We would like to thank the editor and two reviewers for the productive feedback on our manuscript. We have addressed comments related to the sensitivity of the 30-day cutoff in using the USA-NPN data, and included general recommendations for which models performed best in different contexts. More detailed responses to individual comments are included below.

1 Comments from the SME:

General comment: I appreciate the authors' efforts to ensure transparency and repeatability of the analysis.

Thank you.

l. 20: not all the models are process-based; revise?

We changed this to "statistical and processed based models".

l. 41: could you delete "uncertainty in", or replace with "estimates of"? (Isn't it the timing itself that's important, not the uncertainty in the timing?)

We deleted "uncertainty in".

l. 81: fix "quanity".

Fixed.

l. 115: why these dates?

These dates were used in accordance with Crimmins (2017), who used them prior, as budburst is unexpected to be first seen past DOY 172 (June 21) and DOY 213 (Aug. 1) for flowers.

l. 127: phenophase (singular)

Fixed.

ll. 210-211. I don't understand this. Why do some seemingly strong relationships (e.g., "naive" panel of Fig. 2; also some panels of Fig. 3) yield negative R-squared values? This probably just reflects my ignorance... but it's likely that if I'm mystified, other readers will be, too! Please explain more clearly how the ¡0 R-squareds arise.

These values are coefficients of determination which compare the predictions from the model to the observed values. If there is a strong bias in the predic-

tions this can lead to negative values even if the observations and predictions are correlated. Specifically, in cases where choosing the mean value of the observations provides a better prediction than the model these numbers will be negative. While this is not possible in simple regression without held out data (because the model by definition does at least as well as the mean) it is possible for other kinds of models and in cases where the model is not fit to the data it is evaluated on. We have attempted to clarify this in the text.

I found the discussion quite insightful, although there are some additional points that should be addressed and a few areas where clarity could be improved (see Reviewer 1's comments).

Thank you. Please see the replies to Reviewer 1.

Also: l. 289: what do you mean by "internal phenology"? internal physiological processes?

Specifically we meant the date of endodormancy break, but had not used this term prior and did not want to confuse readers unfamiliar with it. We have changed this text to be more descriptive of the Chuine et al. 2016 study results.

l. 294. "warrant caution" -¿ "indicate that caution is warranted"

Fixed.

l. 305: "the" is repeated.

Fixed.

Fig. 4: the labels on the right-hand side ("Metric 1/2") are uninformative. Please use more intuitive labels, e.g. "LTER data/NPN data".

We replaced the labels to reflect the data evaluated. We also adjusted the methods section slightly to hopefully make this more clear.

2 Reviewer 1 Comments to Author:

2.0.1 Comments to the Author

Comparison of large-scale citizen science data and long-term study data for phenology modeling

This is a very practical study and the evaluation of the approach is clear and pretty straightforward to follow. The authors might im-

prove readability a bit by either adding subheads into the Methods & Results or repeatedly using the same terms to refer to the various steps (eg, cross-validation).

After some discussion, we decided not to add subheadings as both sections are relatively short, and each have a single paragraph dedicated to each analysis step. Instead we have made the Analysis introduction paragraph lay out the three primary steps, and also adjusted the wording to be more consistent.

The most critical thing that I see thats lacking is that there is no consideration of how decisions made regarding which NPN data to include in the analysis might be impacting model performance. The authors allowed observations with up to 30d between the last reported no and the first reported yes and chose the midpoint. This could still lead to up to 15d error in the assumed date of yes and the true date of yes; this variance may in many cases be larger than the signal in the model. Though actually addressing this sensitivity is probably beyond the scope of the analysis, the potential and implications for this error should be discussed explicitly in the paper.

We tested the sensitivity of this cutoff time by repeating the analysis with a 15 day cutoff. This change did not influence our results. Three species (flowers and budburst for Vaccinium parvifolium and budburst for Acer circinatum) were dropped from the 15 day analysis due to inadequate sample size. The overall sample size for the 15 day analysis was decreased by 10% (619 observations) and by 1-35% for individual species. We have added notes on this sensitivity analysis to the methods and results as well as reporting the results in new supplementary material.

The authors lay out two goals for this effort to determine inferences about biological processes driving phenology and to develop predictive phenology models. It may be just meusing the term inference in this way confused me.

This is a common usuage in the data science community and we are unclear how to better communicate this distinction between making predictions and understanding processes. To avoid confusion we have added a specific example to describe our usage of inference in the last introduction paragraph.

Regarding the predictions that can result from these different modelsthis is of strong interest to many stakeholder realms. It would be fantastic to see a bit more attention given in the discussion to recommendations for implementation. The authors talk about the tradeoffs between more complex formulations and simpler models and generally reasons these differences may come about. It would be great to also see a bit of more practical discussion along the lines

of, if you can tolerate error of 7-14d, then model X may meet your needs, and its fine to build the models with either NPN or LTER data. However, if your needs are Y, then youd be better off doing blah, blah would this be possible?

We have added additional text to the Discussion describing the best models from our analysis and pointed readers to the supplemental images for specific details of model performance. We have also updated supplemental images S1 and S2 (now S5 & S6) to better show model performance in the cross-dataset comparison, and also added an additional table (Table S2) which gives the overall model performance across all species and phenophases.

Thank you for your serious efforts to understand the USA-NPN data and to contribute to the collective understanding of this datasets potential as well as shortcomings. Im so excited to see this sort of work happen! Im happy to talk with you more regarding this manuscript as well as your other efforts, especially if any of these comments require clarification. Sincerely, Theresa Crimmins, theresa@usanpn.org

2.0.2 Smaller things

Please refer to the Network as USA National Phenology Network and USA-NPN because there are lots of other phenology networks in other countries

We have made this change throughout the text.

In several places data is treated as singular data are plural

Fixed.

L67: the citizen science program run by the USA-NPN that yields the data housed in the National Phenology Database is Natures Notebook Thank you for acknowledging the Natures Notebook contributors for the data!

You're welcome. We're happy to credit all data providers, as this study would not have been possible without them. We have adjusted this line of text to reflect that Nature's Notebook is the citizen science program run by USA-NPN.

Please see the USA-NPN data attribution policy for recommendations for how to properly cite the dataset: https://www.usanpn.org/terms#DataAttribution

We have added the correct citation for the dataset (USA National Phenology Network, 2017)

2.0.3 Abstract

be sure to indicate that the study is focused on the U.S.

This is now made clear by specifying that it's the USA National Phenology Network.

L26-29: models performed best when applied to the same data with which they were built you attribute this to scalethough other things could be at play here, such as local adaptation and species sensitivity to different forcing variables varying across the range.

Spatial scale is the primary difference between the two datasets used here, but Reviewer 1 is correct in that local adaptation or species sensitivity could be the mechanisms behind the differences seen at different scales. We have clarified this in the abstract as well as the 2nd paragraph of the discussion.

L33: you mention that the NPN dataset offers many species though this is a good point, it is actually irrelevant to the arguments that you are making in this paragraph, that is, if your comparison is focused on only a handful of species. Oh, now I see you make this point several times in the paperbut Im still struggling to see how its relevant. Is it that you could have built models with many more species in this study based on what NPN offers, but you were limited by what the LTER datasets had to offer?

Reviewer 1 is correct in that our analysis was limited by species common to the NPN and four LTER datasets. Large scale phenology studies typically employ as many species as are available in a given dataset. Thus we point out the large species pool of the USA-NPN dataset throughout as it is something other researchers should consider when designing their own analysis.

L34-36: this concluding sentence could be made stronger if you kept the focus on the findings of the present paper the strengths and limitations of the two datasets and their suitable applications.

We felt that the strengths and limits of these datasets required more attention in the abstract than could be managed in the final sentence, so we discussed them in the prior sentences primarily and instead looked forward in the final sentence.

2.0.4 Introduction

L77: Natures Notebook has 1,000s of volunteer participants (www.usanpn.org/data/dashboard)

Fixed.

L80: may also want to check out Feldman et al. (2018) https://doi.org/10.1007/s00484-018-1540-4 for further support on cit sci observer skill

Thank you. We have added this reference as there are very few studies looking at observer skill at this level of detail.

2.0.5 Methods

I think its pretty critical to provide your sample sizes for each of the species/phenophases under evaluation for each of the datasets when I read that you held out 20% of observations for evaluation, I was left wondering, was this 5 points or 50 or 500? You could include this info in Table 1

We have added details of the sample sizes to the supplement table S1.

What is the sampling frequency for the LTER sites? This is quite relevant because you chose to use 30d as your cut-off for NPN data - youll allow for up to 30d between the last reported no and the first reported yes. How does this compare to whats possible in the LTER datasets? It could definitely have an impact on the sensitivity of your results.

Three of the four LTER sites (Harard, Hubbard Brook, & H.J. Andrews) have very short sampling intervals (3-7 days). The Jornada LTER has a sample interval of 30 days. We have added these details to the manuscript. With the exception of the Jornada we feel the sampling interval of the LTER sites is very precise, and is one of the strengths of long term data that we point out throughout the manuscript. Given that, the issue of potentially 30 (or 15) days between sampling versus 7 is one of several points relating to volunteer based sampling we discuss in the manuscript.

We note that only 2 of the 38 comparisons are from the Jornada datasets, so it's larger sampling frequency does not affect the overall analysis done in this paper.

After reading about the phenology datasets, I expected to read about the temperature data used. I see now that its nested within the Modeling section the organization just surprised/confused me.

The phenological data are the crux of the study so we felt that they deserved special consideration, while the components used to generate the models could be kept within the modeling section. We will gladly change this if the editor feels it is necessary.

L138-140: Provide some references for your GDD calculation there are multiple ways to go about calculating GDDs

The exact calculations are unique to each model and are provided along with references in Table 2. We have clarified this in the text. We also added references for the GDD and Fixed GDD model.

L143: what is NCD?

This is the number of chill days (mean daily temperature less than 0 degrees C) since Jan. 1 in a given year. We have added the description to the text.

L145: please provide some references to back up your statement that these formulations are the most common

While we feel anecdotally that this statement is true, to our knowledge no formal analysis on model usage has been done. Thus we have removed this line.

L156: Im sure you extracted the daily temp at the lat/long of each observation location maybe just make that a bit clearer

We have clarified this.

I dont think I saw RMSEA RMSEB and RMSEC-RMSED referenced anywhere else in the paper

We removed these specific equations and replaced them with a better description of this analysis.

L196: what does scenario A refer to here?

This referred to the RMSE values derived from LTER Model predictions for held out LTER data. We rewrote this paragraph and changed the labels on Figure 4 to make this analysis more clear.

2.0.6 Results

L231-237: is it possible to quantify these comparisons in any way?

We have added a t-test to show the difference from 0 in these distributions.

2.0.7 Discussion

L244: are you referring to the temporal depth and/or frequency of sampling when you say intense?

We are referring to both of these as LTER sites typically have a long history of frequent and consistent sampling compared to USA-NPN sites.

263: limitations in the sampling of the NPN dataset what sort of limitations? Small sample size, frequency of sampling, or something else?

The frequency of sampling, in a given season and also between seasons, is the primary limit we consider here and we've added this clarification in the text.

L266-268: it would be helpful to add some references for local adaptation and differing sensitivity to phenological drivers across gradients

We added two references as well as some potential drivers here to highlight the potential variation in phenological requirements across different gradients.

L270-272: another option could be to develop local or regional models, especially if a national-scale prediction is not needed

We have briefly included this possibility, as there may indeed be circumstances (population conservation) where the national prediction is unneeded.

L273: define model identifiability and/or provide a reference? I think you are explaining it in the subsequent sentence if so, just make this more explicit.

We have made this clearer by switching to a colon instead of a period between the two sentences.

L276: what do you mean by these analyses?

We have clarified this by changing it to "..using the analyses in the current study".

L293: not sure what youre referring to when you say this previous research

We were referencing the two studies cited in the same paragraph (Chuine et al. 2016 and Basler 2016). We have changed this to "In combination with the aforementioned studies,..."

L337: there are many recent studies that have brought together these various datsets; would be worth adding a couple of references

We have added two additional citations which perform analysis using different phenological datasets.

3 Reviewer 2 Comments to Author:

3.0.1 Comments to the Author

Summary:

While I do think that this manuscript should be ultimately acceptable for publication at Ecology, there are a few major/minor comments and suggestions for which I would like the authors to consider.

Major Comments and Suggestions:

First, for NPN species with larger pools/spatial extents of observations, I wonder if would be useful to calibrate the models using subsets somehow stratified by climate and/or latitude. For example, train the models using only warm region (or southern) red maple and test on cold region maples (or vice versa). Perhaps this could further address the issue of model identifiability (as discussed in Lines 259-277). While the oft-mentioned Basler (2016) study certainly had nice spatiotemporal coverage, it did not have nearly the latitudinal/climate variability that NPN affords.

We agree that this approach is a valuable next step, but ultimately decided that doing this properly would require a follow up paper rather than an additional analysis in this one. That said, we are doubtful that it would address the issue of identifiability. The Harvard and Hubbard Brook LTER share 3 species in common and are relatively close (˜150km), yet their parameter estimates varied widely for most models (see images S4-S5 in the original text, now S8-S9 in the revised text).

Second, given the length of time series of the LTER datasets, I think it would be useful to further examine the performance of models based on other criteria such as the amount of interannual variability explained (correlation or R-square) or bias during anomalously warm or cold years. Perhaps the latter might be outside of the overall realm of the paper, but I think it would be informative to know which models capture this type of information which matters in a global change context.

See the note to Reviewer 1 about general model recommendations. We have updated the Supplemental images S1 and S2 to included Pearson's correlation in addition to the RMSE for all dataset inter comparisons. We decided that more expansive analyses were beyond the scope of this paper, but are hopeful that the open and reproducible nature of the code and data for this project will

allow others (or our future selves) to conduct the important follow up analyses suggested by the reviewer.

3.0.2 Minor Comments and Suggestions:

Line 281: Please verify what is T* (I assume base temperature)

This is correct. We have changed T* to T_base throughout as it is more descriptive.