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.
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 thank the reviewer for his time and thoughtful critique. We agree that existing literature is inconsistent on the relative strength of chilling versus forcing. As the reviewer suggests, there are many studies, especially observational ones, 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., **?????**). Indeed, these inconsistencies, in part, motivated our meta-analysis of growth-chamber experiments.
We have been in contact with the reviewer over email, and have attached 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 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.*

   The reviewer brings up some interesting points about spatial versus temporal variation (e.g., days to budburst may vary geographically as well as across years). This is certainly the case. 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.

   The focus of our study is to understand the effects of chilling versus forcing versus photoperiod as drivers of temporal variation in budburst (i.e. days to budburst), and, to address this, our study synthesizes short-term experiments conducted in controlled environment chambers. There are two main reasons why controlled environment chamber experiments are unlikely to show the substantial temporal and spatial variation described by the reviewer. First, these are short-term studies, with nearly all experiments (27/28, or 96%) lasting less than 1 year. The mean length of chilling treatments was 56 days (range= 1-254 days) and the mean length of forcing treatments was 39 days (range = 0 to 423 days). Second, the fact that these are controlled environment studies means that the chilling, forcing, and photoperiod treatments imposed by these experiments are decoupled from spatial variation in climate.

   Nonetheless, we should try using the within-subject centering approach in Van de Pol and Wright (also in `http://www.stat.columbia.edu/~gelman/research/unpublished/Bafumi_`

`Gelman_Midwest06.pdf`). I suggest starting with simulated data for 1 species across 3-5 experiments that have distinct, non-overlapping forcing treatments to start. We could then see the effects of using within subject centering vs not using it on estimates.

Note: I am not exactly sure what the reviewer means by spatial versus temporal effects for our study. Some studies may have had a couple of years of replicates (?) and others have multiple sites.

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.*

The reviewer brings up a challenge commonly faced in phenology research– disentangling forcing from chilling– and this questions highlights that we did not adequeatly clarify the methods in our meta-analysis. The chilling estimates that we use in our study are simply the chilling treatments applied in the studies we synthesized. My only idea here is to add clarifying language about our methods, the methods of the authors o the original studies, and add support in the form of references to the many studies that have previously been published using the same methods that we did to estimate chilling. I think this comment stems from a misunderstanding of the data and their experimental nature. We use standard chilling metrics, that have been developed through rigorous experiments and are widely used. We left it to the authors of the original studies to decide on chilling versus forcing; I'm not sure what else we could realistically do.

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.

We thank the reviewer for calling attention to a lack of clarity in our previous manuscript. The length of time that chilling treatments were applied (as well as the temperature of these treatments) varied across experiments: chilling treatments d 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 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, at KNB....

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.

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 thank the reviewer for pointing out a need for greater clarification here. We previously cited only one example of a study using proces-based approach that found stronger effects of chilling than forcing. We have added two additional studies that use process-based approaches and also found stronger effects of chilling. We have also added the Roberts et al paper suggested by the reviewer so that the sentence now says: "Process-based phenological models, however, that explicitly model chilling often find this cue to be most critical (e.g., gauzere2019,laube14a,Heide:2005aa, but see roberts et al 2015)."

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 Roberts et al 2015) may be a useful approach for many question. 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 for two main reasons: 1) we wanted to follow the method of previous studies that have reported declining sensitivities (e.g.

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 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 if the reviewer has specific requests

we can try to address them, though they may be beyond the scope of this manuscript, given the substantial analytical work already included in this manuscript.

Should we expand to 20 years? Ask cat/lizzie

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

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

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 for the kind words.