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

Reviewer #1 (Remarks to the Author)

The relative importance of forcing, chilling and photoperiod as cues for budburst is a fascinating one, with clear implications for predicting how species will respond to climate change. Here the authors leverage an exceptional dataset arising from experimental studies using sophisticated statistical analyses and arrive at the surprising conclusion that plants are generally more sensitive to chilling than forcing. I think this study has the potential to make a really valuable contribution that will be of broad interest to readers of this journal. However, I have quite a lot of criticisms/concerns of the study as it stands.

We thank the reviewer for the recognition that the OSPREE dataset is exceptional and that our study can make a valuable contribution to climate change researchers. We have revised the original manuscript substantially to address the reviewer's concerns, as detailed below.

(1) Models: The STAN modelling approach is sophisticated but I think the model is rather incomplete and this could affect the inferences that are reached. For instance why aren't terms included to allow the intercepts and slopes to vary across studies within species?

We completely agree with the reviewer that, ideally, our models would account for both variation in budburst responses among species and among studies. This was an issue we worked extensively on, but we did not clearly present our approach in our initial submission and have revised the text. We now write on Lines 62-67 in the main manuscript:

> Some species are often only represented in one dataset in the OSPREE database, making it impossible to statistically differentiate between species, study, and treatment effects for these taxa. To address this, we combined species found in only one study into "complexes" at the level of genera—such that each taxonomic unit we use in our model occurs across multiple studies (and treatments, see the *The Observed Spring Phenology Responses in Experimental Environments (OSPREE) database* section in the Supplemental Materials for details.)

The main modeling results we report use this version of species/species complexes. Estimates from this model were similar to a model fit to all 203 species (which we report in the supplement, Tables S3 and S4) but we were not clear enough about this in our original draft. In this version have made adjustments throughout the main text (including the abstract) to help clarify this (e.g., line 7, 76-81, 103-104, caption of Fig. 2).

Our aim in building the OSPREE database and using Bayesian approaches that can allow complex models to converge was to cleanly separate out species from study effects. During our data and model development, however, we found studies were generally confounded with species. This is a common problem in this area of research and other meta-analyses have faced similar issues (e.g., **?**, faced the same issue and handled it similarly in consultation with statistician Andrew Gelman). Without strong priors to help differentiate what variation should load onto the study versus species, it is difficult to fit both grouping factors. Indeed, we evaluated models that included studies within species and studies crossed with species but both models were found to be unstable (e.g., estimates varied widely across model runs, chains did not converge).

To address this we developed a model that only included species for which we had multiple treatments across multiple studies. We combined species found in only one study into "complexes" at the level of genera, such that each taxonomic unit we use in our model occurs across multiple studies (and treatments). Thus our taxonomic units of analysis are "species complexes," which are either species represented in > 1 dataset or complexes combining multiple species within a genus that are each singly represented in the dataset. Species represented in

only one dataset with no con-generics in other datasets were excluded from most of our analyses.

Also, I would have thought there is very likely a geographic effect on the effects, and I suggest that you test whether the results are sensitive to inclusion of a spatial random term across which slopes and intercepts vary.

The reviewer makes an excellent point that budburst responses to temperature and photoperiod may vary due to the spatial location of studies or the geographic origin of plant material. In particular, budburst responses are expected to vary by latitude (**????**), and we have added more detailed discussion of this issue to our manuscript in several places. The main text of the manuscript focuses primarily on geographic effects from latitude, because there is strong previous evidence for latitudinal differences in budburst responses, for example via interactions between latitude and photoperiod sensitivity and interactions with chilling responses (**??**). We describe our findings of geographic effects of latitude in the following places in our new version of the manuscript:

- Main manuscript, Lines 79-81: "We fit several additional models, including a model testing latitude effects on cue estimates and one testing effects of chilling study design (see Models in the Supplemental Materials for model equations and other details)."
- Main manuscript, Lines 86-90 "While photoperiod had the smallest effect among the three cues, our results contrast with the extensive literature suggesting photoperiod is an unimportant cue for many species (**??**)—instead we found it was surprisingly large, even when accounting for its interaction with provenance latitude (i.e., the latitude of origin for plant material; see Supplemental Materials for details...)."
- Main manuscript, Lines 134-135: "Further, many additional factors can affect phenological responses, including...latitude (Fig. S3)..."
- Supplement Tables S7, S11, Figure S3, S11.

We agree with the reviewer that there may be other geographic effects beyond latitude that can affect spring budburst responses, such as differences across continents or differences in oceanic versus inland climate (**??**). We attempted to fit a model with continent as a fixed effect in the model, but found that species are confounded with continent in many cases making it difficult to statistically separate effects of continent versus species versus experimental treatments differences.

(2) Meta-analysis: The analysis is described as a meta-analysis, but falls short of being a formal meta-analysis as it seems as though measurement error in the response variable is not incorporated. This should be straightforward to incorporate and I was surprised that it hadn't been given the complexity of the analyses. Also, please report the extent to which the approach followed recommendations made in the PRISMA checklist

This is a good point; we now cite the PRISMA checklist in the Supplement (page 1, first paragraph) and include the full checklist in reference to our study as Appendix 1. Related to this, we have added details on sample size and publication bias to the Supplement (XX). Many of the checklist items are done by our data publication, and we feel this addition greatly strengthens the paper and appreciate the reviewer suggesting it.

We agree with the reviewer that Bayesian approaches offer a straightforward approach for incorporating measurement error, and we would have liked to include this in our model. However, measurement error data were not possible to include for 25 out of 39 experiments included in the OSPREE model with well-represented species. For those studies that do present measurement error, the error was quite small relative to the magnitude of the responses (e.g., standard deviation was, on average 9.9% of the response variable for studies for which standard deviation

was extracted). Thus, it is unlikely that adding measurement error to our analyses would have a large effect on our estimates (**?**).

(3) Methods: The methods seem to be missing from the main ms, and I kept flicking forward to consult a section that does not exist. I thought the Nature letter format does allow a methods section and I found it really to the detriment of the readability of the ms that there wasn't one.

We appreciate the reviewer's concern that methods are not easy to find; this was a concern of multiple reviewers so clearly something we needed to improve. We have worked to now more clearly embed key methods in the text. We now have included a more full model description and the model equation (Lines 66- 74), added a new figure to summarize the experimental methods of studies in our meta-analysis (Figure 1), and added more details on our modeling approaches thoughout the main text with references to extended descriptions in the Supplement.

We provide a separate Methods section with full details of the data and are analyses. This section will be available online, should our manuscript be accepted for publication in *Nature Climate Change*, following the journal's current requirements.

(4) Chilling, forcing and photoperiod: In order for a reader to reach a conclusion about the robustness of the inferences it is vital that the method for quantifying cues is easily understandable. Currently in the main ms it is not (last paragraph of page 3). For instance, we are told the minimum temperature for chilling but not maximum, we are not informed as to when the chilling and forcing periods are and no discussion is given as to how the effect of photoperiod is modeled. It's also unclear in the main ms what a 'standard unit' (I see it is described in the supplement) is and this leaves the reader disconnected with what the analyses are doing. A simple remedy for this would be to include a schematic (as figure 1) that identifies the information used to quantify each cue and relate it to the response. In general the main ms does a very poor job of explaining what was done (the data used, how cues were inferred and vital details about what the models were estimating), instead referring the reader repeatedly to supplementary materials. While the supplementary materials are generally good I still felt disconnected from the data and how the cues were actually quantified. This could be addressed by taking some example datasets and working through in detail how the different metrics were calculated. Without knowing what was done I find it very hard to judge whether the main conclusions are robust.

We thank the reviewer for pointing out that important details on the chilling, forcing, and photoperiod estimates used in our analysis were unclear in the original version. We have added a new schematic figure (the new Figure 1), as the reviewer suggested, which we hope clarifies how the chilling, forcing, and photoperiod estimates were obtained from the original studies for use in our meta-analytical work. We have added substantial new text related to details on chilling, forcing, and photoperiod, including defining a 'standard unit,' to the main text in the following locations:

- On lines 58-63, we describe a summary of how cues were quantified: "The resulting Observed Spring Phenology Responses in Experimental Environments (OSPREE) database includes studies of dormant plant tissue (grown in greenhouses or taken directly from the field) exposed to experimental conditions (**?**) for which we could identify forcing, photoperiod, and chilling treatments quantitatively. Most experiments reported forcing and photoperiod treatments, while chilling occurred mainly in the field, though some studies additionally applied chilling before moving plants into forcing conditions (Fig. 1). "
- Lines 64-68: We now provide the maximum temperature for chilling in the main text: "Because chilling was rarely reported, we calculated an estimate of chilling (both in the field and in experimental conditions), using a common but approximate method (**?**), based on a hypothesis of how chilling accumulates (**?**), with no chilling accumulating below 1.4°C

or above 12.4 °C (throughout the main text we use the term 'chill unit,' see Supplemental Materials, especially Table S3, for details)."

- On Lines 73-74, we have added the model equation for the main budburst model, to make our approach more clear.

- In Lines 84-87, we define 'standard units': "To directly compare the effects of chilling, forcing and photoperiod we fit models using standardized predictor variables (following **?**, , which we refer to as "standard units") and predictors in their natural units (chill units, °C, hours). We further fit several additional models, including a model testing provenance latitude effects, one testing effects of chilling study design, and one testing effects of life-stage (see *Models* in the Supplemental Materials for model equations and other details)." We also have been more consistent in our reporting of estimates from the model using standard units in lines 86-98.

We also provide a full description of the Utah model, including the upper and lower thresholds for chilling, in the Methods and Supplemental Materials, Table S3.

(5) Chilling: I think it's important to know whether the inferences are robust to an alternative model of chilling, e.g., the sequential model that is widely used. From the supplementary materials it is clear that some effort has been made to consider alternatives (chilling portions) but given this analysis underlies the main conclusion of the paper I'd like to see alternative hypotheses considered.

This is an excellent point as our earlier version did not adequately compare results with the Utah model to results with the other chilling model we evaluated, Chill Portions. Now we more clearly compare them in the main text (Lines 116-119), where we write: "We found that applying a different chilling model did not strongly affect our estimates (i.e., 95% uncertainty intervals of estimates for chilling, photoperiod, and forcing overlapped for models using Utah and chill portions with standardized predictors, Table S4)."

We would be happy to add other estimates of chill units beyond the widely used Utah and Dynamic models (e.g., **?**), at the reviewer's request. Parallel versus sequential models (as used in **??**) estimate chilling and forcing effects through a process-based modeling approach and—to our knowledge—are always applied to data with variable temperatures across those periods (e.g., in natural field conditions). Based on this understanding, and in consultation with Isabelle Chuine about applying these models to the controlled environment studies we review here, we do not see a good method for testing the sequential model (or parallel, alternating, etc.) on these data. If the reviewer has specific modeling suggestions or other chilling model he would like applied we would be happy to test them. We also hope that our new Figure 1 may clarify this.

(6) Estimates: It is surprising to see point estimates repeatedly reported throughout the ms without 95% uncertainty intervals, this needs to be rectified. Also at present there is no formal test of whether the chilling response is significantly stronger than the forcing response, though this would be easy to do using the posteriors.

We thank the reviewer for this suggestion and have added 95% uncertainty intervals to all estimates presented in the main text (e.g., Lines 79-91). In figures, we show 50% uncertainty intervals to focus on the most likely estimates (we can change this upon request), and in the supplemental tables we present both 50% and 95% uncertainty intervals.

Statistical artefact with linear regression (Page 5): That application of linear regression to data arising from a growing degree model can lead to biased estimates is a fascinating insight. However, in the supplementary materials it is not clear to me how the temperature sensitivity window

for linear regression (for B. pendula or the simulations) is calculated/defined. How much can the issue of an advancing period of sensitivity be addressed by allowing the sliding window to shift over time? This issue is discussed in Simmonds, E. G., Cole, E. F., & Sheldon, B. C. (2019). Cue identification in phenology: a case study of the predictive performance of current statistical tools. Journal of Animal Ecology.

We thank the reviewer for highlighting the need for additional methodological details, and for this insightful question about the sliding window approach. The window we used was 1 March through 30 April, which we now state in the online Methods, section *Applying our model to Central European data*, (Page 6, paragraph 2).

To address the reviewer's question, we applied the sliding window approach to the same dataset and compared the temperature sensitivities pre-and post- warming. These sensitivities show the same pattern: higher sensitivity pre-warming compared with post-warming that can be explained by differences in mean temperatures across the two time periods (within the selected sliding windows). We have added these analyses to a new section in the "Methods and Supplemental Materials" entitled *Applying the sliding window approach to Central European data*, which we believe strengthen the paper.

Minor comments

Page 2. I suggest changing 'high unexplained variation across' to 'substantial variation among'.

We thank the reviewer for this suggestion and have made the recommended change (now Lines 22-23).

Page 3. All three cues are not generally correlated in longitudinal studies; photoperiod and forcing are, but neither is usually very correlated with chilling.

We thank the reviewer for pointing out that our writing was not clear in this section. In our earlier version of the manuscript, we mentioned correlations between cues but did not specify clearly whether we meant correlations across space or time, nor were we clear about the scale or window at which these cues can be correlated. In this new version of the manuscript, we have clarified our writing, which now says (Lines 40-41):

> Studies attempting to estimate cues using long-term observational data (e.g., **??**) generally fail to overcome the fundamental challenge that cues are strongly correlated in nature (*e.g.*, during the seasonal transition from winter to spring at temperate latitudes, forcing and photoperiod usually increase in step for a given location; mean estimated chilling and spring temperatures can be positively correlated in space).

Though the reviewer states that chilling is often not correlated longitudinally with forcing, we have found that it can be correlated. For example, across the *Betula pendula* PEP725 sites we include in our forecasting analyses (Figure S6), chilling is positively correlated with forcing (i.e., March-April) temperature (r=0.76, 95% confidence interval: 0.75, 0.78).

Page 3. Last sentence of paragraph 2. This is hyperbole. The mean is not expected to shift far beyond historical bounds, though the extremes clearly will.

We understand that the reviewer has concerns about the following phrase from the previous version of our manuscript: "... continued warming pushes climate into environmental regimes far beyond historical bounds." We thank the reviewer for highlighting this phrase, which is a bit vague in its reference to 'environmental regimes' and for which we clearly should have included citations to support. We adjusted the sentence to be more clear and have added references so that it now says (Lines 48-50): "Resolving these discrepancies is critical to accurate predictions

of spring phenology, especially as continued climate change will yield warmer temperatures than have been experienced in at least the last 150 years (**?????**)."

Page 3. Fourth line from bottom. Is interactions the correct term?

We apologize that this was unclear. We have now adjusted this sentence (lines 64-66), which now reads "Our model averages over interactive effects of predictors, including only main effects that we could more robustly estimate given current study designs (see Methods in Supplemental Materials)."

Reviewer #2 (Remarks to the Author):

Spring leaf-out phenology plays a key role in terrestrial carbon and water flux, but the underlying processes are still unclear, especially how the environmental cues, including chilling, photoperiod, and spring warm temperatures, interact and determine the leaf-out processes is still unclear, although most of the phenologist agreed that these three cues are all important. Therefore, quantify the relative importance are valuable and might be important for the phenology modeling and dynamics vegetation models. I carefully read this meta-analysis and found this is an interesting study, but I'm wondering, given the results were reliable, whether the meta-analysis results across experimental studies could reflect the natural plants' response? Or could we rely these experimental results that may inaccurate reflect underlying mechanisms? Because, according to the author (E.M. Wolkovich) previous study, the phenology under warming experiments could not reflect the natural observations (Wolkovich et al, 2012 nature, warming experiments underpredict plant phenological responses to climate change), which might arise from complex interactions among multiple drivers and remediable artefacts in the experiments that result in lower irradiance and drier soils.

We thank the reviewer for pointing out that there are limits to the controlled environment studies synthesized in our meta-analysis. We agree that there are many ways in which experimental conditions differ from observational conditions and we were not clear about this in our previous version. The reviewer's comment also highlights that the previous version did not adequately describe the experimental methods used in controlled environment studies.

To address these comments, we have added a new figure (Figure 1) to show the experimental design of the controlled environment studies in our synthesis. In this new version, we also discuss these issues in the following places:

1. Lines 42-47: "In contrast to observational studies and experimental field warming studies designed to test higher temperatures in natural conditions (**?**), controlled environment experiments can break down correlations between the cues. These experiments, which generally rely on dormant tree cuttings or dormant plants exposed to controlled temperature and light regimes in growth chambers (Fig. 1), have been shown to replicate whole-plant responses in nature (**?**). Such experiments have been conducted for decades (though each experiment generally lasts under a year). "

2. Supplement Pages 5-6, *Applying our model to Central European data*: "We also wished to understand how our findings may apply to conditions more commonly found in nature, where conditions often vary dramatically from those applied in controlled environment experiments. For example, very low amounts of chilling can be applied in experiments compared to the natural chilling found in many temperate areas (Fig. S5). Additionally, chilling temperature and total chilling are more correlated in nature than in experimental conditions (Fig. S5). Further, given the importance of chilling and forcing combined with the fact that seasons do not always warm uniformly with climate change (**??**), we also wished to understand how warming in the winter, spring, or both seasons would shift budburst timing."

We hope that these changes clarify that the controlled environment studies in the present paper differ from the field-based warming experiments in **?**. Controlled environment studies are typically designed to tease apart the role of different cues (chilling, photoperiod, forcing), rather than to replicate current or future natural conditions (the warming studies synthesized in **?** are typically designed to replicate natural conditions).

Furthermore, I'm not convinced that the chilling overweight forcing, and the effect of chilling, photoperiod and forcing might be quantified across more than 200 species based on the various manipulative experiments and MCMC-based Bayesian method, especially considering most of these experimental studies conducted only one year or less than 3 years. The main reasons come from: 1) most of these experimental studies conducted with very different settings, such as using saplings vs. mature tree's cuttings, how the ontogenetic effects play a role or impacts the results? Arbitrary controls in lights/photoperiod length/intensity vs. greenhouse natural light; in addition, for many experimental studies, the temperature and photoperiod were set under extreme climates. I would say this is a response to extreme climate. All of these factors might substantially affect the results.

We agree with the reviewer that there are many factors, not included in our main model, that could explain variation in budburst responses to chilling, photoperiod, and forcing (e.g., ontogeny). In this version we have tried to more strongly indicate this, as well as state that separating chilling from forcing requires more physiological research. For example, in lines 139-143, we state:

> Linking such short-term controlled experiments to natural conditions robustly will require more efforts to understand the complex interactions between chilling, forcing, and photoperiod that we were not able to quantify in this meta-analysis. Most experimental studies do not test for interactions between all three cues (Table S10). Further, many additional factors can affect phenological responses, including ontogeny (Table S9) (**?**), provenance latitude (Fig. S3), and air humidity (**?**).

In addition, in this new version we include a new model to directly test for ontogenetic effects on budburst by adding a predictor of 'life stage' (juvenile vs. adult) to the main budburst model. We found that material from juvenile trees (seedlings or saplings) have similar estimates of days to budburst, on average: 25.3 (95% uncertainty: 15.3-34.7) versus 24.2 (95% uncertainty: 8.9.3-39,8). The effect of stage does appear to vary by species, however: in some, juveniles estimates were earlier, whereas for others, juvenile estimates were later. We have added this model and a table summarizing its results to the supplemental materials.

We believe that there are many additional potential questions and avenues of research to better understand how woody plant phenology responds to different cues, at varying life stages, and in different contexts. Our database and analyses provide a critical step in advancing this work by providing the first comprehensive meta-analysis quantifying responses to chilling, photoperiod, and forcing. It is our hope that our analyses, as well as the freely available OSPREE database, stimulates future experiments and analyses to test additional hypotheses.

2) the interact between chilling, photoperiod and forcing is complicated, and there are still unclear in many important facts. For example, the temperature thresholds of chilling and forcing estimation, and its species-specific values, are largely unknown. For some boreal or alps plants, they may budburst even when air temperature around freezing points, but the temperate trees are still dormancy even air T ¿ 15 degree; the correlations between eco- and endo-dormancy, corresponding the chilling and forcing, whether they are a parallel or a sequential pattern between chilling and forcing? When/how the photoperiod plays its role during the two phase dormancy? Once the endo-dormancy break, continuous chilling accumulation, for example a cold span during spring, is still active? Or entirely depending on the forcing? All these questions are still not figured out;

We completely agree with the reviewer that the interactions between chilling, photoperiod and forcing is complicated. We now discuss this on lines 139-143 (quoted in reply directly above, "Linking such short-term controlled experiments to natural conditions robustly will require more efforts..."). We have also tried to present some of the complexities in our new Figure 1.

We note, however, that our results are strong (e.g., 95% uncertainty intervals for estimates of chilling, photoperiod, and forcing do not overlap zero) and suggest that, despite these complexities, consistent effects emerge. We hope our revised manuscript better balances highlighting the strength of our results with the complexities of both plant physiology and controlled environment experiments.

(3) except chilling, forcing and photoperiod, other cues are also involved with the leaf-out processes, for example air humidity, see Laube et al, 2014 (but recently, Zohner et al, 2019 New phytologist deny this effect) and soil moisture and snow cover. Under manipulative conditions, these effect might be largely ignored as argued in Wolkovich et al, 2012 as well.

As we noted above, we do agree with the reviewer that there are many factors, not included in our main model and not manipulated in most controlled environment experiments, that can affect budburst responses. In this version we have tried to more strongly indicate this, as in Lines 133-135 where we state: "Most experimental studies do not test for interactions between all three cues (Table S10). Further, many additional factors can affect phenological responses, including ontogeny (Table S9, **?**), latitude (Fig. S3), and air humidity (**?**)."

4) species-specific response to chilling, photoperiod and forcing. This has been well reported, for example the pioneer species are opportunistic and photoperiod-insensitive, in contrast the late successional species are sensitive to photoperiod and higher forcing requirements, see the papers, as the authors cited, Korner & Basler 2010;2014; Laube et al, 2014; Zohner et al, 2016 and other studies. Across so large dataset/many species, the mean values, for example chilling effect is 2 times larger than forcing and photoperiod as well as its sensitivity, hold large uncertainty and are no sense.

We completely agree with the reviewer that species-specific differences are important. Indeed our modeling approach is designed to help examine across-species effects but also species-level differences. We now clarify this by including the model equation and writing, on Lines 66-69: "Species are modeled hierarchically, producing estimates of both species-level responses (generally yielding more accurate estimates for well-studied species, such as *Fagus sylvatica* and *Betula pendula*), and the distribution from which they are drawn, yielding an estimate of the overall response across species (see *Methods* in Supplemental Materials)...."

Despite the variation in sensitivity to cues across species described by the reviewer, our models estimate relatively consistent responses in many cases (i.e., overall 95% uncertainty intervals do not encompass 0 for any of the three cues). We now highlight this by including uncertainty intervals for our estimates.

One of the main conclusions is that chilling is over-weight forcing and recent advanced leaf-out is mainly associated with spring warming. However, this is inconsistent with recent study that found the spring phenology did not significantly change during the global warming hiatus, see its figure 1 in Wang et al, 2019 Nature comm, but the spring T is still significantly increase and winter getting colder over the Eurasian (Li, Stevens and Marotzke 2015 GRL)). It seems that increasing chilling and forcing could not explain the dynamics in spring phenology? How to explain this inconsistency?

We thank the reviewer for thinking critically about how our paper may contrast with previous

work, and seeking to understanding the implications of our work. We now cite these two papers. We believe the reviewer is suggesting that, during periods with cooler winters, chilling would presumably increase and this should have dramatic effects on spring phenology, based on our findings that chilling has a strong effect on budburst timing. **?** found that spring phenology did not shift during a period when temperatures were cooler (the global warming hiatus). This, however, does not necessarily conflict with our findings. One finding from our work that we now highlight is that a decrease in temperature is not necessarily equivalent to an increase in chilling, given the thresholds involved in estimating chilling. For example, in this paper we find that warming actually increases chilling in many locations (Figure S6). Thus, the cooler temperatures described by **?** might not be increasing chilling. We discuss this now on lines 155-157, where we write: "In contrast to the common hypothesis that plants experience less chilling with global warming, we found that for many sites total estimated chilling increased with warming (Fig. 4A, C), though this varied with local climate prior to warming (Figs. S6 - S8)."

Minor commons
Line numbers are needed;

We thank the reviewer for this suggestion and have added line numbers.

In methods, the study yielded data from 72 studies across 39 yrs... this is misleading, because for many experimental studies, table S1, the data only for one year, and most less than 3 yrs.

We thank the reviewer for pointing this out and can see how it was unclear. To address this concern, we have removed reference to the 39 years and added that 'each experiment generally lasts under a year' to Line 45.

More description is needed of Bayesian hierarchical model in the main text;

We thank the reviewer for pointing out the need for more detail on our model. We now describe the model in more detail and include the equation for our main budburst model (Lines 61-77).

In the results sections, chilling has greater effect on budburst than forcing?. I would suggest providing the conditions, i.e. under future climate warming, due to the fact that these results come from experimental studies that simulated future warming,

We have added a phrase to clarify that the scope of our results applies to controlled environment studies by saying (Lines 95-97): "Our results, however, suggest that, across 203 species and 72 controlled environment studies, chilling has a greater effect on budburst than forcing."

In the results sections as well, the chilling only occur at warming above 4C? interesting, but does it occur across species? and locations?

We thank the reviewer calling our attention to the need for more detail and nuance here. We have modified the sentence so that it now says (Lines 150-152): "..our results suggest that delays due to decreased chilling only occur at warming above at least 4°C for most sites, though responses vary by species (Fig. 4, S6, S9)."

Zohner, Constantin M., et al. "Rising air humidity during spring does not trigger leaf-out timing in temperate woody plants." New Phytologist (2019). Wang, Xufeng, et al. "No trends in spring and autumn phenology during the global warming hiatus." Nature communications 10.1 (2019): 2389. Li, Chao, Bjorn Stevens, and Jochem Marotzke. "Eurasian winter cooling in the warming hiatus of 1998-2012." Geophysical Research Letters 42.19 (2015): 8131-8139.

Reviewer #3 (Remarks to the Author):

This manuscript addresses the relative importance of the environmental determinants of plant phenology using a meta-analytical approach. Specifically, the authors combine the experimental results of 72 studies and 203 species to estimate the effects of day length, winter chilling, and forcing on spring phenology, using hierarchical Bayesian models. The main finding is that almost all species respond to all three cues, with chilling having the largest, day length the smallest effect. Furthermore, the results suggest that, while all cues are important under experimental conditions, spring forcing will remain the dominant driver of spring phenology over the coming decades. The manuscript is well written and addresses a clear question. However, I have reservations as to the overall importance and validity of these results. That chilling is more important than day length has been shown by previous multi-species studies addressing this (e.g., Laube et al. 2014, Zohner et al. 2016).

We thank the reviewer for his/her review and appreciate s/he found the manuscript well-written and addressing a clear question. We fully agree with the reviewer that other studies have found chilling is more important than daylength, but other studies have stressed the importance of photoperiod effects (**??**, e.g.,). Importantly, our study is one of the very few to provide a comparison of all three cues. Given the current debates in the literature over the relative importance of these cues we feel a meta-analytic approach—as we present here—is critical to advancing research.

Further, our manuscript and related database provide a way to begin to understand why studies may differ: for example, we find that studies using sequential removal of tissue from the field yield different estimates than studies that manipulate chilling in controlled environments (Lines 121-124). We thus believe our manuscript provides firmer evidence for some trends, provides a unique cross-study comparison of all cues, and advances research through new efforts to understand why studies may differ in their findings.

Furthermore, the model output seems to suggest that all three cues (day length, chilling, and forcing) affect phenology in almost all species, leading the authors to conclude that their results contrast with the extensive literature [Zohner et al. 2016, Körner & Basler 2010] suggesting photoperiod is an unimportant cue for many species. [page 4]? Yet, when looking at Table S2, most of the species-level data they use are taken from Zohner et al. (2016) [zohner2016 database]. In fact, 173 (85%) of the 203 species included in this study were already investigated in Zohner et al. (2016). Given that in Zohner et al. (2016), 112 (65%) out of 173 studied species did not react to daylength at all, it is surprising that day length is reported as a relevant, consistent cue across species. This makes me wonder whether their hierarchical Bayesian model is confounded (e.g., giving to much weight to certain species?complexes?) and thus not suitable for exploring the relative importance of the different environmental drivers of spring phenology.*

The reviewer raises a good point, which we believe overlaps in part with Reviewer 1's concerns over clarifying how we handled species and studies in our models. We have now completely overhauled our explanation of this in the main text to clarify that we focus on results from a model of only 36 taxa that are well-represented across studies and designs. We did this to deal with the issue the reviewer is raising—in a meta-analytic approach such as ours we do not want our results to be dominated by any one study. Instead the aim is to present a quantitative synthesis of the literature. In this model, data from **?** make up only 8% of the total data, which may explain in part why we found different results from **?**. To rigorously examine whether **?** affected the estimates in our model, we refit the model, excluding all daa from **?**. Estimates of chilling, forcing, and photoperiod were qualitatively consistent with model estimates from the full dataset. We summarize this model in Table S6 in the Supplemental Materials.

Körner & Basler 2010 clearly is an inadequate reference here, please delete

We have removed this reference from the manuscript in all locations used and replaced it with other references that include more primary research data. Additionally, in our original submission, we cited two papers by Körner & Basler 2010: a 'Perspective' in *Science* (Körner & Basler 2010, *Science* 327, 1461) and a response to a critique of this perspective (Körner & Basler 2010, *Science* 329, 278)—we now have removed and replaced both references from our manuscript.

Apart from that, I take issue with the estimation of the importance of forcing and the attempt to estimate the relative importance of day length, chilling, and forcing. First, I don't see how the effect of forcing can be disentangled from the effects of chilling. This would require knowledge on which temperature ranges are adequate to satisfy chilling and forcing requirements. Yet, as correctly stated in the Supplementary information (page 2), current models of chilling are hypotheses and likely to be inaccurate for many species. Similarly, the effective temperature ranges to fulfill forcing requirements are not known. As such, when comparing the relative importance of winter chilling versus spring warming both factors are likely to be confounded. Also, if a study uses two different forcing temperatures that both lie within the range of optimal forcing conditions, one would see no effect between the treatments and the authors would thus infer that forcing didn't affect phenology, when in fact, forcing has a huge effect, not detected by the study design. Given these considerations, I don't think that a multivariate model, such as the one presented in this study, can adequately disentangle the relative importance of the three main phenological cues.

The reviewer makes an astute and important point that it is impossible to fully disentangle effects of all three cues in many experiments, given the way that these treatments are applied. We completely agree with the reviewer's point and have added a new figure and adjusted our text throughout the manuscript (abstract, main text, conclusions) to address this. We added a new Figure 1 in this version which lays out some of these challenges explicitly for readers. We have now clarified that models of chilling are hypotheses (see Lines 112-116, quoted below) and tried to strengthen this point thoughout the manuscript. As the reviewer points out, there is a great need for specifics-specific information about the effective temperature ranges for fulfilling chilling requirements. One motivation for this paper is to highlight the need for additional work on this and other aspects of spring phenology. We have also added a new sentence to the abstract, where we write, "Further progress to improve budburst forecasts under future climate change will require fully separating chilling and forcing effects at the physiological-level."

In regards to the concern that experiments may find no effect of forcing if different forcing temperatures fall within optimal conditions, we agree that such non-linearieties are important. Indeed, including such non-linearities has been shown to improve models (**?**). Based on our understanding, these non-linearities often occur at very low temperatures ($< 10°C$) or higher temperatures, generally above 40°C, depending on the plant and other environmental factors (**?**). Most of our studies are not at these extremes [give 25-75% values of force and maybe min and max] and fall more within where we expect temperature responses are more linear. Further, we are not aware of any studies in OSPREE that varied forcing (given photo and chill held constant) and did not see an advance. Given these reasons, we did not fit a non-linear model to forcing. Based on these comments though we did attempt to fit a sigmoidal model to forcing. Much like our attempts to fit a sigmoidal model to chilling (see section in Supplement on *Modeling limitations based on experimental designs*), we found these models explained less variation than our linear models and generally did not change other parameter estimates. If the reviewer has specific non-linear models of forcing s/he would like us to explore, we would be happy to.

p.2: What do you mean by 'we included only studies with at least 49.5% budburst?' This is not correct for most of the studies included in your OSPREE dataset. E.g., Heide (1993) and Zohner et al. (2016) defined budburst as the date when 1-3 buds on a twig had opened. Please

clarify.

This is a good point and differences in the definition of budburst are exactly why we have this *Defining budburst* section. We have now clarified our meaning. What now have adjusted text on lines xx-xx to read as follows, "." This issue was only for studies that reported percentage of budburst as a function of time (this was a small number of studies); to make these studies comparable with most other studies, we needed to estimate one day of budburst for each treatment from the % budburst by time figures. In general, we extracted a date close to 90% for most studies of this type, but our code allowed studies to be included as long as the highest % budburst was at least 49.5%. We realize this is confusing and hope we have clarified this point now.

p. 3: Total chilling ranged from -1304 to 4724 Utah units? The Utah model allows for negative chilling units? What's the biological justification for that?

We agree with the reviewer that the Utah model, as well as other chilling models, are often non-intuitive. We now describe this model, as well as the other chilling model we tested, Chill portions, on Lines 114-122 in the main manuscript and in greater detail in the 'Estimating Chilling' section of the methods. We used the Utah model to report chilling because that allowed us to include the greatest amount of studies from the OSPREE database (i.e., because many studies reported their chilling treatments in only this metric), and because it is widely used around the world.

The biological justification for negative chilling accumulation, as described by the developer of the original model (**?**) is that controlled environment studies have shown that exposure to high temperatures "nullfied the effect of the low temp" (**??**). Under these, and numerous other, controlled experiments, plant material is exposed to different temperature treatments and different combinations of treatments to identify the ideal range of chilling temperatures (i.e., those that contribute most toward 'rest completion,' or endo-dormancy break, Fig. 1; for Utah units, this range is usually set to 2.4-9.1 °C, as shown in our new Table S3). **?** bases the Utah model on his own controlled environment studies, as well as those conducted by **?**, all of which were conducted on peaches.

We have also tried to make clear that in the main text that all chilling models are hypotheses on Lines 112-116, where we state: "Models of how species accumulate chilling are poorly developed for forest trees, with few relevant tests evaluating the particular temperatures at which species do or do not accumulate chilling. Instead, researchers generally rely on models developed for perennial fruit trees, *i.e.*, Utah (**?**) and chill portions (**?**), both of which were developed for peach species. These models are themselves hypotheses for how chilling may accumulate and produce dormancy release, but are likely to be inaccurate for many species (**?**)."

p. 4: Latitude model: This model doesn't make sense to me. What is the latitude you refer to here? The location where the experiment took place? You refer to provenance locations, I doubt these are available for most of the studies, especially the ones conducted in botanical gardens or other collections.

We thank the reviewer for highlighting the need for greater clarity about our terminology and the model. We now clarify our meaning and have adjusted the model to include only studies that give locations of the populations from which they sampled. In the initial submission, we did not adequately clarify what was meant by latitude: we now refer to this as 'provenance latitude,' throughout out the manuscript (see below for specific changes by line number).

We also revisited all papers used in this model to check that they refer to the latitude of origin of plant material used in the experiment, which the reviewer is correct in noting is the

traditional definition of provenance. Most studies did give clear provenance, and many were actually studies of multiple provenances (e.g., **?????**). We found, however, that several studies did not include provenance location, were crop species (e.g., commercial blueberries, rather than wild blueberries), or were of unclear provenance (e.g., **?**), which we had intended to remove from this analysis. We refit the latitude model to this slightly smaller group of species (a less than 15% reduction in data) and found that estimates of chilling, forcing, photoperiod, and latitude remain qualitatively consistent with our previous model. We now present this more restrictive model in our manuscript and thank the reviewer for catching this discrepancy.

1. Lines 93-95 in the main text we now state: "While photoperiod had the smallest effect among the three cues, our results contrast with the extensive literature suggesting photoperiod is an unimportant cue for many species (**??**). Instead we found it was surprisingly large, even when accounting for its interaction with provenance latitude (i.e., the latitude of origin of plant material; see Supplemental Materials for details, especially Figs. S3, S11, Table S5)."

2. Supplemental Materials, Page 4: in the description of the Latitude Model, we now say: "we examined the effect of including provenance latitude in a model similar to our main one, but designed to estimate effects of provenance latitude. This model estimated the effects of each phenological cue (chilling, forcing, photoperiod) on days to budburst (as in the main model), in addition to the effect of provenance latitude (i..e., the latitude of origin of plant material used in the experiment) and the interaction of photoperiod and provenance latitude. We include this interaction because photoperiod effects are expected to vary by latitude..." and in the next paragraph we now say "...then subsetted the species and species complexes to include only those that had multiple provenance locations across different latitudes."

3. Supplemental Materials, Caption for Table S5: We have replaced 'latitude' with 'provenance latitude' so that the caption now reads: "Using a model with Utah chilling units and testing the effects of provenance latitude plus the interaction between provenance latitude and photoperiod results in slightly muted effects...."

4. Supplemental Materials, Caption for Figure S3: We have replaced 'latitude' with 'provenance latitude' so that the title of the caption now reads: "Estimates for effects of chilling exceeded estimates for forcing, photoperiod, provenance latitude, and the interaction between latitude and photoperiod, for most species...."

Figures: Figs. 2 and 3, showing a 3-dimensional illustration of the interplay between winter chilling and spring warming, are very hard to read. I would prefer a simpler illustration.

We appreciate the Reviewer's perspective, and struggled ourselves with differing preferences among co-authors and reviewers. We have modified the former Figure 2 (now Figure 3) so that it includes both a simpler 2-dimensional version (at a mean level of forcing) and the previous 3-dimensional version (which shows all possible combinations of chilling and forcing). For Figure 3 (now Figure 4) we refer to 2-dimensional versions available in the Supplement. We have shown both simpler 2-dimensional versions of these figures and the more complex 3-dimensional versions to a number of scientists and found opinions to be split on preferences for 2d versus 3d versions. We hope that showing both versions in Figure 2 (now Figure 3) facilitates understanding and interpretation of the 3-dimensional illustrations.