October 17, 2019

Dear Dr. Findlay:

Please consider our paper, entitled "Winter temperatures dominate spring phenological responses to warming" for publication as a "Letter" in *Nature Climate Change*. This manuscript is a revised version of an earlier submission (). We include a point-by-point response to reviewer comments.

The timing of spring phenology (e.g., budburst, leafout) in woody plants is critical to plant fitness, shapes plant and animal communities, and affects wide-ranging ecosystem services from crop productivity to carbon sequestration. Advances in budburst are some of the most reported—and tangible—biological impacts of climate change, garnering great research and public interest.

Recent warming has ignited debate over the fundamental drivers that determine spring phenology, with far-reaching implications for which environmental cues will dominate future trends (*1-4*). Although most temperate species show responses to spring warming (forcing), the prevalence and relative strengths of responses to chilling (associated with cool winter temperatures) and photoperiod (daylength) could slow or stall advances in spring phenology with continued warming. Indeed, recent work suggests chilling or daylength cues may underlie observed declines in the 'temperature sensitivity' of leafout in Europe (*5-6*).

We address this controversy by synthesizing four decades of controlled environment experiments to estimate overall chilling, forcing and photoperiod responses across 72 experiments and 203 species from around the globe. We find most species respond strongly to all three cues, with chilling being the strongest cue—nearly four times greater than forcing. Yet, when we applied our results to areas with reported declining phenological responses to warming (*5*), we find few sites where chilling or daylength cues would constrain leafout advances under current or near-term warming. Instead, we suggest observed declines may be due in part to a statistical artifact: we show that temperature sensitivities (measured in days per °C) calculated without correcting for warmer daily temperatures will always predict a decline with warming—even with no change in chilling, forcing or photoperiod responses. Our results thus resolve several major debates in plant phenology by showing that most species respond to all cues strongly in experimental conditions, but forcing appears to determine responses to recent warming.

Upon acceptance for publication, the database will be freely available at KNB (*7*; currently meta-data are there); the full database is available to reviewers and editors upon request. This work is a meta-analysis, so data have been previously published; however, the synthesis of these data and the tables, figures, models, and materials presented in this manuscript have not been previously published nor are they under consideration for publication elsewhere.

Sincerely,

Ailene Ettinger
Visiting Researcher, Arnold Arboretum of Harvard University

References mentioned in cover letter

1. Körner, C., & Basler, D. 2010. Warming, photoperiods, and tree phenology response. *Science*, 329: 278-278.

2. Chuine, I., Morin, X., & Bugmann, H. 2010. Warming, photoperiods, and tree phenology. *Science*, 329: 277-278.

3. Zohner, C. M., et al. 2016. Day length unlikely to constrain climate-driven shifts in leaf-out times of northern woody plants. *Nature Climate Change*, 6: 1120.

4. Flynn, D. F. B., & Wolkovich, E. M. 2018. Temperature and photoperiod drive spring phenology across all species in a temperate forest community. *New Phytologist*, 219: 1353-1362.

5. Fu, Y. H., et al. 2015. Declining global warming effects on the phenology of spring leaf unfolding." Nature 526: 104.

6. Richardson, A.D., et al. 2018. Ecosystem warming extends vegetation activity but heightens vulnerability to cold temperatures. *Nature*, 560: 368.

7. Wolkovich, E., et al. 2019. Observed Spring Phenology Responses in Experimental Environments (OSPREE). Knowledge Network for Biocomplexity. urn:uuid:b2ab2746-b830-436b-a7a9-01b3ef3558e4.

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

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.

(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? 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.

(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.

(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.

(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 modelled. 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.

(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.

(6) Estimates: It is surprising to see point estimates repeatedly reported throughout the ms without 95% credible 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.

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.

Minor comments

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

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

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.

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

(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? 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.

(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.

(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.

(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 modelled. 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.

(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.

(6) Estimates: It is surprising to see point estimates repeatedly reported throughout the ms without 95% credible 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.

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.

Minor comments

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

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

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.

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

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.

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. 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; 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. 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, Basler and Korner 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.

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?

Minor commons Line numbers are needed; 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. More description is needed of Bayesian hierarchical model in the main text; 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, In the results sections as well, the chilling only occur at warming above 4C? interesting, but does it occur across species? and locations?

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 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.

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. 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; 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. 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, Basler and Korner 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.

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?

Minor commons Line numbers are needed; 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. More description is needed of Bayesian hierarchical model in the main text; 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, In the results sections as well, the chilling only occur at warming above 4C? interesting, but does it occur across species? and locations?

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.

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).

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, Krner 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) [Zohner16 database]. In fact, 173 (85importance of the different environmental drivers of spring phenology. *[Krner Basler 2010 clearly is an inadequate reference here, please delete]

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.

Supplementary material

p.2: What do you mean by ?we included only studies with at least 49.5

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?

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.

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

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.

Supplementary material

p.2: What do you mean by ?we included only studies with at least 49.5

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?

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.

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.