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 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 distinction 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. As we say in the abstract (line 11-line 13) we believe we have provided some insight into this discrepancy:

> 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 (i.e. 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 spatial factors can drive results (e.g., days to budburst may vary geographically as well as across years). Separating temporal versus spatial vari-

ation 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.

Given these concerns, however, we have added a map to summarize experimental treatments (chilling, forcing, and photoperiod) spatially, so that readers can visualize that there are no strong spatial biases in treatments (Figure S2). With the figure, we also have provided the Moran's I metric for spatial autocorrelation in the European observations (at the first distance class), which show that spatial structuring of the treatment values, when existing, is more different than would be expected taking into account spatial proximity (i.e. closer spatial units aren't necessarily more similar in their treatments than spatially distant units).

We also 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 Table S5 (Utah units with 36 species) in the Supplemental Materials). 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 standardized predictor (z-scored) model.

	mean	25	75	2.5	97.5
μ_α	30.33	29.15	31.49	26.86	33.87
$\mu_{forcing}$	-0.79	-0.93	-0.66	-1.2	-0.38
$\mu_{photoperiod}$	-0.5	-0.64	-0.36	-0.92	-0.09
$\mu_{chilling}$	-2.69	-3.03	-2.35	-3.68	-1.69
σ_α	10.15	9.22	10.95	7.96	13.06
$\sigma_{forcing}$	1.03	0.89	1.14	0.7	1.45
$\sigma_{photoperiod}$	0.87	0.73	0.98	0.56	1.3
$\sigma_{chilling}$	2.28	2.01	2.51	1.65	3.13
σ_y	15.58	15.4	15.76	15.04	16.16
N_{sp}	36				

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 throughout 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 see changes on line 44-line 45 and line 52 and line 57, line 86 and line 103, line 112, line 145, line 148.

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 auto-correlation. 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). We also edited Figure 1 to highlight that the figure shows experiments.

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 precisely disentangling forcing from chilling is a challenge in

phenology research today. Separating these effects is especially difficult using long-term observational data, where correlations between variables and variability in climate make teasing out effects and attributing findings to exact cues tenuous (e.g., **?**). Our meta-analysis, however, synthesizes controlled experiments that are considered the standard and most widely used method for disentangling these effects (**???**). Indeed, decades of work have relied on controlled environment experiments—synthesized here—to estimate chilling requirements, and equally to robustly estimate photoperiod and forcing requirements (e.g. **??**). Current research at the cellular level is working to more precisely separate chilling from forcing by identifying what exactly underlies endodormancy break (e.g., work on the compound callous, see **??**), but until this research is successful and tested across other species the experiments we synthesize here represent the best and most established method to measure forcing and chilling effects.

Thus, we consider this critique to be not directed at our approach, but at the entire field of phenology that uses these experiments—a field with an over 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 will require fully separating chilling and forcing effects at the physiological-level." And in response to reviewer 3's previous comments we have strengthened and clarified this point throughout the manuscript and highlighted it via Fig. 1 also. For example, line 111 - line 114, we state:

> Thus, while researchers generally define "chilling" and "forcing" treatments based on temperatures in controlled experiments (including in the studies used here, see Fig. 1), fully separating out what plants experience as chilling versus forcing will likely require new methods to measure endo- and ecodormancy *(**?**)*.

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.

We thank the reviewer for bringing attention to the potential for measurement error to affect our results. We did aim to scrape sample size and measurement error when collecting data, but we found many studies did not report these values. Given the reviewer's concerns, we have re-checked all of the studies in our database for which we had not previously scraped sample size or measurement error from the original studies. This did not greatly increase the number of studies with these values, however, making it difficult to build them fully into our models. We should note that we reported the sample size per treatment x species (thus, for example, a study with four treatments on three species and a sample size of 8 would have a total of 96 samples); we apologize for this confusion and have revised our text to clarify this. We also note that our current estimate of measurement error is quite small relative to the magnitude of the responses, (e.g., standard deviation was, on average 12.06% of the response variable for studies for which standard deviation was extracted; note that this value increased sightly with the re-checking all of studies that we did) and thus we expected would not qualitatively impact our findings. We note this in our supplement, which we have updated with our current estimates of measurement error and clarified at what level we report sample size.

We have further worked to address the reviewer's concerns through two additional analyses. We first tested whether there was evidence in our database for the reviewer's concern that studies that report standard error tend to be the ones with larger sample sizes. We looked at the relationship between sample size and error

across these studies, and found that, counter to what might be expected, responses for which no error was reported have a slightly *higher* sample size than in those for which measurement error was reported (t=2.92, df=1851.9, p = 0.004): mean sample size for responses with no error was 13.26 (n=2446 responses), whereas mean sample size for responses with measurement error was 11.53 (n = 364 responses).

Next, we evaluated how measurement error might affect our estimates by simulating response distributions for each response in the budburst data used in our main centered model. The simulated response distributions were generated using the reported response as the mean, and the reported sample size and reported standard deviations, whenever possible (substituting the mean n and and mean standard deviation when not possible). A random response from this simulated distribution of responses (rather than the reported response itself) was input into our main centered model. We did this simulation 20 times and compared model estimates to our main model using the reported responses themselves (Figure below). We found model estimates from the models fit to the simulations with measurement error did not differ dramatically from the estimates using the responses themselves (i.e., the 50th percentile uncertainty intervals overlapped). The only exception to this was an increase in σ_y (as would be expected) with simulated response models. .

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 59), 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: "Across 72 studies examined, we found treatments varied uniquely for each study, but some were more common than others, see Fig. S3: chilling treatments averaged 71.4 days (range: 1-182 days) at an average temperature of 4.4 °C (range: 0-16 °C), forcing treatments averaged 15.7 °C (range: 5 to 32 °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).

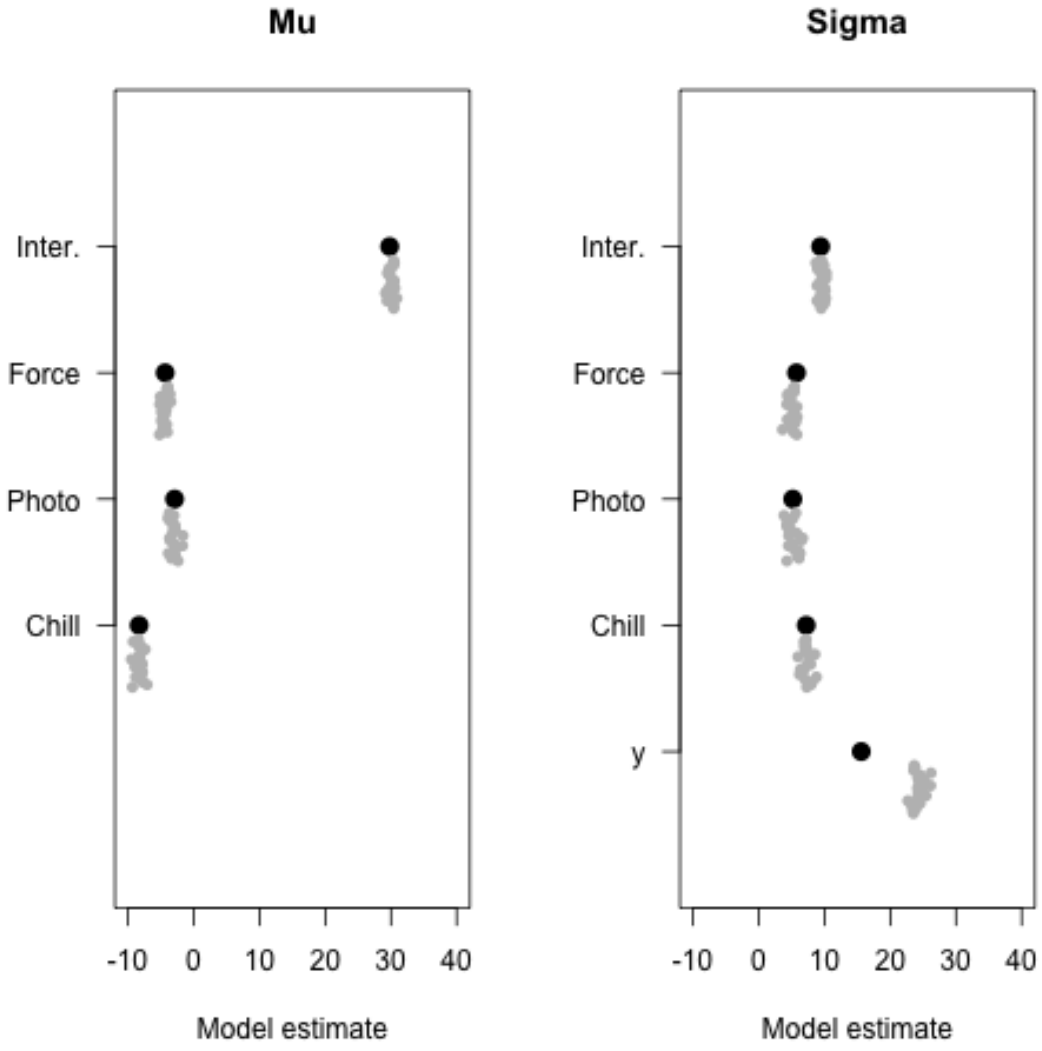

Figure 1: Comparison of main budburst model estimates (black circles) to estimates from models that incorporate estimated measurement error (gray circles) for the main effects in our budburst model:μ_{α} ('Interc.'),$\mu_{forcing}$ ('Force'),$\mu_{photoperiod}$ ('Photo'),and $\mu_{chilling}$ ('Chill'), as well as their associated σ s (σ_{α}, $\sigma_{forcing}$, $\sigma_{photoperiod}$,$\sigma_{chilling}$) and error (σ_{y}). We show means from the posterior distribution for Mus (μ) in the left panel and Sigmas (σ) in the right panel.

The reviewer is correct about our model formulation: we did not originally include these covariances. The use of this type of covariance matrix is generally considered a modelling choice (**?**), unless there are strong covariances, in which case a covariance matrix should definitely be included or the model re-formulated to better address the reasons for such covariances. Our review of our models strongly suggested we were not missing any critical covariances. Based on the reviewer's request, we have added a covariance matrix to our main model to allow random slopes and intercepts to covary. The estimates from this new model are similar to those from our main model (e.g., 50% credible intervals overlap). Please see table below, which compares model estimates from the two models below; we show the correlation matrix for species-level intercepts and slopes below.

Table comparing model estimates from main model (e.g., Figure 2 in the main text, Figure S4 in the Supplemental Materials) to estimates from a model that includes a covariance matrix for species-level intercepts and slopes (i.e., random effects), with 95% uncertainty intervals in parentheses:

	main model	main model, with correlation matrix
μ_α	29.87 (26.43, 33.41)	29.67 (26.08, 33.05)
$\mu_{forcing}$	-4.35 (-6.52, -2.11)	-4.35 (-6.66, -2)
$\mu_{photoperiod}$	-2.92 (-5.38, -0.46)	-3.08 (-5.34, -0.87)
$\mu_{chilling}$	-8.36 (-11.41, -5.28)	-7.62 (-10.89, -4.42)

Correlation matrix for species-level intercepts and slopes (i.e., random effects):

	α_{sp}	$\beta_{forcing_{sp}}$	$\beta_{photoperiod_{sp}}$	$\beta_{chilling_{sp}}$
α_{sp}	1.00	0.05	-0.52	-0.49
$\beta_{forcing_{sp}}$	0.05	1.00	-0.30	-0.00
$\beta_{photoperiod_{sp}}$	-0.52	-0.30	1.00	0.30
$\beta_{chilling_{sp}}$	-0.49	-0.00	0.30	1.00

Minor comments

Line 26. Insert "forcing" before temperature.

We 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 that we refer to experiments, see line 103-line 107:

> 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 177.

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; 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 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 units for 'Windows Open' and 'Window closed' in Table S2 are in ordinal days. We thank the reviewer for pointing out this lack of clarity and now define this in the table legend. 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 six additional versions of our primary statistical model in the Supplemental Materials, Tables S5-S7, 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 59, this figure shows we do have substantial variation across the three factors, as is critical for a robust model). Additionally, to interrogate whether our model estimates may be biased by studies that do not include interactions between multiple cues, we fit our main budburst model to a new subset of experiments: only those that tested at least two interactions between cues. We found that our model estimates did not qualitatively change when fitting the model to this subset of data, and have added a new table to our Supplemental Materials (Table S7).

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 we 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 throughout 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 please see line 44-line 45 and line 52-line 52 and line 57, line 86 and line 103, line 112,line 145, line 148. 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 14, line 181, 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.

References

Basler, D., and C. Körner. 2014. Photoperiod and temperature responses of bud swelling and bud burst in four temperate forest tree species. Tree Physiology 34:377–388.

Caffarra, A., and A. Donnelly. 2011. The ecological significance of phenology in four different tree species: effects of light and temperature on bud burst. International Journal of Biometeorology 55:711–721.

Caffarra, A., A. Donnelly, and I. Chuine. 2011*a*. Modelling the timing of *Betula pubescens* budburst. II. Integrating complex effects of photoperiod into process-based models. Climate Research 46:159–170.

Caffarra, A., A. Donnelly, I. Chuine, and M. B. Jones. 2011*b*. Modelling the timing of *Betula pubescens* bud-burst. I. Temperature and photoperiod: A conceptual model. Climate Research 46:147.

Chuine, I., and J. Regniere. 2017. Process-based models of phenology for plants and animals. Annual Review of Ecology, Evolution, and Systematics 48:159–182.

Fu, Y. H., M. Campioli, M. Van Oijen, G. Deckmyn, and I. A. Janssens. 2012. Bayesian comparison of six different temperature-based budburst models for four temperate tree species. Ecological Modelling 230:92–100.

Fu, Y. S. H., H. F. Zhao, S. L. Piao, M. Peaucelle, S. S. Peng, G. Y. Zhou, P. Ciais, M. T. Huang, A. Menzel, J. P. Uelas, Y. Song, Y. Vitasse, Z. Z. Zeng, and I. A. Janssens. 2015. Declining global warming effects on the phenology of spring leaf unfolding. Nature 526:104–107.

Gauzere, J., C. Lucas, O. Ronce, H. Davi, and I. Chuine. 2019. Sensitivity analysis of tree phenology models reveals increasing sensitivity of their predictions to winter chilling temperature and photoperiod with warming climate. Ecological Modelling 441:108805.

Gelman, A., and J. Hill. 2006. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Harrington, C. A., and P. J. Gould. 2015. Tradeoffs between chilling and forcing in satisfying dormancy requirements for pacific northwest tree species. Frontiers in Plant Science 6:120.

Heide, O., and A. Prestrud. 2005. Low temperature, but not photoperiod, controls growth cessation and dormancy induction and release in apple and pear. Tree Physiology 25:109–114.

Junttila, O., and H. Hänninen. 2012. The minimum temperature for budburst in Betula depends on the state of dormancy. Tree physiology 32:337–345.

Körner, C., and D. Basler. 2010. Phenology under global warming. Science 327:1461–1462.

Laube, J., T. H. Sparks, N. Estrella, J. Höfler, D. P. Ankerst, and A. Menzel. 2014. Chilling outweighs photoperiod in preventing precocious spring development. Global Change Biology 20:170–182.

Rinne, P. L. H., A. Welling, J. Vahala, L. Ripel, R. Ruonala, J. Kangasjarvi, and C. van der Schoot. 2011. Chilling of dormant buds hyperinduces FLOWERING LOCUS T and recruits GA-Inducible 1,3-beta-Glucanases to reopen signal conduits and release dormancy in *Populus*. Plant Cell 23:130–146.

Roberts, A. M., C. Tansey, R. J. Smithers, and A. B. Phillimore. 2015. Predicting a change in the order of spring phenology in temperate forests. Global Change Biology 21:2603–2611.

Rutishauser, T., J. Luterbacher, C. Defila, D. Frank, and H. Wanner. 2008. Swiss spring plant phenology 2007: Extremes, a multi-century perspective, and changes in temperature sensitivity. Geophysical Research Letters 35:L05703.

Sakai, A. K., and W. Larcher. 1987. Frost Survival of Plants. Ecological Studies. Springer-Verlag.

Samish, R. 1954. Dormancy in woody plants. Annual Review of Plant Physiology 5:183–204.

Simmonds, E. G., E. F. Cole, and B. C. Sheldon. 2019. Cue identification in phenology: a case study of the predictive performance of current statistical tools. Journal of Animal Ecology .

van der Schoot, C., L. K. Paul, and P. L. H. Rinne. 2014. The embryonic shoot: a lifeline through winter. Journal of Experimental Botany 65:1699–1712.

Worrall, J., and F. Mergen. 1967. Environmental and genetic control of dormancy in *Picea abies*. Physiologia Plantarum 20:733–745.

Zohner, C. M., B. M. Benito, J. C. Svenning, and S. S. Renner. 2016. Day length unlikely to constrain climate-driven shifts in leaf-out times of northern woody plants. Nature Climate Change 6:1120–1123.