Reviewer comments are in italics; author responses are in plain text.

Referee: 1
Comments to the Author
This is an innovative and creative paper that will be noticed and read. The authors have revised their paper in a reasonable way in response to the comments. Many of the suggestions have been addressed. Figure 3 is now greatly improved. The results are better explained, and the problems with the approach are more directly acknowledged. My view is that this paper is very creative and even novel, but also problematic. Among the phenology community, this article will stimulate conversation about whether or not the results are convincing, and if there was a better way to have done this study, which is a good fate for any paper.

We are grateful to the reviewer for their continued attention on our manuscript, and appreciate that they see the potential for substantial impact. We are pleased that they feel that our latest set of revisions improved this work.

The authors resist examining variation within species, arguing that their sample sizes are too small. However, there are 200 specimens per species, which is very large sample size for these types of analyses, and much greater than the number of species being compared. Readers will wonder why this analysis is not being done.

We appreciate the reviewers point that our inter-specific analyses naturally raise questions about intra-specific variation in flower-leaf sequences, and have expanded text in our Discussion (lines 362-367) to clarify to readers why we do not feel we can adequately address these questions with our current approach.

We suspect that there are separate biological processes driving flower-leaf sequences on the different time scales of inter- and intra-specific adaptation, and that the physiological cues that control individual plasticity are not necessarily the same ones that influence intra-specific adaptation. To understand these factors, we'd need a model that is powerful enough to disentangle the effects of climate at each of the scales. We maintain that sampling bias below the species-level in our current data would over-extend such a model—we'd like to to point the reviewer to Table S1 that indicates our within-species phenology sampling size are not 200/species, but range from 17 to 118 obs/species, with 50% of our species having fewer than 80 observations. These diminishing sample sizes below the species-level are also spread unevenly across the geographic ranges of each species and across the 100+ year time series of our data, and accounting for these biases would require a modeling approach that is well beyond the scope of our current analyses. We appreciate the reviewer's interest in the drivers of variation below the species-level, and agree it would be of broad appeal to readers, but feel it would is best addressed with a study designed to robustly investigate variation at this scale.

The authors resist examining the effects of temperature on hysteranthy. This would be such a simple analysis to do and carry out, as temperature data is readily available. To show the value of this approach, just compare the distributions of the 5 fully hysteranthus species with the 7 partially hysteranthus species. 4 of the 5 fully hysteranthus specie are southern USA species, whereas only 2 of the 7 partially hysteranthus species are southern species. It is a virtual certainty that the March temperatures of the hysteranthus species are warmer than the March temperatures of the partial hyseteranthus species, and these differences explain as much of the variation as aridity. Readers will wonder why such a simple geographic and temperature analysis was not carried out.

Reading this comment, we realized that we did not make it clear enough in our previous response letter that we did in fact follow the reviewer's suggestion to examine the effects of temperature on hysteranthy.

Because the results of this analysis indicated that mean spring temperature did not explain species-level patterns of hysteranthy, and the model explained substantially less variation than our model with aridity, we did not add these analyses into our previous submission. We now realize—based on the editor and reviewer's points—that readers may be interested in seeing this relationship anyway, and have added details regarding this analysis to our Methods, Results and Discussion sections (lines 195-198,264-267,line 363), and now include a new table in our Supporting Information (Table S4).

Based on a more detailed exploration of species' distributions, we were not overly surprised by this lack of of a relationship between temperature and hysteranthy. Several of the species from the southern-most, warmest regions of the US (e.g., *P. texana* and *P. rivilaris*) have only intermediate likelihoods of being hysteranthy (index scores of .51, .44)—less than the likelihoods of the three northern most species *P. nigra*, *P. alleghaniensis* and P. americana (index scores of .55, .58, and .62 respectively). Other more northern species (e.g., *P. maritima* and *P. angustifolia*) are among the most likely to be hysteranthous (index scores of .68 and .76 respectively). This is all to say that relationship between temperature and hysteranthy is clearly more complex than broad biogeographical patterns of *northern vs. southern distrubutions*, and the results from our new analysis in Table S4 can now quantitatively address this—so we are grateful to the reviewer and editor for pushing us to include this in our paper.

For the aridity index, the authors use aridity in June-August. However these plants flower Feb-April. It would be much better to use an index that corresponds to the flowering period.

The June-August period is a standard climatological window developed by climatologists to capture variation in plan-relevant water stress, and—as we highlighted in our the previous letter—this is one of the only publicly available datasets that reconstructs historical aridity on a time-scale most relevant to the questions we are asking in our study.

We also feel June-August is the more biologically appropriate window than the spring window suggested by the reviewer, as selective pressures driving an adaptation to water stress would be strongest in the driest period of the season (June-August in our study system). As we discuss in lines 61 and 318, the water limitation hypothesis suggests hysteranthy evolved to partition hydraulic demand across the season—making it so that leaves and flowers are not both transpiring simultaneously and driving additive water loss during periods of water stress. This doe not indicate that aridity measured during the flowering period only should affect this pattern. **could skip: This is highlighted by the fact that the flowering of hysteranthous species in the dry tropics (where this hypothesis developed) is associated with the seasonal *recovery* of plant water-status (Franklin, 2016)—i.e., the expectation of the hypothesis is that flowering occurs outside of the driest seasonal window, but the dryness of that periods is the selective force on hysteranthy.** We hope this further clarifies our choice of metrics. **Q: Do I need to put some of this into the paper?**

On line 323, it says that the values of aridity range from -0.5 to 0.2, but figure 3 shows values from -1.5 to 1.5. What is the explanation?

The values in Figure 3 have be z-scored in order to compare the estimates for PDSI to those for petal size. We have adjusted the figure caption to make this more clear.

My intention in making these suggestions is to make the authors aware of these issues, even if they choose to keep the paper the way that it is.

We are grateful to the reviewer for their perspective on these issues and felt their comments challenged us to better understand our subject and make our findings and their implications more clear to readers.