Reviewer 1 – comments:

This is an interesting case study comparing the differences in GDD accumulations necessary to trigger leaf-out driven by degree of urbanization, provenance, and microclimate effects, as well as the type of sensor used to measure temperatures. Overall, I really appreciate the study, it makes a valuable contribution to the literature regarding applying GDD thresholds in forecasting phenology. My biggest suggestion is that the authors be cautious about how the findings are interpreted, since the case study is limited to two locations in a small region. And second, perhaps most importantly: the nuances brought forward in this study matter at a small spatial scale and definitely should be considered in local forecasting efforts. However, I would love to see the addition of a bit of language acknowledging that at large spatial scales, simple approaches like GDD thresholds still perform reasonably well, are easy to implement, and easy to understand.

We thank the reviewer for this positive feedback and for helping us improve the manuscript. Based on this review we have updated the language and interpretation throughout. We believe our revised manuscript better addresses our questions and will hopefully be more accessible to a wider audience of readers.

In addition to small tweaks throughout we edited the abstract and discussion to clearly outline these limitations. In the abstract, we now specify that we are referring to one studied urban site and forest, emphasized our findings' relevance to 'local-scale forecasts' and stressed the need to examine impacts at larger spatial. In the Discussion we have added a new paragraph early on that tackles this issue prominently (line 274 - line 279), which reads:

> These findings have relevance to local-scale forecasting. Land managers of temperature forests working at local to smaller regional scales may want to consider how microclimatic, species-level effects and GDD accuracy issues could impact their planning with climate change. As our climatic and species-level data were gathered at a relatively small spatial scale, extending these findings to larger spatial scales (e.g., with remote sensing etc.) to test whether simple approaches like GDD thresholds may still perform reasonably well is an important next step.

To note: I did not install and explore the R Shiny app mentioned on L196; I am wondering whether the authors might consider hosting the app on a website, to enable easier access by readers In general, I really like this study, and I do not have major concerns about the study design or presentation. My comments are mainly in support of making the manuscript as accessible as possible to a range of readers.

Thank you for the idea to host the app on a website! We have now published the Shiny application to site: `https://impactforecasting.shinyapps.io/ecomodels_GDD/` and the link has been updated in the manuscript. This was a very helpful first step to making the manuscript more accessible.

L34: resources and forest management aren't ecosystem services; this sentence might work better if the second half focused on ecosystem functioning, since you mention ecosystem services later in this paragraph

This was a great catch and we have since modified the first paragraph of the introduction (line 34 - line 38).

L52: 'threshold temperature' - I most often see this referred to as the 'base temperature', and 'threshold' used to refer to the amount of GDDs that must be accumulated to trigger the event

Thank you for finding this. We agree that how we were 'threshold temperature' was not appropriate, nor consistent throughout the manuscript. We have updated the manuscript to use 'base temperature' when describing the minimum temperature for GDDs to accumulate.

L53: 'sums' - 'accumulates' might be more accurate - or maybe, 'sums these temperatures each day and accumulates them until. . . '

We thank the reviewer for this tweak in language and have updated the manuscript to be more specific.

L54: might be nice to have a reference or two to back up your statement that different species have different GDD requirements to leaf-out, flowering, etc

We appreciate the reviewers comment and have added in references to support this statement.

L56: 'threshold' - here it appears you are using 'threshold' to refer to the total GDDs necessarily to accumulate to trigger the event - this is how I typically see 'threshold' used in this context - I point this out because it is distinct from using 'threshold' to mean 'base temperature', as mentioned above

Thank you for pointing this out and further clarifying. We agree with the reviewer and have updated the language throughout accordingly.

L60: 'constant' - here I might add 'across individual plants and locations' or similar text, to help the reader follow

We thank the reviewer for this additional context and have added it for further clarification.

L61-62: you state that due to plasticity, individual plants exposed to different conditions could leaf out at different times. . . the same individual could leaf out at different times but actually at the same GDD threshold, depending on the underlying conditions. I think what you mean to say here is that due to plasticity, the same individual could leaf-out at different GDD accumulations.

Thank you for this more accurate wording. We have updated the sentence in the manuscript and further adjusted the following sentences so that our meaning is clearer.

L64: suggest to add 'to trigger a phenological event' to the end of this sentence

We agree with the reviewer and have adjusted the language to be more specific and to also include phenologies other than just leafout.

L64-69: I don't disagree with this text, but it doesn't seem to fully fit with the set-up of the Intro and the argument that GDD thresholds may vary across space. I suggest to drop this paragraph, or shift the focus on how the timing of leaf-out may be different in urban areas to the fact that GDD thresholds may be lower in urban areas and why that might be.

We thank the reviewer for their insight and agree the paragraph feels out of place. We have

updated the language to better fit the rest of the Introduction (line 62 - line 66):

> Climate helps determine the role of chilling and photoperiod—and, thus the required GDD (1). On a large scale climate gradients across space (i.e., latitudinal or continentality effects) and gradients due to anthropogenic impact may thus alter estimated GDD. Urbanization has led to the formation of urban heat islands, which can affect plant phenology and lead to earlier spring leafout—and lower GDD thresholds—due to stronger chilling effects (2).

L81: why use 'vineyard' here?

Thank you for catching this - we have changed 'vineyard' to 'site'.

L80-86: I think this paragraph would fit better above the preceding paragraph - that is, ahead of the paragraph that currently starts on L70 - talk about climate at large/small scales and then move on to provenance

We thank the reviewer for this insight and agree it flows much better by starting at a larger scale and moving to smaller scales. We have updated the paragraphs to better match that progression.

L94: though see recent work by Meng, which suggests that artificial light can counteract the daylength limitation (Meng et al. 2022 PNAS Nexus)

We have added this additional information to our list of hypotheses. Thank you for keeping us updated on the literature.

L101: I think it would be best to indicate in a separate sentence that there are multiple provenances represented at the urban site, rather than as a detail tucked in the middle of the sentence

We thank the reviewer for this comment and have updated the paragraph to better elucidate that the urban site has multiple provenances, whereas the rural site only has one.

L102: I think it should be 'an urban arboretum' rather than 'arboreta'

Yes, this is a great catch, thank you.

L101-103: can you phrase hypothesis 1 more clearly? It is difficult to follow right now... perhaps something like this: GDDs required to trigger leaf-out will be greater for trees at the urban site compared to the rural site, due to lower chill accumulation.

Thank you for pointing this out. We agree the original wording was confusing and we have updated the hypothesis to be easier to follow.

L110: how many species were you able to evaluate in common across the two sites? You mention >3K taxa at Arnold Arb, but it's not clear if you evaluate all of those spp (I'm guessing not) - oh, is it the 15 spp you list in L120-123? If so, this is more important to focus on than the spp at Arnold Arb, I think

We thank the reviewer for this comment and have attempted to update the paragraphs to be more clear. The species are not identical across both sites but there is some species overlap. We have listed out each species observed at both sites and have removed the >3k taxa reference for more clarity.

L116-117: mention/reference Denny et al. (2014) as methods/protocols used to collect phenology observations at Arnold Arb, since you give method info for John O'Keefe's methods (L127)? `https://doi.org/10.1007/s00484-014-0789-5`

Thank you for this citation, we have included here.

L134: for clarity, perhaps add 'at the Harvard Forest site' at the end of this sentence

We have added more information here for clarity.

L147-149: please phrase with parallel language; please also ensure the phrasing of these clearly match the hypotheses stated at the end of the Intro

We thank the reviewer for being mindful of consistency. We have updated the language to better match the tweaks mentioned above.

L155: To minimize potential confusion, it might be good to indicate you assumed each species required a different GDD accumulation or threshold (rather than just GDD)

Thank you for this added context, we have updated the sentence accordingly.

L156: the 'modeled climate data' step is fuzzy to me here, is it possible to be more explicit here?

Yes, the word 'modeled' is probably inappropriate here. We have updated the text to say 'simulated climate data' and have added more information on those methods.

L158: again to minimize potential confusion, I suggest to phrase as 'To test that plants located in urban locales require a larger heat accumulation [or the accumulation of more GDDs] to XXX...'

Thank you, we have modified the text to better align with previous updates and suggestions.

L164: 'different base temperature thresholds' - do you mean different base temperatures?

Yes, that is what we mean. We have removed the word 'threshold' for clarity.

L167: perhaps define sigma here for readers not familiar with Bayesian approaches

Thank you, this is really helpful reminder. We have added more context.

L205: isn't this two simple effects?

Yes! We have updated the sentence.

L248: 'GDD values' - thresholds or accumulations?

Thank you, yes this should be GDD thresholds.

Fig S5 actually seems like it could be valuable to have in the main text, since it illustrates the findings from your hypotheses

We agree with the reviewer and have moved the figure to the main document.

L265: I would suggest to enhance this sentence a bit for clarity's sake, something like: 'Our case study approach, which compared GDD requirements necessary to trigger leaf-out between an urban arboretum and a rural forested site,...'

We thank the reviewer for this added context, we have updated the paragraph for clarity.

The finding that earlier-season species will be estimated with less accuracy is consistent with patterns reflected in the literature from observational studies; it might be nice to tie that literature in here

Thank you for this suggestion, we have added a reference here.

L296-297: this may be just me, but I don't follow how you could accumulate less chill at colder sites

We thank the reviewer for this suggestion and agree that the phrasing is confusing. We have attempted to add more context for clarity and explanation.

L305-322: might be worth a bit more discussion of how different methods of temp logging perform - there most certainly is a big body of literature on this. I imagine HOBOs are more prone to error; it might be good in the future to have multiple HOBO data loggers out at the same site to estimate variation due to the equipment

Thank you very much for all of your comments and suggestions. We have added more information and suggestions for future studies to include multiple types of loggers.

Reviewer 2 – comments:

Combine simulations, observations from an urban and a rural site, the authors assessed GDD requirement of budburst across species, space and methods, and found that the variations of GDD space, species, which provides new sights into the modelling of spring phenology. This study is interesting and well written. Please find below the comments:

Thank you for your positive feedback and helpful comments. We have done our best to address all of your suggestions to further improve our manuscript.

In the present study, the authors point out the variations of GDD, which may be caused by the shifts of chilling and/or photoperiod as the authors indicated in the discussion. The authors have the winter temperature dataset, why add an analyzation of chilling variations?

We thank the reviewer for this suggestion and have carefully considered adding analyses around chilling, but have decided not to for several major reasons. First, we feel this is

already a dense manuscript and we are concerned adding in analyses of chilling would be too much information for readers to interpret. Second (and relatedly), our current understanding of chilling is limited, but suggests a process far more complicated than the thermal sum model of GDD, where most spring temperatures above the base temperature accumulate equally. In contrast, models of chilling accumulation suggest a highly non-linear relationship where some temperatures are optimal, and others contribute very little—but what exactly those temperatures are is not known and very likely varies by species. Including any thoughtful analysis of chilling would thus require the addition of many various models, none of which we could clearly point to as most accurate. Based on this, we have decided that adding in chilling analyses would be overwhelming and take away from the overall message for focusing on GDD thresholds. We discuss, however, the complexity of chilling models some on lines line 292 - line 302.

Lines 33-37: More relevant references should be cited here.

We thank the reviwer for this suggestion and have added references to this sentence.

Line 64-6689-90: The responses of leaf unfolding to temperature and photoperiod are different across latitudinal gradients, which would change the requirement for GDD (Wu et al., 2022, Frontiers in plant science).

Thank you for keeping us up to date on the literature. We have updated the references.

Line 177-180: there were many methods, and why do you selected 0 °C as the threshold? Have you tried other methods, which would greatly increase the reliability.

We have added more language here for clarity to describe our simulations with varying baseline temperatures. Thank you for pointing this out.

Fig.5 Differences in GDD among the combinations should be provide in the figure.

We thank the reviewer for this helpful comment and have added this information into the caption of the figure.

References

[1] Bonhomme, R. *European Journal of Agronomy* **13**(1), 1–10 (2000). https://doi.org/10.1016/S1161-0301(00)00058-7.

[2] Meng, L., Mao, J., Zhou, Y., Richardson, A. D., Lee, X., Thornton, P. E., Ricciuto, D. M., Li, X., Dai, Y., Shi, X., and et al. *Proceedings of the National Academy of Sciences* **117**(8), 4228–4233 Feb (2020). 10.1073/pnas.1911117117.