Comments from the editor

I have received two reviews of your manuscript and have read it carefully myself. While both reviewers believe that the topic investigated is interesting and timely, both also identify some critical problems that prevent my acceptance of the manuscript in its present form. Most of these concern statistical issues, particularly which variables are used as dependent and which as independent. In addition, one reviewer questions the rationale for some of the hypotheses you present.

We thank the editor, Dr. Rausher, for feedback on our manuscript and for the opportunity to revise and resubmit this work. In our updated submission we have addressed the editor's concerns by developed a continuous index to quantify flower-leaf sequences and re-done all of our analyses using this metric. In this broad re-analysis of our data we also developed several sensitivity analyses; a) substituting our dependent and independent variables in our models testing the relationship between hysteranthy and trait predictors and b) ran models with and without day of year as predictor as suggested by Reviewer 2, and c) added the year of observation to our main model as requested by both reviewers to account for the influence of climate change and uneven sampling. This effort resulted in new and improved figures in the main manuscript (Figs. 1b-4), and several additional tables and figures in the Supporting Information. Importantly, even with all of these adjustments to our analyses, our results in support of the two hypotheses we tested in the study did not change. We detail these changes further in our responses the reviewers' comments below.

Referee: 1

Comments to the Author

Summary

This manuscript quantified flower-leaf sequence variation in the American plums, a clade of insect-pollinated species, using herbaria specimens and Bayesian hierarchical modeling and revealed hysteranthy was associated with aridity and smaller floral displays.

General Comments

Overall, I believe this paper addresses a significant and timely topic, and the analyses conducted were intriguing. However, it is crucial to provide additional information as the current explanation is inadequate and challenging to comprehend.

We thank the reviewer for their efforts reviewing our manuscript and are pleased that they found our topic interesting and relevant. We have made numerous change to our manuscript to address the reviewer's concerns including converting our hysteranthy index to a continuous scale and adding several supplemental tables and more text to clarify our framing, methods and scope of inference from this study. We feel that the reviewer's comments have helped us generate an improved manuscript that expresses our analyses and their implications more clearly. We detail the changes we've made below.

Concerns

I feel that the title, insect-pollinated temperate trees, is not appropriate for this study as only Prunus species was included in this study.

We have modified the title to better reflect our study system. The title of the manuscript is now:

> Ecological drivers of flower-leaf sequences: aridity and pollination success select for flowering-first in the American Plums

l.15 Many trees in temperate forests produce flowers before their leaves emerge. ll.36-37 This flowering-first phenological sequence, known as hysteranthy, proteranthy or precocious flowering, is particularly common in temperate deciduous forests around the globe (Rathcke Lacey, 1985). Those sentences are misleading as flowerleaf sequence is a typical pattern for temperate trees. According to Buonaiuto et al. (2021), leaf out

before flowering is a typical model of plant life history.

Thanks for highlighting this. We have adjusted these statements to be more precise. These lines now read:

> This flowering-first phenological sequence, known as hysteranthy, proteranthy or precocious flowering, is apparent in temperate deciduous forests around the globe.

ll.110-114 Please clarify the collected years of sampled specimens. Are those specimens collected in a year or over the years? It may be difficult to pool the data if some specimens collected more than 50 years ago due to phenological sifts influenced by global warming.

We appreciate this point. Our samples span many decades of collection (1844-2020), which is necessary to generate a large enough sample size to give our analyses statistical power. We have added this information at line 108.

Both reviewers raised this issue of the effects of global warming on phenology. In our new version of the manuscript we addressed this by including year of observation as a co-variate in our main model and found year did not have a significant effect on patterns of hysteranthy. This means that while climate change may be affecting both flower and leaf phenology, it has not yet significantly affecting their relative timing, a point we now highlight in lines 340-346. We detail methods for this updated version of our model in lines 128-131

ll.114-115 "In total, we evaluated the phenology of 2521 specimens, but only specimens with visible flowers were included in this analysis (n=1009)." Doing some simple math, 16 species x 200 specimens/species is 3200 specimens in total. Why are there less than 3200 specimens evaluated? Also, please provide the numbers of the analysis-included specimens for each species.

We thank the reviewer for catching this. The total sample size is less that 3200 because the database contained significantly fewer than 200 collected specimens for several species. We have added this point to our Methods (lines 112-117), and now include a supplemental table with the sample sizes for each species (Tab. S1). A major reason we chose models that partially-pool coefficient estimates on species is to account for this lack of balance among species in our dataset.

l.113 Authors are encouraged to add a little more explanations related to BBCH scale so that readers can understand it without referring to the original references.

We agree this would make our text easier to follow and have added a more concrete explanation of this scale to our Methods at line 111 .

ll.134-144 I believe this is an important passage that explains the criteria for determining whether each species is hysteranthous, but it took me some time to understand it. The mixing of similar numerical values such as the percentage of flowering seasonal quantile, the percentage of probability distribution, the BBCH scale, and the aggregated index made it challenging to comprehend. Could you create a diagram using several species as examples to facilitate a more intuitive understanding?

We can see why the development of our initial index was difficult to follow and, in retrospect, we agree that building our index of of seasonal quantiles wasn't the most straightforward approach. In the new version of the paper, we instead present a continuous index based on the likelihood a species' flowers appear before leaf development through across their whole flowering season. We found that this approach is equally robust and substantially more clear and useful.

This index is based on the same underlying data and statistical model as our previous one, and critically, it yields similar results. The major difference is that we predict hysteranthy likelihood across the whole flowering season for each species, rather than just at each major quantile.

We think the these changes suggested by the reviewer have made our methods much easier to follower, even without an additional visual aid. However if the reviewer still feels that a conceptual figure would enhance the clarity of the study, we would be have to discuss this further.

Also, it is hard to understand the followings: -Why is it needed the 0%, 25% 50% and 75% quantiles of their flowering period? -Does probability distribution mean flowering-probability distribution?

With our new continuous index described above, we no longer use these quantiles. We also feel that this improved index circumvents the confusion between the statistical distributions and quantification of flowering probability that the reviewer highlights here.

According to Finn et al. (2007), BBCH 00 is Dormancy: buds closed and covered by scales and 09 is Buds show green tips. I think 01 is suitable for bud development and 07 is suitable for bud break.

We have updated our verbal descriptions of the vegetative BBCH stages to more closely align with how they are described in Finn *et al.* (2007). These changes can be found at lines 148-158.

l. 147 Authors are encouraged to add a little more explanations related to PDSI so that readers can understand it without referring to the original references.

We agree, and have added an explanation of PDSI to the Methods (lines 170-173).

ll.197-206 Please mention the topology of Prunus species on phylogenetic tree in Fig.1b because it is also one of the results of this study.

We now include reference to the original tree topology in the Figure 1b and reference the figure in our Results at line 234.

ll.167-173 Please indicate the number of species included to the analysis.

We have added the sample size to our Methods section (line 204).

Fig.2-4 Please indicate sample sizes of each analysis.

We have added this information to the figure captions.

Fig.2 Fig. S1 I think the flowering season and length are different among those species. Is there any trend such as hysteranthous species bloom in early spring but serathous species bloom in late spring or summer?

This is an interesting question. To address this question adequately we would need additional data about differences in species' ranges sizes, and spatial modeling to account for sampling bias in dataset that are beyond the scope of our current models. We do feeling this is an important point, and have added text highlighting this potentially interesting relationship between hysteranthy and flowering duration/time of flowering at line 125.

Fig.4b Readers with red-green color blindness may have difficulty distinguishing between four colors.

We appreciate you identifying this and have changed our color scheme in all colored figures.

Citation: please check citation style. -l.120 (de Villemeruil P. Nakagawa, 2014)
-and and are mixed.
-Order lists of references in date order (oldest first).

Thanks. We have made these formatting changes.

Referee: 2

Comments to the Author This paper examines the adaptive value of flowering before leaf-out in trees, using data on variation across species from herbarium specimens. and analyses accounting for phylogenetic relationships.

The question addressed is important and interesting, and the approach taken appears basically sound. I do, however, have several concerns with the manuscript.

We thank the Reviewer for providing important feedback on our manuscript are please that they found the

topic interesting, and our general approach to be sound. We appreciate the Reviewer's concerns, and have done our best to address them in the revised version of the manuscript. We detail the changes we've made below.

The two basic ways in which selection might favor hysteranthy are correlational selection, and independent but differential selection on timing of flowering and timing of leaf-out. In the introduction, the authors touch upon this difference, but the reasoning in the manuscript appears a bit confused as Fruit maturation hypothesis, similar to the first two hypotheses, is treated as an example of correlational selection. However, the mechanisms supposed to explain hysteranthy through this mechanism does not involve correlational selection, and thus seems to fall within the category null explanations discussed in the following paragraph.

Overall, the role of fruit size as a third hypothesis to be tested is a bit unclear. While it is brought up as a separate hypothesis in the introduction, very little is said about it in subsequent sections.

We appreciate the reviewer's clear summary of the different ways selection and operate on flower leaf sequence, and agree that the fruit maturation hypothesis fits more broadly within the category of "independent but differential selection". In our revised manuscript we now mention this hypothesis as an example of our null hypotheses (line 78), and have removed our tests of it from our analyses.

The insect visibility hypothesis, as presented, appears a bit simplistic as the argument basically assumes that pollen limitation is equal among species and environments. However, if some species occur in environments that are associated with more severe pollen limitation, then you might expect that these species experience strong selection for increased visibility both through selection for an increased degree of hysteranthy, and through selection for larger floral displays. Responses to this selection would then result in a positive rather than a negative correlation.

We agree that the insect visibility hypothesis feels overly simplistic. We feel this is, in part, because it has not been well developed in the literature, and have added text to advance it further. Based on the reviewer's insights, we now state the possibilities for both the positive and negative associations between hysteranthy and floral display size in our revised manuscript (lines 69-76).

One major problem I have with the analyses used to test the hypotheses, is that the statistical models used do not seem to correspond to the predicted causal relationships. Currently the models examine effects hysteranthy on drought and flower size respectively. However, the logic way to construct models seem to be to examine effects of drought on hysteranthy rather than effects of hysteranthy on drought. It is thus difficult to understand why the authors use the models they do, and no motivation for the chosen model structure is provided.

We thank the reviewer for this point— we agree that we did not present clear justification for some of the modeling choices we made in our original submission. In the new manuscript, we now include several new analyses more similar to the ones suggested by the reviewer and add text to clarify our modeling choices. We address detail these changes below.

Moreover, using a model with hysteranthy as the dependent variable, rather than the current models, would also allow assessing the simultaneous effects of drought and flower size on hysteranthy, something that is not done currently. This is essential as there, as pointed out by the authors, are very good reason that drought and hysteranthy are correlated. Models examining the effects of both traits are therefore essential to disentangle the effects of each trait. In fact, you might expect drought to affect hysteranthy both directly and indirectly through effects on flower size. Anyway, I think that statistical models that better reflect causal relationships, and that examine the effects of drought and flower size simultaneously are necessary.

We agree with the Reviewer that our ideal model would include both traits as a predictors and their interactions to better understand their additive and interactive relationship to flower-leaf sequences. However, our data does not proved a straight-forward way to do this as our hysteranthy index, PDSI and petal size variables were measured on different individuals and have different sample sizes (FLS: 1000, PDSI: 2305 Petal: 2757).

We now explicitly make this point out in lines 175-lines 177 and have executed several new analyses to address this. First, to capture the additive and interactive effects of our predictors on hysteranthy as the reviewer suggested, we developed a model where we use species-level means of PDSI and petal size and test their relationship to our hysteranthy index using a beta regression framework. This is the analysis we now present in our main text and Fig. 3.

This approach captures the important interaction between the predictors the reviewer's comments raised, but the tradeoff is that it cannot account for the within-species variation in the environmental/morphological traits and their phylogenetic structure. For this reason, we also include versions of our original analyses in which we model the relationship between hysteranthy and each trait separately, which allows us to interchange the placement of the dependent and independent variable and used phylogenetic mixed models. We now include these complementary analyses in Fig. S3, and as Extended Methods in our Supporting Information.

We found that the interaction between mean aridity and mean petal size was not a significant driver of hysteranthy in the American Plums, making the qualitative inference from these two approaches similar, with strong associations between hysteranthy and aridity with both modeling frameworks, and weaker associations between hysteranthy and petal size as seen in the comparison between Fig. 3b and Fig. S3.

On the other hand, I find it highly questionable to include day of observation as a covariate in the models. This is because, as stated by the authors, hysteranthy co-varies with flowering time. In fact, selection for hysteranthy is likely to largely occur through selection for earlier flowering. Adjusting for day of observation will thus statistically remove some on the variation in hysteranthy and bias the results. At the least, I would like to see analyses both with and without day of flowering as a covariate in order to be able to judge the effects of including it.

For the updated version of this manuscript, we have followed the Reviewer's suggestion and produced analyses both with and without day of year as a co-variate (see lines 161-164). While these differences do change the hysteranthy likelihood estimates for some species, the relative patterns among species do not change greatly, nor do the relationships between hysteranthy and the trait predictors. We have included the results from the model without day of year as a co-variate, in the Supporting Information for comparison (Tab. S2,Tab. S3,Fig. S2).

We have chosen to keep the model with day of year as a predictor as our main analysis because we feel it a biological relevant explanatory variable that also helps control from temporal observer bias in herbaria records. We have added a sentence explaining this to our Methods section (line 126) and include histograms of observations across the season for each species in the Supporting Information (Fig. S1). Unlike in well-designed observational studies or experiments, hebaria specimen collection dates are not consistent across the season, and, therefor, trying to estimate flower-leaf sequences with accounting fore the unevenness of observations within season could seriously misrepresent the estimates.

I also do not understand why the authors used the approach they did to generate a five-level index rather than using continuous functions and continuous values of hysteranthy. The methods used seem to imply an unnecessary loss of information through categorizing a continuous variable, without stating any reason for doing this.

This concern was shared by Reviewer 1, and we agree that developing this five level index masked important variation that could be captured in a continuous index. In this version, we now provide a continuous index (detailed in lines 148-158), and redone all of our analyses with this metric.

The authors in several places stress that the flower-sequence is likely to change as a result on climate change. However, climate and phenology has already changed considerably, meaning that herbarium specimens for the same species collected during different periods are likely to differ both with regards to absolute and relative phenology. Thus, if collection dates for herbarium specimens are unevenly distributed over the study period, or differ among species, then this might constitute a significant problem for the analyses. It is therefore a major short-coming that the manuscript does not contain any information about how collection dates were dis-

tributed, or even during what time period the used specimens were collected. I think that this key information must be provided, and that analyses need to account for differences in collection date.

This issues was also raised by Reviewer 1. We have added information about collection dates to our Methods (line 108), and now account for collection dates in our model. As we detailed above in our response to Reviewer 1, we have included year of sample as a co-variate in out main model following convention from phenological change studies using 1980 as a hinge point (detailed in lines 128-131 of the revised manuscript). Sample year did not have a significant effect on patterns of hysteranthy (Fig. 2a), which is consistent with recent findings that the interval between these phases has remained relatively stable for most species in the face of recent climate change (Guo *et al.*, 2023).

In several places, the authors refer to the importance for the current study to understand the effects of ongoing global change. For example, on lines 56-57 they state that the study .. offers insights into how shifting flower-leaf sequences may impact species demography and species distributions as climate continues to change. Although, I agree that almost all basic knowledge about the biology of species will be helpful in this respect, I do not see that this study is particularly important in this respect. I thus suggest to down-tune this type of arguments and rely on the importance of this study for our basic understanding of hysteranthy.

We appreciate this point, and agree that this study may have more relevance to understanding the basic ecology and evolution of phenological sequences than the implications for climate change. We have followed the Reviewers recommendation and shifted the context of our study away from climate change in several places our Introduction and Discussion seconds. We now only discuss climate change in the context of our model co-variate (see above) in lines 340-346, and include the new citation from Guo *et al.* (2023) suggesting the stability of flower-leaf sequences we observed in our data is consistent with wider observations.

Other comments, in order of appearance in the manuscript:

Abstract: It would be useful to indicate what hypotheses that were tested, as well as the general methods used to test them.

We have added this information the the Abstract.

Lines 35-36: This is not correct. Flowering before leaf-out occurs also in non-woody species.

This is an important point. We have amended this statement in line 36. It now reads:

> Woody perennials are among a subset of plant types with the unique ability to seasonally begin reproduction prior to vegetative growth.

Lines 37-41: Explain what kinds of functional significance you refer to here.

We have replace this statement with a more precise one: "can confer performance advantages...".

Lines 146-150: I assume that drought indices have changed over this considerable time period. What was the reason for using this time period, and for using the same time period for all records?

We agree that our decision to truncate the drought index at for this period was arbitrary. In the new version, we use the full record provided in our data source.

While we also agree it is certainly possibly that aridity levels at a given location have changed over this time period, we see no clear way coherently assign different evaluation periods for different records, or truncate specific records based on change at a location. While this could be possible achieved with extended spatio-temporal modeling, we do not have any baseline information about how long a population has been present at a given locality, which we feel would make it difficult to fine-tune the time period we evaluate in a way that is ecological sound. If the reviewer has any suggestions for a straightforward way to do this we would certainly be open to trying it.

Lines 151-154: The unit relevant for attracting pollinators is probably not only the individual flower, but

entire inflorescences, or even trees. Would it thus be useful to examine effects of an attraction parameter that includes also, for example, the number of flowers per inflorescence?

We agree with this point, and in our analyses of the larger *Prunus* genus we use do indeed use number of flowers/inflorescence as a predictor. By contrast, the American plums all have solitary flowers, which is why we used petal length as our predictor. We have added text to discuss this nuance in our Discussion at lines 264-269.

Lines 178-185: Why did you use a different number of c categories for this analysis?

This choice was based on data availability. We have added a sentence to explicitly make this point at line 220, which explains that our flower-leaf sequence data for this group come from qualitative descriptions of which only four levels were available.

Lines 209-210: I do not really understand the meaning of this. You also seem to use different wording to explain the same effect in the three paragraphs of the results section.

With the incorporation of our new continuous flower-leaf sequence index, we have re-written the Results sections. We hope that the reviewer finds this presentation more clear.

Line 217: Why However?

This section is no longer included in our revised manuscript.

Lines 220-221: I suggest to skip sentences like these and let the results talk for themselves.

We have removed this sentence.

Line 222: .. through the predictions ??

As the reviewer suggested about we have worked to tighten the language of this summary paragraph of our Discussion, and agree the meaning of this statement was not clear in our original submission. We hope the stylistic changes we've made here (lines 271-279) read more clearly in this version.

Line 225: Strange subheading given that this is the topic of the entire paper?

We thank the reviewer for this point and have removed this subheading.

Line 228: Unclear exactly what trade-offs you refer to here.

We have elaborated on this in lines 273.

Lines 262-291: Here, I think that it would have been interesting to provide quantitative information about how much of the total variation in hysteranthy that was within vs. among species.

We agree with the Reviewer that this would be a very interesting contribution to our understanding of these phenological sequences. However, because our study was based on herbaria records that were collected non-systematically across space and time without any repeat sample, we feel that we cannot robustly partition this variation with the present data. We have added a paragraph clarifying this limitation and suggesting it as an area for further study at lines 336-339.