

Phenological response to climatic change depends on spring warming velocity

Corresponding Author: Dr Robert Guralnick

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Decision Letter:

**** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors ****

Dear Dr Guralnick,

Your manuscript titled "Phenological response to climatic change depends on seasonal warming velocity and species traits" has now been seen by 3 reviewers, and we include their comments at the end of this message. They find your work of interest, but some important points are raised. We are interested in the possibility of publishing your study in Communications Earth & Environment, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication. In particular, they raise that your approach to GDD requirements needs to be more rigorously justified, considering differential responses of different plant species. Furthermore, certain claims in the conclusions should be toned down, and key terms need to be clarified in-text.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response that takes into account the points raised. Please highlight all changes in the manuscript text file.

Please submit your point-by-point responses as a separate file, distinct from your cover letter where you can add responses to the Editors' comments that you do not want to be made available to the reviewers. Word files are preferred.

Important: The response to reviewers must not include any figures, tables or graphs. If you wish to respond to the reviewer reports with additional data in one of these formats, please add them to the main article or Supplementary Information, and refer to them in the rebuttal. Due to current technical limitations, any figures, tables, or graphs embedded in your rebuttal will not be included in the peer review file, if published.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your revised manuscript, point-by-point response to the referees' comments (which should be in a separate document to any cover letter), a tracked-changes version of the manuscript (as a PDF file) and the completed checklist:

Link Redacted

**** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first ****

We hope to receive your revised paper within six weeks; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we may close your file. In this event, we will still be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Earth & Environment or published elsewhere in the meantime.

Please do not hesitate to contact us if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Alice Drinkwater, PhD
Associate Editor
Communications Earth & Environment
@CommsEarth

EDITORIAL POLICIES AND FORMATTING

We ask that you ensure your manuscript complies with our editorial policies. Please ensure that the following formatting requirements are met, and any checklist relevant to your research is completed and uploaded as a Related Manuscript file type with the revised article.

Editorial Policy: [Policy requirements](https://www.nature.com/documents/nr-editorial-policy-checklist.pdf) (Download the link to your computer as a PDF.)

For Manuscripts that fall into the following fields:

- Behavioural and social science
- Ecological, evolutionary & environmental sciences
- Life sciences

An updated and completed version of our Reporting Summary must be uploaded with the revised manuscript

You can download the form here:

<https://www.nature.com/documents/nr-reporting-summary.zip>

Furthermore, please align your manuscript with our format requirements, which are summarized on the following checklist: [Communications Earth & Environment formatting checklist](https://www.nature.com/documents/commsj-phys-style-formatting-checklist-article.pdf)

and also in our style and formatting guide [Communications Earth & Environment formatting guide](https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf).

*** DATA: Communications Earth & Environment endorses the principles of the Enabling FAIR data project (<http://www.copdess.org/enabling-fair-data-project/>). We ask authors to make the data that support their conclusions available in permanent, publically accessible data repositories. (Please contact the editor if you are unable to make your data available).

All Communications Earth & Environment manuscripts must include a section titled "Data Availability" at the end of the Methods section or main text (if no Methods). More information on this policy, is available at <http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf>.

In particular, the Data availability statement should include:

- Unique identifiers (such as DOIs and hyperlinks for datasets in public repositories)
- Accession codes where appropriate
- If applicable, a statement regarding data available with restrictions
- If a dataset has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Data Availability Statement.

DATA SOURCES: All new data associated with the paper should be placed in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific, community-recognized repositories, where possible and a list of recommended repositories is provided at <http://www.nature.com/sdata/policies/repositories>.

If a community resource is unavailable, data can be submitted to generalist repositories such as [figshare](https://figshare.com/) or [Dryad Digital Repository](http://datadryad.org/). Please provide a unique identifier for the data (for example a DOI or a permanent URL) in the data availability statement, if possible. If the repository does not provide identifiers, we encourage authors to supply the search terms that will return the data. For data that have been obtained from publically available sources, please provide a URL and the specific data product name in the data availability statement. Data with a DOI should be further cited in the methods reference section.

Please refer to our data policies at <http://www.nature.com/authors/policies/availability.html>.

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The goal of the paper, to provide a “generalizable model of species phenology response under climate change” is admirable. I agree that such a framework should directly tie into physiological processes that are known drivers of phenological response yet remain simple enough to not require extensive tuning across species, regions or time periods – this would be an amazing contribution. However, I do not think the authors achieve this goal, and several aspects of the study are not sufficiently explained or justified, leading to confusion about methodological choices and interpretation of the results. For instance, no explanation is provided for why data are compared across 5-degree latitudinal bins (which are very large and do not necessarily correlate with climatic conditions in the US); it is unclear what data-driven approach was used to generate temperature windows across latitude (i.e., phenological bins); it is unclear why data were aggregated specifically to (rather coarse) 55km by 55km equal area grid cells; the rationale behind keeping 5 or less observations per grid cell/day/year to reduce bias does not quite make sense – if this is an issue why not just keep 1 per day/year/cell; the rationale for assigning 3 month bins staggered by 1 month across 5-degree latitudinal bins for the spring season is not clear – is there any ecological support for these choices? There were several terms (e.g., ‘phenometrics’) that were not clearly defined, and figures and their legends/captions lacked details. Also, while GDD requirements are considered to be consistent across species in the study, given that previous studies have shown that there is immense intraspecific variation in phenology and how it responds to climatic change, I find this to be unlikely. Similarly, though the authors claim that the specific amount of accumulated heat required to trigger a life cycle event has been quantified for many species, we really do not know what the threshold is for the vast majority of plant species, making the accurate inference of GDD unfeasible. I doubt good information exists even for all 18 species included in the study (which are never named in the manuscript). Further, much of the manuscript was rather confusingly worded, making it difficult to interpret. The study has potential and looking at spring velocity is interesting, but I think presenting the results as an extensible and generalizable model framework incorporating for predicting phenological response to warming is greatly overselling the study. The scope of the study, while broad, definitely does not come close to representing phenological variation globally and across the (plant) Tree of Life (as stated in the manuscript).

More specific comments:

Figure 1: Why is the intercept of the slow velocity line always higher? There can be instances where warming happens more slowly and starts at a lower temperature. The statement “thermal requirements are met earlier in the season in locations with slow velocities than those with faster velocities” may apply to the illustrated example but should not apply to all situations. Also, figure axes are not aligned.

Figure 2: Unclear what ‘phenometric densities’ are. Phenometrics are never defined in the manuscript, and the legend seems to indicate what is being shown are the number of phenometrics measured in each cell but it is not clear how these were counted. Is it the number of different measures (e.g., flowering onset, leaf out) or species, or individuals measured?

Figure 3: Please depict the underlying data – effect plots can be misleading without them. There does not seem to be a pronounced difference between historical and contemporary data besides the intercept contrary to what is in the text.

Figure 4: Please depict the underlying data – effect plots can be misleading without them. How was temperature velocity binned? The early, middle, and late flowering/leafing species seem to show remarkable similar trends within each temporal category. Would it not make more sense to plot different phenological events separately?

Line 64: confusingly worded – given the previous sentence it seems that “a unifying model characterizing generalities and underlying processes determining the mode and tempo of phenological change across space and time” applies to at least plants and animals and the range of seasonal behaviour they engage in – is this feasible? Surely the processes that determine the start of hibernation, for instance, are different from those that determine the timing of spore release in mosses?

Line 158: confusingly worded – are only leaf onset metrics divided into historic/contemporary or does this apply to flowering as well?

Line 161: Figure 2 does not seem to illustrate what is presented in the text.

Line 165: confusingly worded sentence “Linear model testing whether period (historical or contemporary) predicted velocity revealed..”

Line 170: multiple models or just one? Only one seems to be described.

Line 175: what is a “decisively better result”? Please refer to metrics/values.

Line 257: as the authors note in the manuscript, spring temperature variability and velocity are strongly correlated. If so, using velocity to model phenology should not differ much from using variability.

Line 255: The question of sensitivity across latitude has remains unsettled because it is not latitude itself that influences phenology, it is other variables that sometimes covary with latitude. As the authors note on line 257, only ~30% of the variation in velocity can be explained by latitude. Thus the crude latitudinal binning used in the study is not optimal. Perhaps binning by climate could be an alternative?

Line 278: There are many species that leaf out after flowering (e.g., cherry blossoms) – perhaps the authors mean that this is unusual in their dataset?

Line 281: Perhaps I’m missing something here, but this does not seem to sufficiently justify why leaf out and flowering timing were not allowed independent responses to the fixed effects.

Line 293: What data driven approach?

Line 296: The references examine a very small number of woody species and are not sufficient grounds to expect species to have relatively similar GDD requirements across space more broadly.

Line 313: Please elaborate what sorts of differences and biases were present in the data.

Line 324: How was it determined that “The model correctly predicts onset timing for leaves and flowers, capturing variation in sensitivity and responsiveness more fully than other analyses across species and phenological events”? I did not see any

model evaluation or testing on non-training data.

Line 358: Provide the cleaned files please.

Line 376: Research grade observations are not truly vetted – they are just observations that are agreed upon by a certain number of people who may or may not possess botanical knowledge.

Line 384: Unless the entire plant was captured in the image, it is not appropriate to infer absences. For instance, the absence of flowers on a part of one branch on a tree does not guarantee that there were not flowers on other branches or other parts of the same branch.

Line 385: Absences and NAs should not be treated the same.

Line 391: Why 55km by 55km equal area grid cells? Both the phenological data and climate data are higher resolution. Why this particular cell size? This also does not align with the 50km buffers mentioned later.

Line 394: Why keep 5 or less? Why not just consistently keep one? Also do you mean 5 per species? 5 per phenological event? 5 regardless?

Line 398: This sounds like later season data were excluded to fit the authors expectations and make the models fit better?

Line 401: Please elaborate – not clear how these metrics were calculated.

Line 403: which phenometrics?

Line 407: If the observations are in the same 55km grid, why is this 50 km buffering necessary?

Line 428: What is the rationale for this binning strategy?

Line 452: Given that spring was defined as 3 month bins, does this mean the velocity of spring warming was calculated as the slope of only 3 points? This does not seem very optimal.

Line 465: What were these binnings based on? Is there ecological support for these very specific 'early/mid/late' periods? The 'mid' period is only 10 days and plenty of species flower/leaf out across more than that. There can easily be a week's difference in the onset of flowering in the spring across 1 degree of latitude, let alone 5.

Line 475: I thought spring was defined differently across latitudinal bins – why are they consolidated here?

Reviewer #3 (Remarks to the Author):

In this manuscript the authors present a general model framework to model plant species' phenological responses to climate change. They consider the spring temperature, the velocity of warming in spring and the species-dependent thermal requirements (GDD). To test their model, they used phenological data from 18 species collected in the eastern US over 9 years in both the 19th and 21st centuries. The authors found that temperature sensitivity is higher in regions with low temperature velocity and that the greatest realized change in phenological onset was found in northern areas where warming rates have been fastest. They assume that their generalised and simple framework can be very helpful for predicting phenological sensitivity and responsiveness across the Tree of Life.

Major comments:

1. I really appreciate your approach of presenting a simple framework that takes spring warming into account. However, in my opinion some of the statements and conclusions in this paper are a too strongly worded and should be toned down. This begins with the wording of the title. You have only considered one trait (temporal phenological niche coupled with GDD) in the analyses. Furthermore, you often mention that data spans a period of 175 years. In my opinion that is not the case. Two periods from different "climate contexts" were analysed, each lasting circa 9 years. In lines 272-289 you show that their model is applicable to understory and canopy plants and discuss the generality of the results for early flowering species. I am not convinced by these statements, as only two herbaceous species were included in the data set! In fact, I would rather advise you to leave out these species and focus on trees and shrubs and then compare these two growth-forms with each other. This is particularly relevant if, for example, the results from Horbach et al. 2023

(<https://doi.org/10.1016/j.flora.2023.152218>) are taken into consideration. At the end of the conclusions you write that the model presented has wide applicability across the Tree of Life. I would expect this too, but it would have to be tested first.

2. The layout of this journal determines that the section on material and methods is at the end of the paper. In many parts of the "Result" section, you provide brief insights into the methods used to help the reader to understand the results. But especially when it comes to the key results of this study the LMM model with all explanatory variables (Table 1, Figure 3&4), it is difficult to follow all the statements. This is due to the high complexity of the model with many interactions and factors, as well as the relatively small visible effect sizes (Fig. 2). I wonder whether the model can be simplified even further by restricting the design of the model to interactions with comprehensible hypotheses. The rule can apply here: Less is sometimes more. In addition, adding paragraphs and more explanations or definitions for the names of the different explanatory variables to the result section could make it easier to follow the findings.

3. The following point certainly does not require much text to be written, but I consider it to be important. Throughout the manuscript, the GDD concept is described as universally valid. I mostly agree with this statement. However, the discussion

should briefly address how the species-specific choice of base temperature would affect the relationships shown. It is known that phenological events should be predicted in different species with different thresholds for the base temperature.

Minor comments:

- L1: I would revise the title of this paper. You have only considered one trait "temporal niche" in the analyses. The title suggests that several traits were analysed.
- L75: What do you mean with "temporal record"? Is this the period over which the phenological data was collected?
- L75: "Quality" is rather a vague term. You should be more precise here.
- L106: "(the rate of warmth accumulation)" I think this can be deleted. It should be clear to readers from the previous explanations what is meant by this.
- L109: I really like the graphical framework! It might be useful to reference what is meant by "cool" and "warm" context. An example like before with the lake for velocity would be helpful.
- L135: I would already mention here briefly that the species were trees, shrubs and herbaceous species.
- L137-138: "[...], who established what is believed [...]". I would suggest to move this sentence to the M&M.
- L141: Add a space: "model, "
- L143: Do you mean the "cold" and "warm" context mentioned before with "spring temperature context". If yes please be consistent in the wording.
- L158/Fig. 2: The term "phenometric" has not yet been explained. Since M&M are at the end of this paper (Journal style), I think it should be made clear in the results or in the caption of the figure what is meant by this.
- L161: The most intensive sampling years can't be seen in Fig. 2.
- L166: You should write "0.28 °C". In addition you may change this to °Kelvin.
- L169: Before (e.g. L165) you wrote " $r=0.564$ ". Please be consistent with the spacing (again in L263)
- L174: Please be consistent with the number of digits shown in Table 1. The values of the deviation should have the same number of digits as the effect size. Furthermore, in my opinion, tenths of digits are sufficient for the effect size, as we are talking about days in a year that probably cannot be measured more precisely anyway.
- L182: Please be consistent with the number of digits shown in Supplemental Table 2
- L187: "middle" or "mid" instead of "midde"
- L291: In Line 214 you added no ":" after the sub header but a line break. Same issue again in L311.
- L312: Why are you using the term "replicated "experiments""? I would rather use "space-for-time-substation" (different locations across a spatial scale and "time-for-time" (data from the past and from 21st century)
- L327: I would say that it CAN be applied to insects. But that would also have to be tested.
- L337-339: To what extent would this be possible? In particular, could the effects of extreme weather events such as warm temperatures in late winter be explained using your framework?
- L410-411: It might be interesting for the reader to see what original number of observations you have reduced by filtering.
- L460-461: If I have understood your approach correctly, you are using the timing of flowering or leaf out as a proxy for whether little or a lot of GDD is needed? If this is the case, it should be made clear here and, in the introduction, that the "trait" considered is the magnitude of the GDD.
- L480: Why were the continuous predictors scaled? Does this method refer to lines 317-319? If so, it should be mentioned there again.
- L708: How were the three groups defined for temperature velocity? I couldn't find anything about this in the M&M.

Reviewer #4 (Remarks to the Author):

I really enjoyed reading this paper, it was compelling to include temperature velocity of change as a factor that would differ across biomes yet allow a generalizable variable in a model. However, I felt the paper needed more discussion on the utilization of species' thermal requirements and how to appropriately estimate them for this sort of model.

I understand that the GDD requirement is based on whether you accumulate fewer or more GDD by the time you start some phenological events, but is this taken from 1 year of data to estimate if a plant was early or late and assigning a GDD threshold to it? What level of multi-season GDD estimates do we need to average? Have studies done enough thermal performance curves using GDD to estimate GDD thresholds (T_{opt} or T_{min})? Finally, climate-relevant measures of phenology (e.g., the number of GDD accumulated by time of flowering) can shift with climatic conditions (e.g., Wadgyamar et al. 2018 New Phytologist) – can these threshold account for shifts in the thresholds themselves?

Minor comments:

- L42: Unclear what mode means here.
- L66: There are multiple global meta-analyses and large-scale studies trying to determine generalities regarding phenological shifts since the 2000s; these statements generally need to be toned down.
- L296: Populations often differentiate in their thermal performance, suggesting different GDD requirements across space. Could be worth trying latitudinal-band-specific GDD requirements?
- L406: I appreciate the similar scaling of the contemporary and historical datasets here.
- L415: Somewhere in this section please explain how the GDD requirements were determined for each species.
- Fig. 3: Suggest indicating which slopes are statistically different from each other.
- Fig. 4: It'd be nice to include a Fig. 5 with a similar structure, but velocity on the x-axis.

** Visit Nature Research's author and referees' website at www.nature.com/authors for information about policies, services and author benefits**

Communications Earth & Environment is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the [Platform Support Helpdesk](http://platformsupport.nature.com/).

Version 1:

Decision Letter:

** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors **

Dear Dr Guralnick,

Your manuscript titled "Phenological response to climatic change depends on spring warming velocity" has now been seen by our reviewers, whose comments appear below. In light of their advice we are delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Earth & Environment.

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers, which concern clarity of figures and generalised language. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table".

*****Please take care to match our formatting and policy requirements. We will check revised manuscript and return manuscripts that do not comply. Such requests will lead to delays. *****

Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed at the end of the Editorial Requests Table; the list of required files is also available at <https://www.nature.com/documents/commsj-file-checklist.pdf>.

OPEN ACCESS:

Communications Earth & Environment is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit <https://www.nature.com/commsenv/open-access>

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to submit the above items:

Link Redacted

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Alice Drinkwater, PhD
Associate Editor
Communications Earth & Environment
@CommsEarth

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

We appreciate the effort that went into addressing the issues raised. The manuscript is clearer than before, provides more details regarding the methods employed, and claims and expectations have been toned down to be more realistic.

However, there was still confusion around the figures and results (our lab discussed the manuscript as a group, in confidence). Several found fig1 confusing and non-intuitive. Perhaps changing the y-axis to GDD would make what the authors are trying to convey clearer. Likewise we found fig2 to not be very informative and difficult to reconcile with the analyses that were done on these data. Not using a log scale and clarifying whether the phenometric count is species specific or not would help. Also it would be useful to parse the map out by the phenological event recorded, as one would not directly compare leaf out in one region with flowering in another, and it is important to be able to see which areas are consistently sampled and which are not. The rugs in fig3 are a welcome addition, though the raw data would still be preferable. However the rugs for contemporary and historic data overlap making it hard to see how the data are distributed. Perhaps it would be possible to have two separate rows of rugs? Additionally, while we understand the differences depicted in fig3 are significant, the visualization does not illustrate this effectively. We suggest it might be better to just show these results as a table of coefficients. Another alternative could be to combine the information 3 panels into 1 using different line styles. This would at least make the comparison between early-middle-late categories easier visually. We also noted that capitalization and axis labels were not consistent and figure dimensions seemed odd, but recognize that some of this may be due to the conversion process. Finally, we again strongly suggest the inclusion of the list of species examined in the main text. It would greatly facilitate our interpretation of the results.

Reviewer #3 (Remarks to the Author):

I appreciate the great effort you have made to revise this manuscript. Below are some suggestions for improvement, which I think should be taken into consideration:

L80/L94: I know it is the style of the journal not to give the author's name in the text. But from my point of view it looks a bit strange to write 'e.g.' followed by the superscript numbers. I think you can also omit "e.g.", as it should be clear that quotes given are often only examples.

L159: Space after "vs."

L176: Add a "." after Fig.

L190: Delete space after "= 0.34" to be consistent.

Reviewer #4 (Remarks to the Author):

I appreciate the authors' time spent clarifying the language and decision making in this manuscript. However, I am still concerned about the estimates of GDD "thresholds" (versus plastic shifts in phenology) and think it's still trying to generalize much more than an ENA dataset on just plants is able to.

Estimating the mean GDD accumulated at time of leaf-out (and other phenophases): How is this a set threshold? GDD accumulated at time of life history event will likely be a plastic trait. To get a threshold, I think you would need a critical minimum GDD required for annual growth to begin (in the case of leaf-out). Since these species do not have TPCs for every one (instead, records that allow estimates of mean GDD accumulation at time of phenophase), the framing should be adjusted.

Generalizing language: I highly recommend removing ectotherms, and potentially, as another reviewer mentioned, focusing on the trees and shrubs that make up the majority of the dataset. There are also still a lot of arguments for universal models when this is really testing effects of rates of warming in a much narrower spatial, climate, and species context.

** Visit Nature Research's author and referees' website at <http://www.nature.com/authors> for information about policies, services and author benefits**

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Below please find responses to reviewers comments, with our responses in blue to each reviewer comment.

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The goal of the paper, to provide a “generalizable model of species phenology response under climate change” is admirable. I agree that such a framework should directly tie into physiological processes that are known drivers of phenological response yet remain simple enough to not require extensive tuning across species, regions or time periods – this would be an amazing contribution. However, I do not think the authors achieve this goal, and several aspects of the study are not sufficiently explained or justified, leading to confusion about methodological choices and interpretation of the results.

Thank you for your thorough and constructive review of our manuscript. Please see our detailed responses below. We have worked to address concerns, clarify justification of methods and generally toned down the interpretation of our results.

For instance, no explanation is provided for why data are compared across 5-degree latitudinal bins (which are very large and do not necessarily correlate with climatic conditions in the US); it is unclear what data-driven approach was used to generate temperature windows across latitude (i.e., phenological bins);

We agree this choice needs further clarification. Our study area is focused in the Eastern USA across a broad latitudinal gradient of approximately 15 degrees. This is a region where Hopkin’s bioclimatic law was developed and generally holds: 1 degree of latitudinal change corresponds to 4 or 5 days later onset. Indeed, our empirical data and analyses support Hopkin’s Law. Given Hopkin’s law, one can use that to choose latitudinal windows to allow examination of climate proximal to spring onset timing. 5 degree blocks shift the onset dates nearly a month, from January-April in the South, to February-May in the Mid-Atlantic region, to March-June in the North. This choice means we are using climatic data proximal to the onset event in the region, and in doing so we remove any correlation between latitude and temperature. Another way to think about this: If we had just looked at climatic drivers of onset in the Southern part of the Eastern USA, Jan-April makes sense as a “spring climate” window, given onset timing, while we would never use that window in the North, where March-June makes more sense given Hopkin’s Law. This choice also aligns well with resolution of climatic data available from 1850-1859. We have adjusted the methods to add a clearer rationale. The whole paragraph starting with the sentence “To avoid this issue, we used a sliding window approach to define spring time periods across latitudes used for

determining climate context.” has been rewritten to clarify these decisions and their rationale.

it is unclear why data were aggregated specifically to (rather coarse) 55km by 55km equal area grid cells; the rationale behind keeping 5 or less observations per grid cell/day/year to reduce bias does not quite make sense – if this is an issue why not just keep 1 per day/year/cell; the rationale for assigning 3 month bins staggered by 1 month across 5-degree latitudinal bins for the spring season is not clear – is there any ecological support for these choices?

One unique aspect of this work is combining phenometric data from across different sources. The approach for these phenometric calculations differ depending on the source and each source has reported and unreported sets of uncertainties. This is one reason we are most interested in relative rather than absolute shifts in timing changes. Because the sources are different, the phenometrics themselves have different spatial resolutions. The only phenometrics aggregated to 55km by 55km are those from iNaturalist. The 1850-1859 phenometrics are reported for locations that are typically municipalities, which of course also have associated (and unreported) spatial uncertainty. The National Phenology Network phenometrics are generated from data at the site level.

The choice of this resolution for iNaturalist follows best practices for generating phenometrics from incidental data outlined in a paper where the lead author on this paper was involved in as senior author (Belitz et al., 2023 - <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2435.14173>). That work clarifies that there are always tradeoffs between spatial resolution and data density and that there are minimum data density requirements for fitting phenometrics successfully. The choice of 55km by 55km for the iNaturalist data basically was the best optimization for meeting minimum data requirements while limiting spatial extents across all species used in this study. We have added text to discuss this optimization in the appropriate section of the Methods e.g. “Our choice of aggregation was based on trade-offs between data densities needed to fit phenometrics and spatial resolution, and which reflect best practices for fitting these types of models⁴⁷”. Key here is that all climate data used with iNaturalist phenology data are aggregated at the same resolution as the phenometrics. We therefore have site level climatic data for historical and NPN, and 55km by 55km aggregated climate data for iNaturalist. Because our main interest is comparing trends at broad spatial and environmental extents, we argue these issues with differing extents and uncertainties simply add limited levels of unbiased error to these models.

In addition to identifying best practices, in our previous work, our lab group has used multiple ways to reduce bias due to oversampling of iNaturalist data over some days of the year, mostly during the City Nature Challenge when much higher numbers of observations are reported than during the rest of the year. The choice of how to reduce bias may result in

very modest impacts on downstream phenometric calculations, however these impacts are small compared to doing nothing, e.g. uncontrolled bias, which have a much larger impact.

However, to assure ourselves and the reviewer that iNaturalist data that are aggregated at broader spatial extents are not influencing our analyses, and also effectively responding to a comment from Reviewer 2, we re-ran models removing the iNaturalist species (and in doing so, also removing herbaceous understory species) to see if this impacted results. While we have less power to detect signal, the results are quite similar when just using NPN and Historical data and show nearly identical results for the key questions of the trends in phenology due to spring temperature velocity. We have added a supplemental analysis (Table 3, Figures 4 and 5) covering this and added in more details in the manuscript regarding these issues.

The additional text for the new analysis is in the Results and reads, “Both because we have very limited taxon sampling of herbaceous plants, and because iNaturalist phenometrics are at a more aggregated spatial resolution than the other resources used here, we also fit a set of models for a reduced set of 13 species whose present-day phenometrics were only from National Phenology Network, which also excludes herbaceous species. We found effectively the same results in both the reduced and full set analyses (See Supplemental Table 3 and Supplemental Figures 4 and 5).”

There were several terms (e.g., ‘phenometrics’) that were not clearly defined, and figures and their legends/captions lacked details.

We have updated the manuscript text to define phenometrics as empirically determined measures of the timing of leaf-out and flowering onset in plant species included in the study, and we have also included more information in the Figure 2 legend to provide clarity to the reader. We address the specifics of individual changes to Figures below.

Also, while GDD requirements are considered to be consistent across species in the study, given that previous studies have shown that there is immense intraspecific variation in phenology and how it responds to climatic change, I find this to be unlikely.

Thank you for pointing this out. A key component of our framework is the incorporation of variability in GDD requirements across species and we include this trait in our model analyses. We assigned Low and High GDD requirements based on expert knowledge combined with observations from the data. In this context, we do not consider GDD requirements to be consistent across species but highlight the importance of considering this mean species-level variation when examining phenological response. This is a central component of our general modeling framework presented in Figure 1.

If the comment is not referring to the mean timing of leaf-out or flowering/species mean GDD requirements, but to the slopes of response for individual species across

geography or environment, the modeling approach used here does fit independent slopes and intercepts of species; each species is specified to have its own slope and intercepts in response to temperature and latitude. The Figures show the main fixed effects slopes after accounting for these species-specific slopes. Since the fixed effects part of this model is accounting for the vast majority of the variation, we opted to not include the random effects plots. Finally, if the reviewer is making a similar comment to Reviewer 3 regarding the fact that our model is a simplification of a more complex reality, we do agree and have toned down our statements and clarified that we are not attempting to model all aspects of phenological sensitivity or response.

Similarly, though the authors claim that the specific amount of accumulated heat required to trigger a life cycle event has been quantified for many species, we really do not know what the threshold is for the vast majority of plant species, making the accurate inference of GDD unfeasible. I doubt good information exists even for all 18 species included in the study (which are never named in the manuscript).

Ranges of values for accumulated heat required to trigger leaf out and flowering are well known for many of these species, since they are very commonly used for phenology monitoring and form the basis for spring timing models, e.g. *Acer rubrum* (Red Maple) or *Syringa vulgaris* (Lilac). Timing of spring events and GDD requirements are highly if not completely autocorrelated by definition (if a plant has lower GDD requirements it is generally going to leaf out or bloom earlier) so we can use that timing as a proxy for requirements, and it is what we use here and refer to in our models (both conceptual and empirical). We clarify this in the Methods (more details about text added below). Given the above, we have made only minimal adjustments to the text. Regarding the 18 species, please see Supplemental Table 1 for a complete list of the species included in the study. We are not sure that simply listing the species in the main text would add value. However, we do refer to some of the species specifically where needed within the main text.

Further, much of the manuscript was rather confusingly worded, making it difficult to interpret. The study has potential and looking at spring velocity is interesting, but I think presenting the results as an extensible and generalizable model framework incorporating for predicting phenological response to warming is greatly overselling the study. The scope of the study, while broad, definitely does not come close to representing phenological variation globally and across the (plant) Tree of Life (as stated in the manuscript).

Thank you for pointing out the need to “tone down” the generalizability of the model framework. Throughout the manuscript, we have removed reference to being able to generalize across the Tree of Life and we have updated the language in several places. Specifically, we have updated the abstract to now state:

Our framework offers enhanced utility for predicting phenological sensitivity and responsiveness across multiple plant species with varying thermal requirements in the middle and eastern United States.

We have updated the last sentence of the Introduction to state the following:

Testing our simple framework within the context of differential responses to warming conditions demonstrates the potential to generalize this framework across a greater number of species and geographic areas.

We have updated the last paragraph of the discussion to state:

Because this model is fundamentally grounded in growing degree day accumulations, it has potential for broad applicability across additional species and geographic areas.

More specific comments:

Figure 1: Why is the intercept of the slow velocity line always higher? There can be instances where warming happens more slowly and starts at a lower temperature. The statement “thermal requirements are met earlier in the season in locations with slow velocities than those with faster velocities” may apply to the illustrated example but should not apply to all situations. Also, figure axes are not aligned.

We have modified this figure based on the reviewer’s comment. We agree that the “warmer” and “slower” slopes showing in Figure 1 could shift relative to one another in specific instances. But typically, there is a positive correlation between warmer starting conditions and “slower” advancement. Our goal in this Figure is not to show every possible contingency, but to keep parameters such as T-base and total GDDs accumulate more constant, so that the outcomes of varying other parameters are clear. Indeed, it is possible for the reader to derive their own estimates of how changing parameters in this model may lead to other outcomes. However, a key finding of this work is that our empirical results are strongly supported by our conceptual model, although error is still present, which likely reflects that the conditions set here do not demonstrate (nor is meant to demonstrate) all variability that can occur in nature. We have updated the Figure 1 caption to state that Tbase and overall GDD amounts are set to be consistent for both faster and slower spring velocity ramp.

Figure 2: Unclear what ‘phenometric densities’ are. Phenometrics are never defined in the manuscript, and the legend seems to indicate what is being shown are the number of phenometrics measured in each cell but it is not clear how these were counted. Is it the number of different measures (e.g., flowering onset, leaf out) or species, or individuals measured?

Thank you for pointing out the need for clarification. We have updated the manuscript text to define phenometrics as measures of onset of leaf-out or flowering in plant species included in the study, and we have removed the term “phenometric densities” from the Figure 2 legend. The legend now states:

Distribution of the number of phenometric estimates of leaf-out and flowering onset for historical (A) and contemporary (B) plant species included in the study across Eastern North America. Phenometrics are aggregated to a 55km resolution in order to match the spatial grain of iNaturalist phenological estimates. Latitude bands defining different temporal windows for spring temperature context and velocity are shown in grey, and the spring timing for latitudinal windows are shown in grey on the left panel.

Figure 3: Please depict the underlying data – effect plots can be misleading without them. There does not seem to be a pronounced difference between historical and contemporary data besides the intercept contrary to what is in the text.

When fitting mixed models, where all species are fit in the same model, depicting underlying data is often not illustrative or useful for the reader. However, we do understand the reviewer’s point, and we have instead added “rugs” to the Figure plots to show data density along the X-axis. Regarding the “pronounced difference between historical and contemporary data besides the intercept contrary to what is in the text”, we did not use the term “pronounced”. The differences in slopes between past and present are significant according to model results (at the $p < .05$ value), leading to relatively earlier onsets in present in the North than in South. In places where we could, we have attempted to temper our words

Figure 4: Please depict the underlying data – effect plots can be misleading without them. How was temperature velocity binned? The early, middle, and late flowering/leafing species seem to show remarkable similar trends within each temporal category. Would it not make more sense to plot different phenological events separately?

Thank you for this suggestion, and see above for our response adding rug plots. We did not bin temperature velocity - it is a continuous variable. Effects plots are designed to show “low, mean, and high” exemplar lines, as is done here. We agree that the trends hold across temporal periods but we also note there are important differences for species flowering/leafing early, middle or late that are well described in the text and that directly relate to the conceptual model presented in Figure 1. Here, one of the key objectives was to observe similarities and differences between past and present phenological responses and we feel plotting these together provides a visualization that allows the reader to observe this aspect of the study without having to move between outputs. Besides the rug plots, we haven’t changed anything here.

Line 64: confusingly worded – given the previous sentence it seems that “a unifying model characterizing generalities and underlying processes determining the mode and tempo of phenological change across space and time” applies to at least plants and animals and the range of seasonal behaviour they engage in – is this feasible? Surely the processes that determine the start of hibernation, for instance, are different from those that determine the timing of spore release in mosses?

Our argument is geared towards examples where heat accumulation triggers phenological events, and we think contextually it is clear we did not intend to include every possible phenology scenario. We also were careful to exclude endotherms from our discussion. However, we see the point that it is better to be cautious, and we have updated the text to narrow the focus a bit. The sentence now states: This rapidly growing body of work provides a foundation for understanding the shape of phenological responses across multiple plant species and other ectotherms along environmental gradients”

Line 158: confusingly worded – are only leaf onset metrics divided into historic/contemporary or does this apply to flowering as well?

Thank you for pointing out the need for clarification. The sentence now states:

In total, there were 1,714 historical and 2,391 contemporary flowering onset metrics, and 1,689 historical and 3,179 contemporary leaf onset metrics.

Line 161: Figure 2 does not seem to illustrate what is presented in the text.

Fig. 2 provides a map of the study area and distribution of the number of onset metrics used in the analyses partitioned by historical and contemporary data records. The text describes the overall number of records available to calculate onset metrics for the historical and contemporary data sets and provides additional detail about numbers of records available for flowering and for leaf-out for each time period. We have removed a reference to the Figure and point only to the Supplemental Table 1 for sampling effort across years.

Line 165: confusingly worded sentence “Linear model testing whether period (historical or contemporary) predicted velocity revealed..”

Thank you for pointing out the need for clarification. The sentence now states:

The linear model testing whether period (historical or contemporary) had a strong effect on velocity revealed a small (0.28°C), though significant, decrease in velocity (i.e. slope) in the

contemporary observations across all latitudinal bands; this effect was strongest in the Southeastern USA ($p < 0.001$, adjusted $R^2 = 0.34$).

Line 170: multiple models or just one? Only one seems to be described.

Good point. We have changed this to read “Conversely, we found that warming has been fastest in the North, based on a linear model showing a strong interaction between time period and latitude ($p < 0.001$, adjusted $R^2 = 0.588$)”

Line 175: what is a “decisively better result”? Please refer to metrics/values.

Thank you for pointing out the need for clarity. We have removed this sentence because it introduces a new topic in the Results section and we do not elaborate on it further in the manuscript.

Line 257: as the authors note in the manuscript, spring temperature variability and velocity are strongly correlated. If so, using velocity to model phenology should not differ much from using variability.

We appreciate this comment and realize we didn't do as good a job as we should have making clear what we meant regarding “variability” in this section! Our intention with these statements in the Discussion was to suggest a separation between spring temperature velocity, which here is a modeled slope of spring warming, and direct variability around that fit, effectively the residuals of this fit. Higher variability reflects a poorer fit, which could be due to multiple reasons, including extreme weather events. Just using standard deviations of the spring climatic time series does not disentangle those two factors. We really don't know how much spring climate velocity and extreme events are related when it comes to phenology and phenology drivers, and it is therefore a frontier area we are interested in exploring more fully. We now say, “We have reframed the question of sensitivity away from variability, which often is ambiguously defined, to spring temperature velocity, i.e. the rate of warming over the spring season. In the context of the velocity of warming, variability can be thought of as the residuals of the model fit of spring velocity. This aligns with how extreme events are defined, e.g. conditions that are far outside the bounds of typical conditions during a defined period. In this formulation, the importance of extreme events represents one axis of pressure on plants and how they phenologically respond 38, and spring warming velocity is another. Those different axes are likely intimately connected, and may themselves interact and should be the target of further study.”

Line 255: The question of sensitivity across latitude has remains unsettled because it is not latitude itself that influences phenology, it is other variables that sometimes covary with latitude. As the authors note on line 257, only ~30% of the variation in velocity can be

explained by latitude. Thus the crude latitudinal binning used in the study is not optimal. Perhaps binning by climate could be an alternative?

We agree with the reviewer and our point is the same as theirs - latitude itself is not influencing sensitivity. Our analyses suggest spring temperature velocity is likely important. But we disagree about the latitudinal binning. The binning was only used to define the proximal time windows for measuring climatic variables used in this model (spring temperature velocity and spring average temperature) and latitude itself is a continuous variable in the models. The value of the definition of these appropriate windows is to reduce a serious problem in phenology studies across latitudinal gradients - an autocorrelation between temperature and latitude that is mostly removed here. We have tried to further clarify that binning was to define temperature windows but otherwise have not changed manuscript text.

Line 278: There are many species that leaf out after flowering (e.g., cherry blossoms) – perhaps the authors mean that this is unusual in their dataset?

We have removed the phrase “somewhat unusual” and now state:

For example, *Acer rubrum* flowers precede leaves, while many other taxa leaf-out before flowering.

Line 281: Perhaps I’m missing something here, but this does not seem to sufficiently justify why leaf out and flowering timing were not allowed independent responses to the fixed effects.

We did allow leaf out and flowering timing to independently vary in the model. What we are saying here is that we did not fit an interaction term in the model between timing (early, middle, late) and phenological event (flowering and fruiting) because this particular interaction showed modest variance inflation, but enough to warrant not fitting that particular interaction. For many other main effects and interactions, leaf out and flowering independently vary. We realize the wording was unclear so we have been more explicit and now say, “We did not directly examine this question because while there was a significant interaction between species-specific timing of onset (e.g. early, mid, or late) and phenological event (leaf out or flowering), the variables were moderately collinear, so we excluded this interaction in our final models.”

Line 293: What data driven approach?

Here, we leveraged our data set to inform the sliding window approach. We have made this clearer and removed “data-driven”. The sentence in question now reads, “Here we have used known rates of change in onset over latitude to define appropriate temperature windows and

verify these windows given our onset data.” The remainder of the paragraph describes our use of the data set to help inform the sliding window approach (i.e. data driven), and additional details of the approach.

Line 296: The references examine a very small number of woody species and are not sufficient grounds to expect species to have relatively similar GDD requirements across space more broadly.

We have adjusted this sentence to say “While it is possible that adaptation has acted to shape populations differentially in different regions, we expect species to have relatively similar GDD requirements across space, although more work is still needed to test this supposition”.

Line 313: Please elaborate what sorts of differences and biases were present in the data.

We have noted the key biases throughout the manuscript, e.g. observer effort, differences in phenophase definitions among the programs, and the ease with which phenological status can be determined. We have expanded this section to make it clearer that, in particular, the historical data from Hough is more limited in reporting exact information on phenophase and observer effort. We have already made the point that these issues have led us to tread carefully here about absolute rates of change. The added sentence reads, “In particular, the historical dataset used here, while providing yearly replication across latitude, lacks detailed metadata about observer level of effort for determining onset, and what exact phenophase e.g. budburst or unfolding leaves or both were used to define onset.”

Line 324: How was it determined that “The model correctly predicts onset timing for leaves and flowers, capturing variation in sensitivity and responsiveness more fully than other analyses across species and phenological events”? I did not see any model evaluation or testing on non-training data.

Thank you for pointing out the need for clarification. We have updated the text of this sentence to state:

“Empirical observations of onset timing for leaves and flowers were consistent with expectations under our model framework and captured variation in both sensitivity and responsiveness across the study area.” We have also generally toned down the concluding paragraph regarding broad applicability.

Line 358: Provide the cleaned files please.

Please refer to our data availability statement, which indicates that the files are available through our Zenodo repository.

Line 376: Research grade observations are not truly vetted – they are just observations that are agreed upon by a certain number of people who may or may not possess botanical knowledge.

We have removed the word “vetted” and updated this sentence to reflect the iNaturalist research grade identification process: iNaturalist RG observations are community-collected digital image vouchers of organisms in nature with species identification consensus from two or more users.

Line 384: Unless the entire plant was captured in the image, it is not appropriate to infer absences. For instance, the absence of flowers on a part of one branch on a tree does not guarantee that there were not flowers on other branches or other parts of the same branch.

We agree that it is not possible to infer absences without observing the entire plant. Here, we use a presence-only data approach to estimate the timing onset metrics and do not consider absences in this calculation. This approach is commonly and successfully used to estimate phenometrics and we point the reviewer to simulation results showing this, if of interest here: <https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13448>. We have made clear that our approach is presence-only based in multiple places in the manuscript.

Line 385: Absences and NAs should not be treated the same.

We agree that absences and NAs should not be treated the same. Here, we use a presence-only approach and do not include absences or NAs in the calculation of onset metrics. We have updated the sentence to state the following to clarify to the reader:

“In rare cases, images were difficult to score for a variety of reasons and we excluded these from the data set.”

Line 391: Why 55km by 55km equal area grid cells? Both the phenological data and climate data are higher resolution. Why this particular cell size? This also does not align with the 50km buffers mentioned later.

See above for our explanation regarding spatial resolution of iNaturalist phenology estimates. Regarding the 50km buffers used later for filtering: Because phenology sampling in the contemporary dataset is more spatially extensive than in the historical dataset, we performed a filtering step to only include contemporary phenometric estimates within 50km of historical estimates. We placed a 50km buffer around the historical sampling sites and used the union of those buffers to create a new polygon, and then intersected the contemporary data that fell within that polygon. This step reduced the number of phenometric estimates to 3,403 historical and 5,570 contemporary estimates and helped to ensure that historical and contemporary

locations analyzed are within the same general area. We addressed these above; please see other points in the revision letter for changes made.

Line 394: Why keep 5 or less? Why not just consistently keep one? Also do you mean 5 per species? 5 per phenological event? 5 regardless?

Please see above. Here, we set a ceiling on the number of iNaturalist records per day, randomly selecting 5 if there were more than that number. Presence-only models are sensitive to sampling biases over time, and in the case of iNaturalist data, the City Nature Challenge can lead to oversampling early in the season. We used this ceiling to help reduce this bias to improve estimates.

Line 398: This sounds like later season data were excluded to fit the authors expectations and make the models fit better?

This is not in the case. This is purely about using the best approach to fitting onset phenometrics from iNaturalist data for 3 woody species. For these 3 species, and given that duration and offset are not of interest in this study, we can focus on scoring presences of unfolding or unfolded leaves through the spring and early summer to capture the onset timing from those data. Adding data from later time periods does not improve estimates of onset timing. We also have provided a supplemental analysis without iNaturalist-derived phenometrics that shows very similar results to the full analysis.

Line 401: Please elaborate – not clear how these metrics were calculated.

In addition to the citation to the *phenesse* package, we have added an additional sentence to the text:

The *phenesse* package provides a set of methods for accurately estimating different percentiles of a phenological distribution using used presence-only data.

Line 403: which phenometrics?

We have updated the text to state 5% onset phenometrics where necessary to be absolutely clear which metric was calculated.

Line 407: If the observations are in the same 55km grid, why is this 50 km buffering necessary?

See above for the response. The filter used at a 50 km distance helps to ensure that the historical and contemporary locations analyzed are within the same general area and helps

to reduce potential biases from greater sampling extent during the contemporary time period. Please see the paragraph in the Methods section.

Line 428: What is the rationale for this binning strategy?

Please see above for the response. We have covered this elsewhere in this response letter.

Line 452: Given that spring was defined as 3 month bins, does this mean the velocity of spring warming was calculated as the slope of only 3 points? This does not seem very optimal.

We calculate spring velocity based on the underlying climate data as described in the Methods. Each onset estimate is associated with a velocity measure (the slope of that year's spring warming for the defined temporal window). There are ~2500 unique velocity measurements in this study (note: this value is fewer than the number of phenometrics since multiple species could have phenometrics in the same region and year).

Line 465: What were these binnings based on? Is there ecological support for these very specific 'early/mid/late' periods? The 'mid' period is only 10 days and plenty of species flower/leaf out across more than that. There can easily be a week's difference in the onset of flowering in the spring across 1 degree of latitude, let alone 5.

The binning was based on clearly defined breaks in species flowering/leaf out mean onset timing as described in the text. The mid period is ten days, but the early period ranges from a mean onset time of 89 days for *Sanguinaria canadensis* leaves to *Acer saccharinum* flowering at 101 days. So the early period only spans 12 days. The late period has a longer duration (from 118-157) but this reflects a longer tail of a couple species that have very late onsets (e.g. *Tilia americana*); most of the late species are between 118-128 days. We have opted to not change anything in the text.

Line 475: I thought spring was defined differently across latitudinal bins – why are they consolidated here?

We have changed this sentence to, "For mean spring temperature models above, we used a single range - from February to June - rather than our sliding window approach." This statement only applies to the above models which do not include species onset as a response variable. We have clarified this in the manuscript and thank you for catching the need for clarification.

Reviewer #3 (Remarks to the Author):

In this manuscript the authors present a general model framework to model plant species' phenological responses to climate change. They consider the spring temperature, the velocity of warming in spring and the species-dependent thermal requirements (GDD). To test their model, they used phenological data from 18 species collected in the eastern US over 9 years in both the 19th and 21st centuries. The authors found that temperature sensitivity is higher in regions with low temperature velocity and that the greatest realized change in phenological onset was found in northern areas where warming rates have been fastest. They assume that their generalised and simple framework can be very helpful for predicting phenological sensitivity and responsiveness across the Tree of Life.

Major comments:

1. I really appreciate your approach of presenting a simple framework that takes spring warming into account.

Thank you for the positive feedback!

However, in my opinion some of the statements and conclusions in this paper are a too strongly worded and should be toned down. This begins with the wording of the title. You have only considered one trait (temporal phenological niche coupled with GDD) in the analyses.

Thank you for pointing out the need to tone down the statements and conclusions in the manuscript. Regarding the title, we decided to simplify rather than add too many details, opting for: **Phenological response to climatic change depends on spring warming velocity**. We cover the context for how spring temperature velocity interacts with other factors but this brings home the main idea and keeps it pithy. In addition to updating the title of the manuscript, we have toned down statements in the text throughout the manuscript.

Throughout the manuscript, we have removed reference to generalizability across the Tree of Life and we have updated the language in several places. Specifically, we have updated the abstract to now state:

Our framework offers enhanced utility for predicting phenological sensitivity and responsiveness across multiple plant species with varying thermal requirements in the middle and eastern United States.

We have updated the last sentence of the Introduction to state the following:

Testing our simple framework within the context of differential responses to warming conditions demonstrates the potential to generalize this framework across a greater number of species and geographic areas.

We have updated the last paragraph of the discussion to state:

Because this model is fundamentally grounded in growing degree day accumulations, it has potential for broad applicability across additional species and geographic areas.

Furthermore, you often mention that data spans a period of 175 years. In my opinion that is not the case. Two periods from different "climate contexts" were analysed, each lasting circa 9 years.

We agree that the wording of the study period should be more explicit to indicate two different time periods. We have updated the study period throughout the manuscript, e.g.

"We tested this framework using a unique collection of phenology observations spanning the Eastern USA and covering two time periods separated by more than 150 years."

"We also evaluate the rate of change in phenology across historical and contemporary datasets spanning >150 years and along a broad latitudinal gradient in the Eastern USA, a region where temperature is known to be a dominant driver of phenological response"

In lines 272-289 you show that their model is applicable to understory and canopy plants and discuss the generality of the results for early flowering species. I am not convinced by these statements, as only two herbaceous species were included in the data set! In fact, I would rather advise you to leave out these species and focus on trees and shrubs and then compare these two growth-forms with each other. This is particularly relevant if, for example, the results from Horbach et al. 2023

(<https://doi.org/10.1016/j.flora.2023.152218>) are taken into consideration.

Thank you for your comment. We find this to be an annoying data limitation, and it turns out there are many more herbaceous plants that could be used in the historical data but only for flowering. However, we did want to keep leaf and flower comparisons between the two time periods for each species. We have implemented this suggestion as a new Supplemental analysis, where we removed all species where onset metrics were generated from iNaturalist in the contemporary time period, and tested whether results changed. Removing iNaturalist species also removed both herbaceous plant species. Results indicate that our key findings hold with some adjustments in parameter estimates and p-values for specific variables (please see relevant Supplemental Table and Figures). We opted to leave for

future work comparing shrubs and trees and herbaceous plants to not overly complicate these analyses presented here.

At the end of the conclusions you write that the model presented has wide applicability across the Tree of Life. I would expect this too, but it would have to be tested first.

Please see our response above about overall “toning down” the language in the manuscript. For the conclusion, we have updated the last paragraph of the discussion to state:
Because this model is fundamentally grounded in growing degree day accumulations, it has potential for broad applicability across additional species and geographic areas.

2. The layout of this journal determines that the section on material and methods is at the end of the paper. In many parts of the "Result" section, you provide brief insights into the methods used to help the reader to understand the results. But especially when it comes to the key results of this study the LMM model with all explanatory variables (Table 1, Figure 3&4), it is difficult to follow all the statements. This is due to the high complexity of the model with many interactions and factors, as well as the relatively small visible effect sizes (Fig. 2). I wonder whether the model can be simplified even further by restricting the design of the model to interactions with comprehensible hypotheses. The rule can apply here: Less is sometimes more.

We agree that the model presented here is complex, but our design for these models do basically follow the conceptual framework we lay out in Figure 1. One strategic choice was to examine both leaf out and flowering, which definitely adds complexity but we hope it also extends generalizability. Per our conceptual model, we started with a reduced set of predictor variables based on biological and ecological knowledge about plant phenology, and we presented a set of expectations prior to the analyses (Burnham and Anderson 2002). We opted to use backward stepwise selection as our model selection approach in order to identify results from the model most strongly supported by the data. We did not change the manuscript directly as a result of this comment but we hope the other changes have helped clarify some issues.

Burnham KP, Anderson DR, Burnham KP. Model selection and multimodel inference: A practical information-theoretic approach. 2nd ed. New York: Springer; 2002. xxvi, 488 p.p.

In addition, adding paragraphs and more explanations or definitions for the names of the different explanatory variables to the result section could make it easier to follow the findings.

We have added some more explanation of the predictor variables in the model fitting section to clarify a few things, e.g. “Next we fit a single linear mixed effects model (LMM) to test our predictions using the lme4 package in R where the response is day of year of the onset of either flowering or leaf out and predictors included mean spring temperature, velocity, both defined based on the sliding window approach described above, as well as latitude and longitude as

continuous fixed effects variables. We mean centered and scaled all continuous predictors using the `scale()` function in R. We also included “state” (flowering or leaf out), “period” (historical or contemporary), and “timing” (early, mid or late) as categorical fixed effects predictors.”

3. The following point certainly does not require much text to be written, but I consider it to be important. Throughout the manuscript, the GDD concept is described as universally valid. I mostly agree with this statement. However, the discussion should briefly address how the species-specific choice of base temperature would affect the relationships shown. It is known that phenological events should be predicted in different species with different thresholds for the base temperature.

Thank you for this comment, it is an important one, and we absolutely agree. We realize in that trying to create a “simplified” conceptual model, we glossed over the many areas of complexity that are very worth raising. We have added a key sentence into the Conclusion to surface this point and to clarify that these base temperature differences are critical here as well, and likely reflect some of the underlying uncertainty in our models. The sentence reads, “Finally, while we have not addressed species-specific adapted differences in base temperature, they are likely to impact the shape of responses predicted here and will be important to consider in further refining this framework.”

Minor comments:

L1: I would revise the title of this paper. You have only considered one trait “temporal niche” in the analyses. The title suggests that several traits were analysed.

We have updated the title:

Phenological response to climatic change depends on spring warming velocity

L75: What do you mean with “temporal record”? Is this the period over which the phenological data was collected?

We have updated the sentence to state:

Individual studies evaluating phenological change vary over time in observation methods and metrics examined, impacting findings⁷.

L75: “Quality” is rather a vague term. You should be more precise here.

We have removed the term “quality” from the sentence.

L106: “(the rate of warmth accumulation)” I think this can be deleted. It should be clear to readers from the previous explanations what is meant by this.

We have deleted “(the rate of warmth accumulation)”.

L109: I really like the graphical framework! It might be useful to reference what is meant by "cool" and "warm" context. An example like before with the lake for velocity would be helpful.

Thank you for this comment. We have updated the the text to state the following:

Locations with low and high warming velocities can occur in both warm and cool climate contexts, e.g. in regions where spring is on average generally warmer or colder.

L135: I would already mention here briefly that the species were trees, shrubs and herbaceous species.

We have updated the sentence to state:

We test the predictions of the model using historical and contemporary phenology data, focusing on 18 common plant species, which include trees, shrubs, and herbaceous species in Eastern Temperate Forests in the USA that vary in their timing of leaf-out and flowering.

L137-138: “[...], who established what is believed [...]”. I would suggest to move this sentence to the M&M.

Thank you for this suggestion. Given that the M&M are located at the end of the manuscript, we would like to highlight early on how we are leveraging historical data with contemporary records. We think this provides context to the reader and helps orient them to the value of the work presented here.

L141: Add a space: “model, “

Done.

L143: Do you mean the “cold” and “warm” context mentioned before with “spring temperature context”. If yes please be consistent in the wording.

Thank you for pointing out the need for clarification. We have updated the sentence to state:

We use these historical and contemporary phenology data to test predictions of our model, in particular, that species’ GDD requirements, spring temperature velocity and spring temperature context (i.e., warmer vs. cooler) interact to determine spatial and temporal trends in onset phenology.

L158/Fig. 2: The term "phenometric" has not yet been explained. Since M&M are at the end of this paper (Journal style), I think it should be made clear in the results or in the caption of the figure what is meant by this.

We have updated the manuscript text to define phenometrics as measures of the timing of onset of leaf-out and flowering onset in plant species included in the study, and we have also included more information in the Figure 2 legend to provide clarity to the reader. The definition is included at the start of the Results.

L161: The most intensive sampling years can't be seen in Fig. 2.

Thank you for this comment. We have removed a reference to the Figure and point only to the Supplemental Table 1 for sampling effort across years.

L166: You should write "0.28 °C". In addition you may change this to °Kelvin.

Thank you for pointing out this detail. We have updated the text. Here we opt to remain in units of Celsius because we feel the reader is more familiar with these units than Kelvin.

L169: Before (e.g. L165) you wrote " $r=0564$ ". Please be consistent with the spacing (again in L263)

Thank you for pointing out this detail. We have checked the manuscript for consistent spacing and updated where needed.

L174: Please be consistent with the number of digits shown in Table 1. The values of the deviation should have the same number of digits as the effect size. Furthermore, in my opinion, tenths of digits are sufficient for the effect size, as we are talking about days in a year that probably cannot be measured more precisely anyway.

Good point! We have adjusted this table.

L182: Please be consistent with the number of digits shown in Supplemental Table 2

Done.

L187: "middle" or "mid" instead of "midde"

Good catch. Fixed.

L291: In Line 214 you added no “:” after the sub header but a line break. Same issue again in L311.

Thank you for pointing out the need for consistency. We have updated the sub headers throughout the manuscript by removing the “:” and adding a line break.

L312: Why are you using the term “replicated “experiments””? I would rather use “space-for-time-substation” (different locations across a spatial scale and “time-for-time” (data from the past and from 21st century)

We have changed this wording and removed the “replicated experiments” statement. What we mean here is that we have multiple years (8 years between 1852-1859) and 13 years (2010-2022) in two different time periods where we can look at yearly latitudinal (and environmental) trends in phenology. This yearly replication is valuable. We now say “A strength of the approach here is multi-year temporal replication both in the historical and contemporary dataset across latitudinal gradients in order to examine trends in phenology.”

L327: I would say that it CAN be applied to insects. But that would also have to be tested.

We have removed the mention of insects and updated this sentence to state:
Because this model is fundamentally grounded in growing degree day accumulations, it has potential for broad applicability across additional species and geographic areas.

L337-339: To what extent would this be possible? In particular, could the effects of extreme weather events such as warm temperatures in late winter be explained using your framework?

We love this question and we think it is very testable, but we haven’t done it here and we think the next step is combining these extreme events with the framework developed in this work. But we think this is maybe best stated (as it is currently) as a future step. We hoped this framework would be extended, tested and re-examined! We haven’t made a change to the text but appreciate the comment!

L410-411: It might be interesting for the reader to see what original number of observations you have reduced by filtering.

We have added a sentence covering these numbers to the main manuscript. We reduced about 30% of the current observations via spatial filtering.

L460-461: If I have understood your approach correctly, you are using the timing of flowering or leaf out as a proxy for whether little or a lot of GDD is needed? If this is the case, it should be made clear here and, in the introduction, that the "trait" considered is the magnitude of the GDD.

Yes, you are right and agree that we should clarify this. We have added this as requested as follows, "In this study, we focused on a single species trait: species mean timing of flowering or leaf out, which serves as a proxy for the amount of GDD required for a phenological event to occur."

L480: Why were the continuous predictors scaled? Does this method refer to lines 317-319? If so, it should be mentioned there again.

Scaling is a common practice in model fitting to compare effect sizes and to help reduce the potential for numerical errors because of differing units in predictor variables. We have clarified this where suggested above. The sentence now reads, "We mean centered and scaled all continuous predictors using the scale() function in R in order to compare effect sizes and improve model fitting."

L708: How were the three groups defined for temperature velocity? I couldn't find anything about this in the M&M.

These are not really "grouped" - the effects plotting package has a default to show three lines representing low, mean, and high values when plotting interactions. The values for temperature velocity are still continuous but plotted as grouped for readability.

Reviewer #4 (Remarks to the Author):

I really enjoyed reading this paper, it was compelling to include temperature velocity of change as a factor that would differ across biomes yet allow a generalizable variable in a model.

Thank you for your positive feedback!

However, I felt the paper needed more discussion on the utilization of species' thermal requirements and how to appropriately estimate them for this sort of model. I understand that the GDD requirement is based on whether you accumulate fewer or more GDD by the time you start some phenological events, but is this taken from 1 year of data to estimate if a plant was early or late and assigning a GDD threshold to it?

It is based on well established literature that spring forcing based on GDD accumulations is the single most important thing determining onset timing. So for each year, region and species, we collected spring climate data relevant to our conceptual model and hypotheses and used those to test the model and hypotheses. With regards to species GDD requirement calculation, we used mean species onset timing as a proxy for this, over all the years in the historical and contemporary datasets, spanning approximately 22 years (9 for historical data, and 13 for current). We covered this elsewhere in revisions for the other reviewers re: manuscript edits that clarify this.

What level of multi-season GDD estimates do we need to average? Have studies done enough thermal performance curves using GDD to estimate GDD thresholds (T_{opt} or T_{min})?

Luckily these are well sampled species that are routinely monitored, but we hopefully now make clearer via our revisions that we do not actually assemble GDDs for these analyses. We only use spring velocity and spring climate context, which we calculate from gridded climate data. These data are available, in the 1850-1859 time period and in present.

Finally, climate-relevant measures of phenology (e.g., the number of GDD accumulated by time of flowering) can shift with climatic conditions (e.g., Wadgyman et al. 2018 New Phytologist) – can these threshold account for shifts in the thresholds themselves?

This is an excellent point also raised by Reviewer 2 and we refer to that answer above. We agree, and we are not suggesting that the conceptual model we present, the predictions it generates, or the empirical analyses cover all the ways that plants can respond phenologically over short or longer time scales. This effort is more geared towards a simple but useful framework and we hope new parameters such as changes in GDD requirements can and will be incorporated. We note a sentence or two in the conclusion that we hope to help make this clear.

Minor comments:

L42: Unclear what mode means here.

One author has long been influenced by G.G. Simpson who wrote “Tempo and Mode of Evolution”. While this work is all about “tempo” or timing, the mode part is more about processes underlying the timing, which we do want to cover here. Because of this underlying homage and its relevance, we have opted to keep this wording.

L66: There are multiple global meta-analyses and large-scale studies trying to determine generalities regarding phenological shifts since the 2000s; these statements generally need to be toned down.

Agree, we never meant to oversell but we do think this framework could have general applicability. We have worked hard to tone down these statements.

L296: Populations often differentiate in their thermal performance, suggesting different GDD requirements across space. Could be worth trying latitudinal-band-specific GDD requirements?

We appreciate this idea and its related to other comments and responses. We have tried to keep our focus on a conceptual framework that makes key predictions and testing just those predictions here, but we do think more exploration and including more realism in follow up work is warranted.

L406: I appreciate the similar scaling of the contemporary and historical datasets here.

Thanks!

L415: Somewhere in this section please explain how the GDD requirements were determined for each species.

We have mentioned this above in response to reviewer #1 and #2 and repeat it here since the response is the same. "Ranges of values for accumulated heat required to trigger leaf out and flowering are well known for many of these species, since they are very commonly used for phenology monitoring and form the basis for phenology models e.g. *Acer rubrum* (Red Maple) or *Syringa vulgaris* (Lilac). Timing of spring events and GDD requirements are highly if not completely autocorrelated by definition (if a plant has lower GDD requirements it is generally going to leaf out or bloom earlier) so we can use that timing as a proxy for requirements, and it is what we use here and refer to in our models (both conceptual and empirical)."

Fig. 3: Suggest indicating which slopes are statistically different from each other.

This is challenging since it is the main effects and their conditioning by the interaction terms that determine these slopes. Several interaction terms presented in Table 1 show significant effects, and we feel the correct approach here is to refer to Table 1 so that readers can observe the direction and significance of values across different terms included in the model.

Fig. 4: It'd be nice to include a Fig. 5 with a similar structure, but velocity on the x-axis.

One thing one of the authors has never loved about effect plots is that changing which axes are displayed can, on first look, appear quite different. However, including velocity on the x-axis and plotting similar terms on the other ones is basically showing the same result, and given our key predictions, we do not feel plotting additional interactions is particularly useful here.

Reviewer #1 (Remarks to the Author):

We appreciate the effort that went into addressing the issues raised. The manuscript is clearer than before, provides more details regarding the methods employed, and claims and expectations have been toned down to be more realistic.

Thank you for the positive feedback!

However, there was still confusion around the figures and results (our lab discussed the manuscript as a group, in confidence). Several found fig1 confusing and non-intuitive. Perhaps changing the y-axis to GDD would make what the authors are trying to convey clearer.

We admit to some surprise that the figure was shared amongst a whole lab group. However, the feedback that it was found confusing pushed us to revisit it to consider further clarifications. The purpose of this figure, as explained in the caption, is to present a relatively simple model varying three separate parameters: *spring temperature velocity*, *temperature context* and *species' GDD requirements*. These are the key variables we measure in our empirical models, and this linkage is essential for providing a test of the predictions of this model. A final key aspect is that base temperature has been kept constant. Day by day (or hour by hour) temperature above a threshold is what determines GDD accumulations and warmer contexts mean that the base temperature is exceeded faster than in cooler contexts, which is why temperature is on the y-axis and GDD accumulations happen earlier in warmer contexts. Because increasing *temperature velocity* has a direct relationship to *GDD accumulation rate*, increasing temperature faster increases GDD faster as well. So, in this model, it is *temperature contexts* and *spring temperature rate increases* that have a large part in determining onset of heat accumulation and the point where a phenological event happens. Given how critical it is to link our framework with empirical analyses, we feel we cannot adjust this figure. *However*, we have updated the caption of Figure 1 to be clearer, and in particular to be explicit that increasing temperature rates must increase GDD accumulation rates and amount. Here is significant revision of the caption:

“Fig 1. Simple model where base temperature has been held constant, showing how temperature context (warmer versus cooler), species GDD requirements, and velocity of increasing spring temperature impact the relative timing of phenological events. Green lines represent locations exhibiting slow spring temperature velocity; orange lines represent locations with fast temperature velocity. Blue triangles represent a base temperature above which warmth begins to accumulate. The faster the spring velocity, the more rapidly GDD accumulates after the base temperature is reached, but especially in warm contexts, that base temperature is reached later. The timing of onset for phenological events is shown in green and orange dots. In cooler climate contexts, locations with slower temperature

velocities lead to thermal accumulations being reached later in the season than locations with faster velocities. Conversely, in a warmer climate context, thermal requirements are met earlier in the season in locations with slow velocities than those with faster velocities.”

Likewise we found fig2 to not be very informative and difficult to reconcile with the analyses that were done on these data. Not using a log scale and clarifying whether the phenometric count is species specific or not would help. Also it would be useful to parse the map out by the phenological event recorded, as one would not directly compare leaf out in one region with flowering in another, and it is important to be able to see which areas are consistently sampled and which are not.

Thank you for pointing out the need for further clarification on Figure 2. We have updated the figure to now include 4 panels showing historical and contemporary sampling effort for flowering and leaf-out. We have also added a flower symbol or a leaf symbol to each panel to provide clarity to the reader. We have opted to keep the log scale used in the previous figure in order to show more clearly the variation in sampling using our continuous color scale across time periods and now across flowering and leaf-out samples. A version without log-scaling did not visually pull forward those differences as well. We feel the log-scaled version provides the clearest visual to the reader and allows a visual comparison across the data.

The rugs in fig3 are a welcome addition, though the raw data would still be preferable. However the rugs for contemporary and historic data overlap making it hard to see how the data are distributed. Perhaps it would be possible to have two separate rows of rugs?

Thank you for these suggestions regarding Figure 3. The rugs do represent the raw data on the x-axis of the figure. After investigating multiple options, it turns out a “double rug” is not non-trivial with existing tools, packages, and software. While there is some visual overlap in the distribution of the rugs between contemporary and historic data, we feel that the rugs provide a visual guide to the reader as a reference for distributions. With this in mind, and after multiple attempts in R and Adobe Illustrator, we prefer to keep the rugs as they are and again, note that the data used in the analysis is made available through our github repository. We did however fix other minor issues with Figure 3 re: figure dimensions etc. as requested below.

Additionally, while we understand the differences depicted in fig3 are significant, the visualization does not illustrate this effectively. We suggest it might be better to just show these results as a table of coefficients. Another alternative could be to combine the information 3 panels into 1 using different line styles. This would at least make the comparison between early-middle-late categories easier visually.

We do provide a table of coefficients - it is Table 1. What we are showing here are the effect plots *based on the best-fitting model* for the key variables of interest related to: 1) latitude; 2) past versus present datasets; and 3) flowers versus leaves. Because models with interactions can complicate interpreting model coefficients, we feel effect plots are by far the best way to understand these results. We have opted to not change this figure.

We also noted that capitalization and axis labels were not consistent and figure dimensions seemed odd, but recognize that some of this may be due to the conversion process.

Thank you for this comment. We have tried to fix any and all issues with figures.

Finally, we again strongly suggest the inclusion of the list of species examined in the main text. It would greatly facilitate our interpretation of the results.

Our main question in this paper is proposing and testing a model that can explain timing of temperate plant phenological events based on species' GDD requirements, along with spring temperature velocity, and temperature context. In the context of this study, species replication is particularly critical because that allows for varying GDD requirements. Given this, we simply don't think that adding a table of 18 species is worth the space in the main text, especially since Figure 2 covers sampling rates over space, time and now flowering versus leaf-out. We have instead added this sentence in the first paragraph of the results: The 18 species used here represent commonly observed species (e.g. *Acer rubrum*, *Liriodendron tulipifera*, *Sambucus canadensis*, *Prunus serotina*, *Syringa vulgaris*; see Supplemental 1 for full list).

Reviewer #3 (Remarks to the Author):

I appreciate the great effort you have made to revise this manuscript. Below are some suggestions for improvement, which I think should be taken into consideration:

L80/L94: I know it is the style of the journal not to give the author's name in the text. But from my point of view it looks a bit strange to write 'e.g.' followed by the superscript numbers. I think you can also omit "e.g.", as it should be clear that quotes given are often only examples.

We will leave this one to the Editors to adjudicate.

L159: Space aber "vs."

Done

L176: Add a "." after Fig.

Done

L190: Delete space after " $= 0.34$ " to be consistent.

We couldn't find the space! We think this is ok.

Reviewer #4 (Remarks to the Author):

I appreciate the authors' time spent clarifying the language and decision making in this manuscript. However, I am still concerned about the estimates of GDD "thresholds" (versus plastic shifts in phenology) and think it's still trying to generalize much more than an ENA dataset on just plants is able to. Estimating the mean GDD accumulated at time of leaf-out (and other phenophases): How is this a set threshold? GDD accumulated at time of life history event will likely be a plastic trait. To get a threshold, I think you would need a critical minimum GDD required for annual growth to begin (in the case of leaf-out). Since these species do not have TPCs for every one (instead, records that allow estimates of mean GDD accumulation at time of phenophase), the framing should be adjusted.

We have simplified our modeling framework for the purposes of testing predictions such that GDD accumulations are not plastic. We do mention in the Discussion, based on the great set of first reviews, that this assumption is probably not warranted and it likely represents unaccounted for error in our empirical models. We think this is less a problem than an exciting opportunity to begin teasing apart these aspects more fully down the road.

Generalizing language: I highly recommend removing ectotherms, and potentially, as another reviewer mentioned, focusing on the trees and shrubs that make up the majority of the dataset. There are also still a lot of arguments for universal models when this is really testing effects of rates of warming in a much narrower spatial, climate, and species context.

We have removed "ectotherms". We did actually do a narrower analysis with all the herbaceous taxa removed (they were all part of the iNaturalist-derived phenology metrics), which we included as a supplemental analysis and which does not alter our results, so we think this covers the other part of the reviewer's comments. We agree in general that we

should not overgeneralize but we also think this modeling framework, which is new, can now be further explored and tested to determine how broadly it applies.