Below we give all original reviewer comments and our responses to each review under the 'Editor' or 'Referee' subheadings. Comments are given in *italics* while our responses are in regular print.

Editor:

The reviewers were overall quite positive about this submission and, highlighted the innovative perspective and potential biases that researchers may have when interpreting data and patterns of plant phenological changes with climate change. At the same time, I was really pleased to see that all the reviewers had some interesting suggestions for improving the impact and generality of the review for plant scientists, particularly with respect to the incorporation of more text focused on recent advances in molecular mechanisms and the potential for integrative research to address research gaps in our current approaches to plant phenology and climate change.

We too were pleased by these reviews, which were very helpful in improving the paper.

We have reviewed the molecular literature and now include it throughout the manuscript (while working to not greatly extend the length). We now mention molecular pathways early on (line 33-line 36), include a new paragraph to explain them (in the section 'Why observational studies are not enough to robustly estimate cues', line 90-line 102), and added another new paragraph (line 287-line 301) in the section 'Incorporating our understanding of physiology into forecasts of phenology' explaining how insights from molecular pathways could improve future studies. Additional small changes related to this were made throughout (detailed below), and new paragraphs were added in the supporting information that review the molecular basis of chilling (we can move these to the main text if requested, but were concerned about length).

Referee: 1:

This is a review with an important message—we need better data to forecast the consequences of climate change on spring phenology. The authors correctly argue that associations of phenology with changes in external conditions have been oversimplified and have not sufficiently considered the complication of multiple covarying cues. The literature to date on tree phenology is reviewed and the authors conclude that three factors chilling and forcing (temperature cues) and photoperiod together determine the transition from dormancy to growth each year. In general, the temperature and photoperiod cues are abstracted into average exposure over a temporal window and used to model outcomes of a warming climate. Interactions of these different cues have been studied in controlled environment experiments, which have suggested that the temperature accumulated over the fall and winter interacts with the photoperiod experienced in the spring to determine budburst. The authors argue that predicting the influence of climate change on phenology requires understanding how these interactions will change and that will require much more data and different modelling techniques. They collate the published work on phenology analysis in woody species and conclude insufficient temperature variables have been examined.

They argue that detailed controlled environment studies are required with the most useful future experiments for forecasting designed to identify threshold effects, optimal temperatures/photoperiods, and non-linearities of interacting cues. They also argue that the results of these experiments will still be limited in their utility for prediction without a deeper understanding of how major phenological cues act physiologically, how the different temperature cues influence dormancy, and how they interact with photoperiod mechanisms.

The review has some important messages and improving predictions of phenology as the climate warms is essential. However, the review misses out on recent advances in molecular understanding driven through research in reference organisms but in many cases now translated to tree species (work for example of Bhalerao et al and Strauss et al) - where the role and action of specific proteins is understood. This molecular understanding helps crystallize how different aspects of temperature and photoperiod cues are integrated at specific regulatory nodes in the gene network. This is the aspect that the authors conclude is missing in the modelling. Future controlled environment experiments in specific woody species could then include measurement of expression of specific RNAs–eg FT copies. This would significantly help accurate modelling of phenology in different temperature x photoperiod environments.

We agree with the reviewer and thank them for their suggestions. We have now added more on molecular pathways throughout, including adding references from the suggested authors (e.g., Maurya & Bhalerao, 2017; Azeez *et al.*, 2021). As mentioned in our reply to the editor, we added new paragraphs early on (line 90-line 102) in the section 'Why observational studies are not enough to robustly estimate cues,' and later (line 287-line 301), in the section 'Incorporating our understanding of physiology into forecasts of phenology,' as well as mentions in the abstract and other sections (e.g., line 119-line 121). Finally, we have also worked to differentiate interactions estimated statistically from the developmental pathways that create them (line 15, line 67, line 69), including: "Such studies have been critical both in our understanding of interactions from a more simple statistical perspective—where we estimate how the effect of one cue depends on another cue—and from a deeper understanding of the developmental pathways that produce observed interactions" (line 33-line 36).

The other recent conceptual breakthrough in seasonal timing has been to understand that average temperature is not monitored over a time interval but both acute and long-term changes in temperature are used to infer seasonal progression. The different temperature inputs are integrated at regulatory nodes in the gene network. This understanding significantly changes the inputs into the modelling. Thus, the missing information for accurate phenology prediction in woody plants is identification of these key regulatory nodes (building from the knowledge in the other reference plants).

Agreed, we now mention this early on in a new paragraph (line 90-line 102), where we write:

> Molecular studies show that plants may monitor temperatures—especially cool temperatures— over long time periods, while also responding to short-term fluctuations, and new work on budbreak suggests epigenetic control, with multiple regulatory layers (Azeez

et al., 2021).

I would thus recommend addition of a section in this review highlighting recent molecular advances in the field of phenology (focusing on woody plants but including some of the important concepts made in other species). An important outcome of this review, which would improve predictions going forward, would be a synthesis of ecological, physiological and molecular understanding of phenology.

We agree and have added several paragraphs relating to this now, as detailed above.

Referee: 2

Changes in the timing of spring events such as budburst are widely considered to be an important fingerprint of climate change, and may have important implications for plant population persistence, species distributions, and ecosystem function. Predicting future phenological responses is therefore critical. Unfortunately, we still lack critical information for most species about how plants integrate multiple seasonal environmental cues to control phenological transitions. This review is therefore very timely. The authors argue persuasively that experiments in controlled environments are necessary to decompose the interacting effects of cues such as chilling accumulation, heat sums, and photoperiod, which are often correlated in nature. They review and synthesize data from a comprehensive survey of experimental studies of plant phenology, revealing large gaps in the information needed to understand and predict responses to interacting environmental cues. I agree with the authors that comprehensive experimental studies of multiple cues are badly needed. However, I think that the paper would benefit from a more thoughtful discussion of what specific information is needed about underlying mechanisms to improve predictive performance in the face of novel climate conditions. The authors have an opportunity to set an important agenda for future integrative phenological research, and I would like to see them develop a clearer, more integrative vision for the path forward.

We appreciate the reviewer's positive feedback, and also their concerns. We have made changes throughout the manuscript to address these concerns, as we detail with line numbers and quoted text below in response to each concern. We believe the manuscript is much improved and thank the reviewer for their help.

Phenological prediction is done on different scales by different groups of researchers to answer different questions. At one end of the scale, the goal is to predict regional patterns of green-up from coarse climatic variables like mean seasonal temperatures. The problem is that simple statistical models can't account for interactions among temporally correlated environmental cues, as discussed in the introduction to this paper. At the other end, the goal is to predict the phenology of specific species, genotypes, or cultivars using process-based models of development, integrating detailed information about dynamic responses to multiple daily environmental cues. The challenge here is getting accurate estimates of model parameters quantifying underlying biology (e.g. chilling response functions). The fundamental challenge is to design experiments to elucidate the actual physiology underlying process-based models, and then scale up that knowledge

to large scale predictions of phenology from climate variables. We need more than power to test for statistical interactions. We need to understand the underlying biology of chilling, forcing, and photoperiod response and their dynamics in seasonal environments. This is where properly designed experiments can make a real contribution.

We agree and have tried to add this to the current manuscript. In particular, see line 312-line 331, which include one previous paragraph expanded to two paragraphs to highlight these needs.

Here are some specific comments:

Line 57-58: effects of warming on chilling can either delay or accelerate spring phenology depending on the shape and position of the chilling accumulation function (Fig. S2 is a key figure that probably belongs in the main text).

Yes, we completely agree. We have added a note on this, see line 63. We elected not to include Fig. S2 in the main text as revisions have already increased our word length, however, we would be happy to move this figure to the main text if the editor requests.

Line 60-61: Even if temperature is the only forcing factor, linear models may not be appropriate. The relationship of developmental rate with temperature may be linear at lower temperatures, but like most thermal performance curves, it may max out at an optimum temperature and decline at higher temperatures.

We agree, and have worked to clarify this point throughout.

Lines 73-81: Crop models of phenology often use photothermal units of development, with daily developmental progression happening only during daylight hours. So in this sense daylength can be an integral cue, independent of critical daylength cues (which are separate parameters)

The point that these processes are averaged over longer temporal windows in long term data in many studies is a major limitation. If cue responses are non-linear, it might be better to look at the distribution of daily temperatures over that window, rather than the mean. For example, if the chilling effectiveness function is known, it could be used to estimate accumulation of chilling units which is more biologically informative than mean temperature. Of course, this requires finer-grained temperature data which may be harder to come by- but this should be a high priority and that is worth mentioning.

These are good points. We tried adding this complexity to this particular part of the manuscript, but found it greatly lengthened this section, so we have not addressed it here directly. However, other changes throughout the manuscript we believe make this complexity clearer. For example, we now address long- versus short-term pathways for temperature in molecular studies in a new paragraph on molecular insights (line 90-line 102):

Combined with molecular approaches controlled environment studies have led to our current models of the pathways that control phenological events, and create interactions. The most well developed of these, the flowering time pathway for *Arabidopsis thaliana* (reviewed in Satake *et al.*, 2022), links closely to chilling, photoperiod and forcing: flowering time is mediated by cool temperatures—through a vernalization pathway, daylength—through circadian clock control, and together these pathways limit or promote development driven by warm temperatures. Similar multi-controlled pathways appear across other species and events (Maurya & Bhalerao, 2017; Satake *et al.*, 2022), including in peach flowering and *Populus* budbreak (Azeez *et al.*, 2021). Across species, these pathways almost always have far greater complexity than generally considered in observational studies of spring events. Molecular studies show that plants may monitor temperatures—especially cool temperatures—over long time periods, while also responding to short-term fluctuations, and new work on budbreak suggests epigenetic control, with multiple regulatory layers (Azeez *et al.*, 2021).

Line 85: Limited understanding of chilling is a critical challenge. The discussion in the supplement is really important and it would be great if some of that could make it into the main text.

We completely agree and have worked to address this by:

(1) Extending the text in the supporting information to include a new paragraph, as well as several new lines, on the complexity of chilling and new information from molecular studies.
(2) Referencing this section of the supporting information earlier in the manuscript (line 63).
(3) Adding an extended discussion of this to the main text section, 'Incorporating our understanding of physiology into forecasts of phenology' (line 287-line 301)

Our understanding of the molecular pathways, however, for dormancy and dormancy-break is very poor (especially when compared to our understanding of flowering time pathways, Azeez *et al.*, 2021), with most recent breakthroughs focused far downstream of the signal and receptor pathways. Thus far, research has been on a very limited number of species (Singh *et al.*, 2017; Rinne *et al.*, 2018), making extrapolation to other species difficult.

Our inchoate understanding of endo- and ecodormancy pathways is likely the greatest current impediment to improved models of spring phenology. Without markers or other methods to clearly identify the start or end of chilling—or whether there even is a clear start or end (as opposed to potential parallel models of chilling and forcing, Harrington & Gould, 2015)—experiments will continue to operate with a set of latent chilling and forcing treatments (i.e. the underlying values of chilling and forcing that the plants actually experience) some distance from the 'chilling' and 'forcing' treatments researchers discuss in their papers. In such a murky experimental world, we may be unlikely to see rapid advances from new experiments.

Lines 98-102, Figure 2. Here's a more biological way to think about this interaction, inspired by process-based models of flowering that I'm familiar with. You can think of forcing in terms of developmental rate (developmental progress toward budburst per GDD), which is modified by

chilling and photoperiod. For example, I would guess that if you plot Figure 2 as 1/GDD vs chilling units, you will see a more-or-less straight line, which describes the acceleration of development (or release of repression) by accumulation of chilling, up to an asymptote at the point where the chilling requirement is saturated and additional chilling won't be effective. Similarly, developmental rate might be accelerated by more inductive daylengths. To visualize how this works, see Figure 1 in Chew et al. 2012 (New Phytologist, 194:654-665)

Thanks, we added this, including a reference to Chew *et al.* (2012) (see line 119-line 121).

Fig. 2 also illustrates a very important point: cue responses may vary among populations, especially on large scales across different climates or latitudes. This is an important challenge for phenological prediction and I would like to see that point acknowledged in this paper.

Agreed, we have changed one of the titles of the later sections to 'Building population- and species-rich predictions' and now discuss population variation there (line 345-line 356, these lines are quoted in response to the last comment by this reviewer, see below).

Lines 163-164. Statistical tests of interactions are great, but even more important is to design experiments that provide insight into the causes of those interactions. I would argue that the first experimental priority should be detailed characterization of response functions of developmental rate to chilling, ambient temperature, and daylength for the species of interest, which could then inform selection of treatments for tests of interactions.

We thank the reviewer for this important point and we now address it later in the manuscript, where have changed text (line 241, and line 246-line 249):

> Identifying threshold effects and optimal temperatures or photoperiods generally translates into response function designs, which require many different levels of a single cue—making such experiments difficult to cross with other cues. Yet, where a threshold or optimum occurs is likely dependent on the level of other cues (Stearns & Olson, 1958; Flynn & Wolkovich, 2018), making studies of interactive cues critical for robust forecasting. Ideally, studies testing for non-linearities due to interactive cues would be fully-crossed (i.e., every combination of levels is present in treatments), resulting in a robust response surface design. Such designs, however, can quickly require a large number of controlled environments. Researchers may be able to reduce such numbers by a modified response surface design that aims to characterize the response function of one focal cue across several relevant levels of the other cue.

In particular, as the authors point out in section 2 of the supplement, chilling responses are very poorly understood. There is a need for controlled experiments measuring developmental rate across a range of chilling temperatures and durations to characterize chilling effectiveness functions and saturation thresholds for estimation of chilling accumulation in different climate scenarios (Figure S2 shows how key this is for prediction, but experimentally parameterized models are needed). Of particular interest is whether and when plants "forget" past chilling exposure

at higher temperatures, as in the Utah model. It would be great if this sort of experiment could be combined with expression studies to understand the mechanisms and dynamics of endodormancy cycles.

We completely agree and have worked to address this; see our response to line 85 comments.

Similarly, we clearly need experimental measurements of developmental rate across a range of temperatures (preferably in plants that have already been subjected to saturating chilling accumulation to release endodormancy completely). If development toward budburst is slower at supraoptimal temperatures, this could potentially produce nonlinear responses to warming.

Agreed, see our response to line 85 comments, especially the new text in the supporting information.

Lines 219-220 Where a threshold or optimum occurs is likely dependent on levels of other cues.. Possibly, but do you need a full factorial to test this, or can it be first approached by accurately characterizing the response function of interest across a few relevant levels of other cues?

Good point, we now address this (line 246-line 249, and changes throughout the paragraph, which we quote more fully above):

> Researchers may be able to reduce such numbers by a modified response surface design that aims to characterize the response function of one focal cue across several relevant levels of the other cue.

Line 231 and section below: This whole section is absolutely key. We really need to understand the dynamics and markers of endodormancy and ecodormancy to inform predictions. If we don't know when the transition between these states occurs, or whether it is abrupt or graded, we don't know when to expect different seasonal cues to be important. For example, must a chilling requirement be saturated before the GDD cue can begin to accumulate, or can both cues accumulate at the same time? Can you say a little more about how you can experimentally address these questions?

Yes. We have extended this section (see quoted paragraphs above).

I would also like to see a few sentences discussing the challenges presented by genetic variation in cue responses across the landscape

This is a great suggestion, and we now mention this point early in the manuscript (line 46), and towards the end (line 345-line 356):

> Given the efforts and data involved in models for a single population or ecotype, building up to multi-population and multi-species predictions may appear daunting, but such models are crucial for accurate forecasts that can apply to diverse regions

and large-scale vegetation models. Our current understanding of pathways for phenological events suggest strong population differences (Wilczek *et al.*, 2009; Tanino *et al.*, 2010), while population-level studies of spring budbreak suggest less local adaptation than budset (Aitken *et al.*, 2008), chilling is often strongly differentiated by population (Junttila & Hänninen, 2012). Thus, models that assume constant cues across a species range, could make grossly inaccurate predictions.

At the species-level, molecular studies show temperature pathways diverge often across species (on contrast, photoperiodic control appears highly conserved, Satake *et al.*, 2022) suggesting a critical need for more molecular data across species....

Referee: 3

This review highlights the importance of experimental studies for accurately predicting changes in plant phenology with climate change, while also describing current information gaps and future challenges for experimental studies. The authors synthesize findings from phenological experimental studies and present a convincing argument that experimental data, in addition to observational data, will be crucial to predict phenological changes outside historic ranges of variability. The challenges addressed here are very relevant for plant ecologists and phenologists, and the figures do a good job of illustrating gaps in phenological cue space that could be addressed in future studies. I have only minor suggestions and comments detailed below.

My only larger concern regards the database search. Why were the terms 'flowering', 'forcing' or 'chilling' not included as topics in the search? Additionally, the search does not seem like it would pick up any studies that just examined forcing temperatures and did not mention photoperiod or dormancy. I understand that the authors wanted to focus on studies that looked at interacting cues, but there may be many greenhouse studies on annual plants (that do not necessarily experience dormancy) that are missed by this search.

We thank the reviewer for their positive comments and helpful edits. Our database was developed for the original aim of our research—to examine how cues together affect spring phenology shifts with warming—and thus we developed search terms for this purpose. In the end, however, we picked up many single cue studies (we address this below in greater detail in response to the related specific question). In early searches (as developed the search terms), we found few studies on flowering and thus focused on budburst and leafout. We expect it would take a number of additional specific searches to include studies of annual plant flowering from controlled temperature greenhouses (in our experience with this review, few greenhouse studies report consistent greenhouse temperatures, and more commonly give a range that would be hard to use for our purposes).

Summary

In the last sentence of the Summary, I suggest adding a comma after a new generation of lab experiments.

In addressing concerns of referees 1 and 2, we have adjusted this sentence such that a comma is no longer needed (we believe).

Main Text

l. 7. Do these citations show that spring phenology has stopped advancing everywhere? I would perhaps change this to say advances have appeared to slow or even reverse in some plant species (and or regions, etc.).

Good point. We actually have found this 'decline' in every long-term dataset with warming, but the citations are specific to two regions so we have adjusted the sentence

l. 23 What are the values the authors are referring to here? The date of a phenological event?

We have struggled with how to refer to cues and levels of cues, which is what we meant here. We have changed to 'cue levels.'

40. Should this be written as how treatment effects compare to shifts in cues?

We meant treatments, but we see now it is not very clear and so have added '(i.e., levels of cues).'

l. 83-87. I'm not sure that the argument for photoperiod entirely makes sense here. Photoperiod will not be directly altered by climate change, and if a plant's physiology is altered by climate change such that it experiences different photoperiods during different physiological or phenological stages, then wouldn't that indicate that photoperiod is not as important of a phenological cue? Perhaps if this is reworded, I will understand this argument better.

We thank the reviewer for highlighting that our previous language might be unclear to some readers, and we see the reviewer's point. We agree that—based on the reasoning of the reviewer (which is our basic meaning)—one could interpret that "that photoperiod is not as important of a phenological cue," however we still believe it is a relevant cue. Previous research tells us that, for most species, chilling, forcing, and photoperiod together determine budburst and other phenology (Ettinger *et al.*, 2021). Because these cues can sometimes compensate or interact with one another, it is not necessarily correct to assume that because shifts in phenology have already occurred, photoperiod is not an important cue or that it will not be an important cue in the future (or for some species, such as *Fagus sylvatica*, where it is very strong). We have tried to reword this sentence to clarify without adding too much text. We now write (line 107-line 109), "Additionally, the relevant photoperiod a plant experiences during important physiological transitions (e.g., when dormancy ends) may change dramatically with warming."

l. 155. This trend

Thanks. And apologies for the typo, it's fixed now.

l. 209-210. I would add spring before phenology in the first sentence of this section. The authors make the case mainly for spring phenology, but don't really touch on how experiments could help increase accuracy of fall or mid-summer phenological events. I'd suggest either explicitly including any information on controlled experiments for later events (I'm guessing there aren't many) or focusing this review on spring phenological events rather than trying to generalize to predictions of phenology'.

We have added 'spring.'

l. 239. Did the systematic literature review encompass studies that only looked at forcing temperatures? It appears from the database search terms that studies would only be included if they involved at least two cues. But it would be interesting to know if there are other 'forcing only' type experiments. These would obviously also involve interactions with photoperiod even if all plants are exposed to ambient photoperiod while under different temperature regimes (e.g. in greenhouses), but such studies might not explicitly include the topic 'photoperiod' or 'daylength', and thus not be picked up by the current search terms?

We agree that it would be interesting to look at 'forcing only' type experiments and we understand the reviewers' concerns. Our original aim was to examine interactive cues, and thus we developed search terms specific for this aim. However, we found many single-cue studies; in the section 'How studies and their experimental treatments vary globally' we review that most studies manipulated only one cue. Further, our search terms even picked up studies that used controlled environment studies to look solely at provenance effects (that is, without different treatments for chilling, forcing or photoperiod, see Figure S3). We did conduct preliminary searches to narrow in on search terms that addressed our aim, while gave us a reasonable number of papers to review, and believe we have a fairly representative sample of these studies—across interactive and single cue experiments.

Fig. 1. The lines in the panels could be slightly more distinct colors to help distinguish them from each other.

Good point, we have produced a new figure with more distinct colors.

References

Aitken, S.N., Yeaman, S., Holliday, J.A., Wang, T.L. & Curtis-McLane, S. (2008) Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications* **1**, 95–111.

Azeez, A., Zhao, Y.C., Singh, R.K., Yordanov, Y.S., Dash, M., Miskolczi, P., Stojkovic, K., Strauss, S.H., Bhalerao, R.P. & Busov, V.B. (2021) Early bud-break 1 and early bud-break 3 control resumption of poplar growth after winter dormancy. *Nature Communications* **12**.

Chew, Y.H., Wilczek, A.M., Williams, M., Welch, S.M., Schmitt, J. & Halliday, K.J. (2012) An augmented Arabidopsis phenology model reveals seasonal temperature control of flowering time. *New Phytologist* **194**, 654–665.

Ettinger, A.K., Buonaiuto, D.M., Chamberlain, C.J., Morales-Castilla, I. & Wolkovich, E.M. (2021) Spatial and temporal shifts in photoperiod with climate change. *New Phytologist* **230**, 462–474.

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.

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

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

Maurya, J.P. & Bhalerao, R.P. (2017) Photoperiod- and temperature-mediated control of growth cessation and dormancy in trees: a molecular perspective. *Annals of Botany* **120**, 351–360.

Rinne, P.L.H., Paul, L.K. & van der Schoot, C. (2018) Decoupling photo- and thermoperiod by projected climate change perturbs bud development, dormancy establishment and vernalization in the model tree populus. *Bmc Plant Biology* **18**.

Satake, A., Nagahama, A. & Sasaki, E. (2022) A cross-scale approach to unravel the molecular basis of plant phenology in temperate and tropical climates. *New Phytologist* **233**, 2340–2353.

Singh, R.K., Svystun, T., AlDahmash, B., Jönsson, A.M. & Bhalerao, R.P. (2017) Photoperiod- and temperature-mediated control of phenology in trees–a molecular perspective. *New Phytologist* **213**, 511–524.

Stearns, F. & Olson, J. (1958) Interactions of photoperiod and temperature affecting seed germination in *Tsuga canadensis*. *American Journal of Botany* **45**, 53–58.

Tanino, K.K., Kalcsits, L., Silim, S., Kendall, E. & Gray, G.R. (2010) Temperature-driven plasticity in growth cessation and dormancy development in deciduous woody plants: a working hypothesis suggesting how molecular and cellular function is affected by temperature during dormancy induction. *Plant Molecular Biology* **73**, 49–65.

Wilczek, A.M., Roe, J.L., Knapp, M.C., Cooper, M.D., Lopez-Gallego, C., Martin, L.J., Muir, C.D., Sim, S., Walker, A., Anderson, J., Egan, J.F., Moyers, B.T., Petipas, R., Giakountis, A., Charbit, E., Coupland, G., Welch, S.M. & Schmitt, J. (2009) Effects of genetic perturbation on seasonal life history plasticity. *Science* **323**, 930–934.