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.

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.

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 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the within-subject centering approach in Van de Pol and Wright (also in `http://www.stat.columbia.edu/~gelman/research/unpublished/Bafumi_Gelman_Midwest06.pdf`).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 can't think of a reason not to, beyond the work involved, and it would be interesting to know.

   I am not exactly sure what the reviewer means by spatial versus temporal effects, though, and I don't think that using the within subject centering will address this. Some studies have multiple years and others have multiple sites. If we use within-subject centering, it will account for differences among studies but will not separate spatial versus temporal variation. To do this, we would need to add random effects of year and site, I think, and this seems unlikely to be possible to fit, given that study id was not possible to include. It seems worth reaching out to this reviewer since he signed the review and asking for clarification.

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.*

   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 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. Perhaps this is another comment for which it would be beneficial to communication with the reviewer.

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 simiulated 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.*

   add language about the variation in time for chilling treatments (e.g., min-max days). A figure for each study seems a bit overkill but perhaps add a table with this information? or just explain that the data are publicly aailable.

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).*

Should we just try adding these correlations to our main model and see if it affects estimates?
Minor comments
Line 26. Insert "forcing" before temperature.
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 1in Roberts et al.
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.
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, ordinal days. shoulD we expand to 20 years?
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.
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.