XX and XX in Supplement). As expected, if our data do not have the bias the reviewer is concerned about, our slope estimates were unchanged, while our intercept estimates converged on those of our standarized-predictor (z-scored) model. Give that these estimates are so similar to the estimates we already present, we have chosen not to add this to the supplement, though we can if requested.

Instead, we have changed our language throughout the manuscript to better define what types of studies our meta-analysis focuses on and to provide more clarity on how we apply model estimates to climate data from Central Europe. In re-reading the manuscript, we could see this was often unclear. In particular we have reduced our use of the phrase 'controlled environment studies' and more often refer to 'experiments with controlled temperature and/or photoperiod conditions.' This occurs throughou the paper, including in the abstract (line 6), where instead of referring to 'controlled environment studies' we now call them experiments, and mention again 'controlled conditions'), also pleas see line 45-line 46 and line 53-line 53 and line **??**, line 86 and line 103, line 112,line 146, line 149.

We additionally edited **all** figure captions for clarity. In particular, in the caption to Fig. 4 we can see how it would be easy to think we are using phenological data with potentially important spatial autocorrelation. Instead, we are using estimates from our meta-analysis of experiments applied to climate data from Central Europe. We have worked to clarify this in the caption and text (see above for line numbers changes in the main text).

*2. Is it really chilling? My gravest concern relates to a point raised by reviewer 3 on whether the approach taken is adequately estimating chilling or whether it instead contains a forcing signal. The authors attempt to address this with a sprinkling of caveats about the chilling portion being a hypothesis (though this is not apparent in the abstract) but I think this issue greatly undermines what can be inferred from this approach and the key finding of the study. There is a lot that is good about this study, but the limitation of the methods for robustly teasing apart chilling from forcing means that I think it confuses our understanding more than it advances it.*

We completely agree with the reviewer that disentangling forcing from chilling is a major challenge in phenology research today, however, our meta-analysis uses an experimental design that is widely agreed to be the best current widely used method for disengtangling these effects (ADDCITES). Current research at the cellular level is working to address this challenge by identifying what exactly underlies chilling (e.g., work on the compound callous **??**), but until this research is successful and tested across other species the experiments we use here represent our best method to attempt to disentange forcing and chilling effects.

Thus, we consider this critique to be not one directed at our approach, but at the entire field of phenology that uses these experiments—a field with over a 60-year history. Indeed, one motivation for this paper is to highlight the need for additional work on this and other aspects of spring phenology. As we note in the current abstract, "Further progress to improve budburst forecasts under future climate change will require fully separating chilling and forcing effects at the physiological-level." And in response to reviewer 3's previous comments we have stregthened and clarified this point throughout the manuscript and highlighted it via Fig. 1 also.

We do believe this concern may come in part from a lack of clarity that we are focused on short-term experiments that manipulate temperature and/or photoperiod and have worked to

clarify the design of or study throughout the manuscript (see reply to 1. above for more details).

*3. Measurement error. The fact that 75% of studies had a sample size ¡ 8 suggests that measurement error is likely to be substantial. On page 7 of the methods it is stated that measurement error averages just 9.9% of the response variable. However, if the studies that report a standard error tend to be the ones with larger sample sizes then this issue may be worse than suggested by the authors.*

Here are the ideas that I have to deal with this:

- We could simulate effects of adding measurement error for each study, using a range of variance and sample size (perhaps min, max across range of studies for which there is this information?). I could imagine doing this in a couple of different ways. We could add a step prior to fitting the bb model in which we randomly draw budburst responses from a distribution (simulated using the reported response as the mean, the reported variance and n from the study). We then fit the model. We could do this 100 (1000?) times and see how much it alters the estimated effects.
- Alternatively, we could include the full simulated distribution into the data fit to the model and add a new random effect of "study."
- Either way, we could/should check the 25/39 studies that do not have sample size and/or variance currently in OSPREE in case the information was reported but we failed to capture it- I'm worried that these data were inconsistently entered into OSPREE.

*4. Chilling and forcing time: I apologise if I have overlooked this but I still cannot find in the methods or main text a clear statement of the dates over which chilling units (and forcing units) were calculated. Figure 1 is helpful but does not include a specific statement about timing. If timings are idiosyncratic to each study this should be made clear and it would be really helpful to have a figure that shows for each study the time period of chilling and forcing. This would also help the reader to evaluate whether the 'chilling' metric is distinct from forcing.*

Again, we apologize as we suspect this concern may be related to a lack of clarity on our meta-analysis focusing on short-term experiments that manipulate temperature and/or photoperiod, which we have attempted to clarify throughout the manuscript (see reply to 1. above for more details). As these are short-term experiments there is no consistent temporal window of when chilling was applied. We now mention that treatments vary by study early on in the caption to Fig. 1 and in the main text (see line **??**), where we reference a heatmap figure that shows the treatments we have via our meta-analysis.

The length of time that chilling treatments were applied (as well as the temperature of these treatments) varied across experiments: chilling treatments from 1 to 182 days in duration (mean = 71.4 days) and temperatures ranged from 0 to 16 °C (mean = 4.4°C). The predictor variable "chilling" in our model is derived by applying standard chilling calculations to estimate the amount of chilling applied in these chilling treatments (we use both Utah units and Chill Portions in separate models, to compare the effect of using different chilling metrics). The predictor variable "forcing" is simply the forcing temperature applied; this also varied across experiments. To clarify this, we have also added the following to the legend of Figure 1, "Studies in the OSPREE database applied chilling treatments that ranged in duration from 1 to 182 days (mean = 71.4 days) and temperatures ranged from 5 to 32 °C (mean = 4.4°C)."

More detailed information, e.g., the temperatures and durations for forcing and chilling in each study can be found in the OSPREE database, which will be publicly available upon publication via the Knowledge Network for Biodiversity (part of DataONE, which should make these data discoverable through multiple portals).

*5. Random regression covariances. In the random slopes model it looks as though the variance in slopes across species for one driver is fitted as being independent of the variance across other drivers. I think the covariances between these random slopes and the with the random intercept should be estimated i.e. estimate a 4 x 4 covariance matrix (alphasp, betaforcing, betaphotoperiod, betachilling).*

We have added parameters to our main model to allow slope and intercept to vary, and added a table to the supplemental methods comparing the estimates of both models. We did not find that adding the covariance matrix altered parameter estimates very much.
*Minor comments*

*Line 26. Insert "forcing" before temperature.*

We thank the reviewer for this suggestion, and have made this change.

*Line 107. I'm not convinced that it is often found to be the most important cue? it may be highly dependent on how you define importance. If importance is defined as it's influence on year to year variation then in the UK we find chilling to be a less important cue than forcing –see fig 1 in Roberts et al.*

We agree with the reviewer, as we say in the abstract (see line 13), "in current environmental conditions the dominant signal of climate change is from increased forcing," and we have worked to clarify our methods throughout (see reply to 1. above for more details). We have also adjusted the text to further clarify our meaning by adding additional citations and clarifying we refer to experiments, line **??**-line **??**:

> This has not been widely suggested previously, perhaps because little experimental work has directly manipulated chilling, and the few studies that have were designed to compare chilling versus photoperiod effects *(e.g.,* **????***)*, not chilling versus forcing effects. Process-based phenological models, however, that explicitly model chilling often find this cue to be most critical *(e.g.,* **???***).*

We now also cite **?** on line **??**.

*Last paragraph of methods: The start date of GDD models does not have to be specified by the researcher, it can be estimated from the data.*

We thank the reviewer for pointing this out, and estimating a start date for GDD (as done in **?**) may be a useful approach for many questions. For this aspect of the methods, the goal was to examine potential statistical artifacts in estimating changes in forcing sensitivity. To evaluate this potential, we chose a specific start and end date because this is an approach used by many studies (e.g., XXX) and because this simplification allowed for a more straightforward understanding of the potential effects of statistical artifacts.

*Table S2. Are the window open and closed in ordinal days? I'm also skeptical as to the informativeness of fitting a sliding window to just 10 years of data.*

Yes, the windows are opened and closed in ordinal days; we have added the following XX to the Table legend to clarify this. To do this analysis, we followed the methods in **?**, the reference suggested by the reviewer in the previous version. We were happy to incorporate this and feel it has strengthened our manuscript. Additional sliding window analyses may be interesting, and we think would be an excellent topic for a paper focused on sliding-window methods, which are designed for observational data. Our manuscript is focused on short-term experiments, which cannot easily be used with the sliding window approach—we have worked to clarify our methods now (see reply to 1. above for more details).

*Signed*
*Ally Phillimore*
*(I sign all of my reviews)*
*References*
*Roberts AMI, Tansey C, Smithers RJ, Phillimore AB (2015) Predicting a change in the order of spring phenology in temperate forests. Global change biology, 7, 2603-2611.*
*Van De Pol M, Wright J (2009) A simple method for distinguishing within- versus between-subject effects using mixed models. Animal Behaviour, 77, 753-758.*

*Reviewer #2 (Remarks to the Author):*

*Thanks for the revision. The authors made a good response, and most of my concerns were responded. As I pointed before, this is an interesting study in quantifying the relative importance among the most important 3 cues in spring phenology, and thus be valuable for global change ecology studies.*

*However, I still not fully convinced the chilling effect overweight forcing and photoperiod. Could the uncertainty in the experimental studies be quantified in the hierarchical Bayesian model? The experimental studies were theologically designed to estimate one cue effect, but interactive effect with other cues were actually not excluded, thus the solely effect of one cues might be overestimated.*

We appreciate the reviewer's concern, as this is one we had ourselves when we saw the results and we have worked to interrogate our model fully (we present XX additional versions of our primary statistical model in the supplement for this very reason). We have addressed this in several ways. We now reference a heatmap figure that shows the treatments we have via our meta-analysis in caption to Fig. 1 and in the main text (see line **??**). Additionally, we ... add a new model ... and update the table.
reference Table S10 that most studies (36/42) included more than 1 cue. Should we add the studies that included only 1 cue? I thought they had to include more than one...but the numbers in the table only add up to 36. also relates to Rev 1 interest in uncertainty
*In addition, the authors argued that the decreased winter temperature during hiatus is not necessarily resulting an increase in chilling, but warming winter reduced chilling as most studied reported, thus both warming and cooling winter would reduce chilling? This is tricky, and may overestimate the chilling effect.*

We believe we may agree with the reviewer but did not clarify well enough in our previous drafts when we were we referring to chilling in short-term experiments that controlled temperature and photoperiod versus long-term climate data from Central Europe. In response to this Reviewer and Reviewer 1, we have worked to clarify this throughout our manuscript. In particular we have reduced our use of the phrase 'controlled environment studies' and more often refer to 'experiments with controlled temperature and/or photoperiod conditions.' This occurs throughou the paper, including in the abstract (line 6), where instead of referring to 'controlled environment studies' we now call them experiments, and mention again 'controlled conditions'), also pleas see line 45-line 46 and line 53-line 53 and line **??**, line 86 and line 103, line 112,line 146, line 149. We additionally edited all figure captions for clarity.

*Anyway, this is valuable investigation in quantifying the environmental effects on spring budbreak spring, but the reliability is still need further estimation.*

We completely agree further work is needed and have aimed to stress this in our manuscript (see line **??**, line **??**, for example).

*Reviewer #3 (Remarks to the Author):*

*The authors have done a great job in revising the manuscript. The additional analyses and figures they present have clarified the points raised in the previous review round and greatly improved the overall presentation of the data. I agree with all the conclusions and have no more comments.*

We thank the reviewer for the time spent reviewing our manuscript, and hope it will encourage more work on separating out chilling and forcing effects.