Reviewer: 1

This manuscript presents an interesting common garden experiment that examines phenotypic differentiation in plant growth and physiological traits along a latitudinal transect, the relevance of plasticity for local adaptation, and the agents of selection associated with adaptation. Although such studies of local adaptation are not necessarily novel, this study has several original merits that warrant upmost consideration of the manuscript in *Evolution*. First, I am not aware of any other studies that examine the importance of climatic neighborhood vs. individual locality on phenotypic differentiation and find that climatic neighborhood is better predictor of phenotype than climate at an individual locality. Second, the finding that there is little variation in plasticity between populations and thus plasticity must not be important for local adaptation is interesting and not frequently tested. Third, very few plant studies have examined physiological tradeoffs at a population- scale across large spatial gradient despite the acknowledged importance of this these phenotypes. Finally, this manuscript does an impressive job considering different agents of selections – specifically few studies consider variance in climate factors among seasons as well as average climate factors as potential agents of selection. Despite these merits, there are a number of issues which need to be resolved to clarify the manuscript.

Major Comments

1. One of the major conclusions in this manuscript is that climate averages at a neighborhood scale is more highly associated with phenotypic variation than climate variation at a local scale. However, there are no data summaries or comparisons between these two scales in the results section and it is currently difficult for readers to compare the three alternative hypothesis that explain latitudinal clines (Line 65-68). Figure 4 only contains summary data from the neighborhood scale. It is difficult to believe that there are not significant or strong associations between traits and climate means/variance at the local scale given the associations at a neighborhood scale. I understand that the random forest regression is not ideal for visualizing comparisons between scales, but including a summary of univariate models examining associations for traits with both local and climate neighborhood scales of critical climate factors would be useful for addressing whether individual climatic factors explain more variance than latitude and whether neighborhood is a better predictor than local site.

We agree that summarizing this results intuitively is challenging and struggled with it ourselves. We have added analysis (Methods lines 230-7, Results lines 306-10) and a figure (Figure 5) showing that for both Climate-TraitPC1 and Climate-Latitude comparisons, correlations with neighbourhood climate are generally stronger than local climate, confirming what we observed using VSURF.

2. Analyzing the two treatments (temperature, drought) within a single fixed factor makes it difficult to tease apart how each treatment impacts the response variables. Table 3 suggests that the two treatments were treated as separate fixed factors; however, the methods (line 164-166) suggest that the treatments were treated as a single factor. Please clarify in the methods. A main text (or supplementary) figure that includes reaction norms for each family for each treatment with separate panels for each response trait would make it easier for the reader to eyeball the data/trends.

We have clarified the methods (lines 183-6) to explain that we tested for separate temperature and water effects, as well their interaction. This was what we did in previous manuscript, so this does not alter the results. We have also added a figure of reaction norms (Figure S5).

3. It looks like low and high elevation populations were sampled toward the southern end of the transect, while only low elevation populations were sampled at the northern end of the transect. Since high elevation populations typically have greater variance between years for precipitation and temperature, does this impact your conclusions on the importance of coefficients of variation or seasonality on precipitation or temperature? How about on neighborhood effects (62km in the Sierras is much different than 62km in the Willamette Valley).

The reviewer correctly points out an interesting pattern in the natural history and biogeography of this species: its elevation range compresses slightly from south to north, such that there are no very high elevation populations in northern California and Oregon. However, this pattern is relatively weak. There was no significant correlation between Latitude and Elevation among the 16 focal populations ($r = -0.25$, $P = 0.34$). This is not really different than for all 358 range-wide occurrences ($r = -0.17$, $P = 0.001$), which is significant because there are many more populations. Furthermore, we chose populations to maximize climatic variation, and thus we did sample some high elevation northern populations (e.g. Little Jamison, 39.7 north, 1603 m elevation, Wawona, 37.5 north, 1224 m elevation). Given our intentionally dispersed population sampling, we do not think the paucity of high elevation, northern populations is driving the latitudinal pattern in this case.

While it may be true that, in general, high elevation populations have greater variance between years, this does not appear to be a major factor in this system. For example, for range-wide precipitation of the wettest quarter (bio16), elevation explained ~0% of the interannual variation compared to 89% for latitude. For interannual variation in

4. Additional information on the life history of *E. cardinalis* would be useful given the conclusions on plasticity and the hypothesis on storage vs. growth in the discussion. Is there any idea of how long *E. cardinalis* lives or when most reproduction occurs (Year 1, Year 2... Year 8)? Do droughts actually have significant mortality impact in the field? A citation or personal observation for this question would be useful on Line 394-395.

We have added a description of latitudinal variation in vital rates and life history, based on annual surveys of permanent census plots along a latitudinal gradient (lines 437-9). We also explain that droughts have caused population declines and even local extinctions in recent years (lines 429-31).

Other comments

Lines 18-20 – The sentence beginning "Alternatively, genotype..." could be clearer. Specifically, the reader may be tripped up by confusion between genotype X environment vs. genotype X site interactions. Could be simpler to say that ability to respond to a particular environment or the magnitude of response to an environment could be adaptive.

Changed as suggested (line 21-3).

Line 22 – differences in 'photoperiod responses', not 'photoperiod'.

Changed as suggested (line 26).

Line 251 – please add a sentence describing the relationship between population latitude and traitPC1 value here.

We have added a sentence explaining the TraitPC1 cline (line 289-90).

Line 320-329 – The statement that strong fitness tradeoffs do not lead to adaptive differentiation seems paradoxical. If fitness tradeoffs are extremely strong between two different demes, spatially heterogeneous selection would generate strong patterns of local adaptation as migrant alleles would be strongly selected against in each deme. Please explain your logic here and/or add citations to Line 322.

We agree that this a subtle point that was not adequately explained and we have therefore clarified the theory (lines 365-9). Basically, the theory requires that habitats within demes are variable. If tradeoffs are strong, then selection favors habitat specialists regardless of how common the habitat is. Hence, selection remains constant, even as the frequency of habitats (wet versus dry soil, different host plants, etc.) changes over space.

Line 351-377 – the terminology used for drought resistance strategies is confusing and does not fit conventional definitions. Growing faster and high photosynthetic rates is not a drought tolerance strategy, it is a drought escape strategy (Ludlow 1989; Kooyers 2015). Two different predictions about enhanced drought resistance could have been made at the onset of the experiment. First, allocating more resources toward storage (a drought tolerance strategy) could be more common in areas of high or frequent drought stress. Second, rapid growth/reproduction could be common in areas of high drought stress (a drought escape strategy). These two hypotheses as well as reasons for observing one pattern over the other should to be more clear in this section.

We have clarified our terminology in the Discussion (lines 406-9). We think that our hypothesis could be described as a form of drought escape in which plants invest in growth and reproduction in the present season to mitigate the negative consequences of possible severe drought in subsequent seasons. This is not the classical definition of drought escape. Indeed, we do not have good evidence that populations from drier locations speed up time to flowering, despite growing faster, so this did not seem to fit the classic drought escape syndrome. Based on observations in the field, it does seem like drought speeds up flowering, perhaps indicating plasticity to escape drought, but we did not have evidence for variation in plasticity. Although many plants did flower by the end of the experiment, we did not record flowering time because we were trying to measure solely vegetative growth. Therefore, it remains an open question whether there might be variation in drought escape with respect to flowering, though we found no evidence for this in terms of growth rate. This could be a result of the fact much of the growth we measured occurred before severe drought set in (reaction norms show little effect of watering treatment on growth, but strong effects on mortality).

Line 423-425- Along this same line, including information on time to first flower and/or flowering throughout the experiment would provide support for the rapid growth in unpredictable environments hypothesis – after all, this strategy is only viable if reproduction occurs rapidly. I am a bit confused about the sentence on Line 423-425. Is information about phenology in these populations known from the experiments in this manuscript? If so, please include.

We address this point above.

Figure 4 – The dependent variable is confusing on the climate-latitude and climate-traitpc1 importance graphs. Please add a sentence to the figure legend describing what these values are and how they relate to the random forest regressions.

We added the following clarification to the legend: "Variable importance is defined as the average amount a climate variable reduces mean-squared error in the predicted response (TraitPC1 or Latitude), compared to a randomly permuted dataset, across all trees in the random forest (see Geneur et al. [2015] for further detail)."

A supplementary figure depicting mortality in each of the treatments would help readers get a better sense of the intensity of drought stress in the absence of any soil moisture measurements.

We have added this as a panel in the reaction norm figure (Figure S5) mentioned above.

Reviewer: 2

Comments to the Author

In this study the authors used Erythranthe cardinalis to highlight a latitudinal cline in photosynthesis & growth and identify selection pressures underlying this cline. There are really two components to this study: a common garden experiment and a set of correlative analyses. They propagated individuals from 16 wild populations and applied temperature and drought treatments, revealing significant between-population and between-treatment variation in several traits relating to germination, growth and mortality. The traits were summarised using principal components and the first axis used to investigate correlations with latitude and the underlying climatic factors that vary with latitude. They suggest that there is little evidence for variation in plasticity between the selected populations and that interannual climatic variation within climatic neighbourhoods (climate averaged over a 62-km radius around populations) best explains trait and latitudinal variation. They conclude that interannual variation in precipitation is an important selective pressure.

The manuscript was difficult to navigate through, the introduction is a bit stilted with some errors and a bit of fuzzy logic to suggest that the method can really test the hypotheses. The assertion that local adaptation is one of the most ubiquitous observations in nature is not really correct, largely because it is so hard/time consuming to identify and prove. In particular the paragraph from ln 43-62 does not seem to make a lot a sense. A major assumption made at the beginning of this paragraph is that there is an underlying climatic driver behind (physiological) local adaptation. I think this is a much bigger assumption than the authors realise, there could be selection due to the abundance, particular pollinators/pests/competitor species or to the underlying geology. There may also be much finer scale local adaptation than is captured at the scale imposed and the climatic variables may not be appropriate summaries of the climatic variablilty. These things should be more carefully considered and explained as they are

critical to the interpretation of the results. This is really the second reason that latitude may correlate more strongly than multiple climatic variables. The "alternatively...migration" statement (ln50-53) is really just stating that migration may be swamping selection so that it is not possible for local adaptation to occur at that scale. Merging all the variation in the neighbourhood is a total fudge as the physiological measurements were made for specific populations and there doesn't seem to be a valid basis for choosing any particular scale to merge data into a neighbourhood without genetic information (I would prefer to see this neighbourhood angle completely removed, see below). Ln 56 this statement is just plain wrong. I think the distinction between intrinsic and plastic variation that are both dependent on genotype could be better/more clearly explained as it is a possible point of confusion. It should also be pointed out somewhere that at least some of this local variation in physiological traits may be mediated by epigenetic modifications.

The reviewer raises several points here that we have addressed:

1. We have revised our statement about generality/ubiquity of local adaptation in both Abstract and Introduction (lines 1-6).

2. We have revised the Introduction (lines 77-82) and Discussion (lines 358-60 and 442-4) to note additional limitations and further justify the focus of our study. We agree that local adaptation to other factors such as competitors, pollinators, and edaphic conditions are also important in general. However, we did not have compelling observations from which to make testable hypotheses at the outset of the study. The study species lives in sparsely vegetated habitat (gravel and cobble bars along rivers), making competition an unlikely selective agent. The species is hummingbird pollinated throughout its range. There are only two species of hummingbirds in the region and they have similar morphology. Thus, although there could be subtle differences in visitation rates, there is nothing about the natural history to suggest that variation pollination is a likely selective agent, and population genetic data suggest little variation in mating system. Nor did we have compelling observations that these factors could explain the latitudinal clines since they do not vary strongly with latitude, independent of climate (e.g. climate determines both when the plants and pollinators are most active). More importantly, local adaptation can occur along multiple environmental axes and spatial scales simultaneously. The fact that this study does not measure local adaptation to biotic factors or microsite climate does not alter the conclusion that a genetically based cline is consistent with local adaptation to climate and that we have identified possible selective agents. We focused on climate because climate often determines and where species are found and also can exert strong selection on populations within species. There is also a compelling need to know how populations are (or are not) locally adapted to climate so as to predict how they'll respond to climate change (Aitken and Whitlock 2013, now cited). We disagree with the reviewer on one point: contrary to their statement, we did not ignore the possibility of small scale (~ 90 m^2 - the grid size of the digital elevation model) climatic adaptation. Climatic factors like average temperature did not vary strongly with latitude and we could easily have observed that populations

A significant issue with much of the logic and interpretation is that the authors blur the lines between correlation and causation. While finding a causative selective pressure was the goal, this is very difficult to identify, especially without genetic work. The authors assume that the selective pressure is climate and do not examine any other potential drivers of the observed latitudinal cline. The most they can say with this experiment and analysis local adaptation at the chosen scale does seem to take place (the populations are significantly different in the common garden) and that certain climate variables are correlated with traits that vary among populations (regardless of experimental temperature/water treatment) and little about the causes of local adaptation. The experiment and the climate analysis seem quite disconnected and more effort could be

made to show how they fit together. The authors should be a bit more circumspect about what the experiments are set out to show and what they can really tell us.

We have done our best to modify our statements throughout the manuscript with this criticism in mind. We recognize that field experiments are necessary to establish causality between selective pressures, traits, and fitness differences. Based on results from this experiment, we are undertaking such experiments now.

The discussion had several points that feel like conjecture rather than discussing what the results of this study can show. For example, that there is a "fitness tradeoff between low versus high latitude environments", which was not tested in the study. Some statements important to the discussion are lacking references e.g. "We also know from previous work that population size does not vary strongly with latitude."

In several portions of the Discussion, we attempt to explain why there might be local adaptation with some traits, but not others. This is an important result of our paper that merits discussion as we would like to know what sorts of traits mediate local adaptation, which do not, and why (indeed, Reviewer #1 asked for further discussion of these points). However, we have attempted to more carefully distinguish throughout the Discussion exactly which conclusions are based on our results and where we are speculating about possible explanations for future work.

The authors are unclear about their definition of scale in local adaptation. "As this hypothesis predicts, climatic neighborhoods (62-km buffer around populations) correlated with TraitPC1 and latitude of occurrences better than local climate" which seems to indicate that populations are not locally adapted. There are also confusing statements that are not backed up with evidence which seem to undermine discussion points and warrant further explanation e.g. "This may not greatly constrain local adaptation because local and neighbourhood climate values were generally similar in E. cardinalis populations (data not shown)." They then go on to detail how "Riparian habitats where E. cardinalis live are highly heterogeneous at small spatial scales." This was interesting but added an extra layer of complexity that again was not addressed in the study and seems unnecessary for the discussion.

We have addressed the first point above - the closer correlations with neighbourhood climate suggests that while local adaptation may not be perfect because of gene flow in spatially heterogeneous environments, at broad scales (100s of km) populations are locally adapted. We have also clarified that, at the resolution of the climate data (90 m^2), local and neighbourhood climate are highly correlated. However, within several meters of a population, there is a lot of small scale variation in vegetation from riparian areas with water available all year, to chaparral and pine forests that become extremely dry by the end of summer.

In the final paragraph the authors write "These differences between populations suggest different trait optima in different climates." The populations do have different trait means

under controlled conditions but this is different from having different optima, particularly as there are many traits that may be involved in trade-offs that are not measured.

We have changed this statement to a prediction: "We therefore predict to find different optima for these traits in different climates."

Perhaps the most interesting part of the paper is the significant population*water interaction on leaf expansion, the strong positive relationship of TraitPC1 vs. interannual variation of precipitation and latitude vs interannual variation of precipitation. This suggests that populations may having differing capacity to respond to different watering regimes, which could give some indication that the populations are responding this way to interannual variation of precipitation in nature. Although this link cannot be proven directly from the data in this study it is plausible, however, the design of the experiment could introduce this difference as an artefact of measuring leaf expansion under the treatments at different times during the life of the same plants.

We agree that this is potentially interesting and deserving of future study, but are reluctant to draw many conclusions at this stage given that the interaction was not that strong (P = 0.024), especially considering the number of tests performed. As discussed in response to the next comment, we timed growth measurements to capture comparable phases of growth across all plants, so it seems unlikely differences in leaf expansion are an artefact of measuring at different times.

The common garden experiment was confusing as it was unclear how many individuals per population per treatment were used, why these treatments and measurements were performed at certain times during the experiment and how these temporal differences would affect the downstream analysis (does growth rate of these plants differ during a treatment because of life stage changing during treatments?). While the authors aim to assess variation in plasticity, this seems to ignore that any standing genotypic variation present in each population may lead to different responses among individuals in a single population which in turn may lead to variation in plastic response within a population. This experiment is really the most solid part of the study so it should be more clearly described as it does affect interpretation of the result.

We have added details on numbers of individuals per population (line 108), number of families (line 108), and number of individuals per treatment (line 142).

We now explain on lines 154-65 that leaf expansion rate measurements began as soon as leaf expansion began (shortly after the emergence of true leaves) and continued until leaf size began to asymptote and the stem began elongating. We continued measurements of stem elongation throughout the vegetative phase and ceased measurements once plants started producing buds. Therefore, we do not think that changes in life stage effect these measurements because all were taken prior to the transition to reproduction.

Because of low replication within families, we did not have power to assess within population genotype-by-environment interactions. We note this limitation in the Discussion (lines 347-8). Therefore, we cannot say that GxE isn't important, only that it does not appear very strong across populations over the environmental gradients we tested.

The authors used an extensive and complex array of tests (which appear suitable but I am not familiar with all of them – the justification for using these approaches should be presented, particularly the forest) to attempt to find correlated climatic variables. Despite this, the authors do not consider the most important Climate-TraitPC1 variable (diurnal range) because it "did not have any obvious similarity to Climate-Latitude variables" yet when the most important climate-latitude variable did not overlap with any Climate-TraitPC1 variables they chose to pursue this further. This seems to be a double standard and the authors should either change their analytical approach or stop considering variables that do not fit their current approach. It is also unclear how well the dataset used in the climate-latitude analyses can represent the 16 focal populations chosen for experiment/climate-trait analyses. Some additional work should be done to clarify this. Bioclimatic variables are often correlated and I think the authors could try removing highly correlated variables before running their random forest regression to see how this affects the output and downstream results. The logic behind using the climate-latitude variables also needs to be explained more clearly.

The reviewer raises several points here that we have addressed:

1. We have added justification for using VSURF and some details about the algorithm (lines 251-60). In particular, we explain that VSURF automatically tries to eliminate unimportant and redundant climatic variables. We believe that an algorithm that uses the data to eliminate collinear predictor variables is better than arbitrarily eliminating variables before the analysis.

2. The decision to ignore diurnal range (the most important Climate-TraitPC1 variable) was based on criteria we established prior to the analysis and explained in the Methods. At the outset, we were concerned that VSURF might identify spurious correlations between TraitPC1 and climatic factors. This is why we complemented the Climate-TraitPC1 analysis with the Climate-Latitude analysis, as explained on lines (238-48) and reiterated here for convenience:

"To help eliminate potentially spurious correlations between TraitPC1 and climate, we tested for overlap between climatic variables that best predict latitude of all *E. cardinalis* occurrence records (see detail below), not just the 16 focal populations. We refer to these climatic factors as `Climate-Latitude' variables. The logic is that climatic factors associated with both TraitPC1 and latitude for all populations are more likely to be important selective agents than climatic factors that happen to correlate with TraitPC1 but do not covary with latitude throughout the *E. cardinalis* range. Therefore, we did not consider Climate-TraitPC1 variables to be candidate selective agents unless the same or very similar variable was found in the Climate-Latitude analysis."

We therefore do not believe we are applying a double standard, but rather following through with an approach aimed at avoiding overinterpretation. We decided a priori not to consider further Climate-Latitude variables that did not overlap with Climate-Trait variables because the goal was to explain the latitudinal clines in traits, not all aspects of climate that vary with latitude. We hope these clarifications suffice, but we would be glad to clear up any more lingering misunderstanding if provided with more specific criticism.

3. We have clarified on lines 104-6 that "These populations were intentionally chosen to span much of the climatic range of the species based on all known occurrences (see below)." We did not use a formal method because we had several additional constraints (seed availability, overlap with other studies, and we wanted half of populations from Sierras, half from Southern ranges). However, within these constraints we carefully selected populations that spanned much of the range of bioclim variables across all known occurrences. We therefore believe this to be a representative sample.

4. As mentioned above, we have now clarified that VSURF is a multistep algorithm that progressively eliminates redundant variables.

The climatic neighbourhood has potential to summarise a broader set of variation but is only really a useful concept to test if multiple populations (that are tested in the common garden) are included in a single neighbourhood and gene flow estimated between them, otherwise there is little real value/justification for doing this. The choice of 62km seems to be based off an unpublished paper, more explanation is needed to say why exactly this was chosen. It could be interesting to see whether the relative importance of climatic neighbourhood variables vs local variables changes when you increase the radius.

The 62-km buffer was chosen based on the signature of spatial autocorrelation in allele frequencies; this is the distance at which allele frequencies are no longer significantly autocorrelated, while at shorter distances they are. The spatial autocorrelation analyses are based on ~25,000 SNPs identified in a genotyping-by-sequencing assay, conducted on 49 populations from throughout the range and 10-20 individuals per population. From extensive, range-wide field surveys of presence and absence (Angert et al., ms in revision) that validate ecological niche models built for this species (which were used to weight the neighbourhood climate estimates in this ms), we know that the 62km buffer is highly likely to contain multiple populations.

The results contain a *lot* of discussion. Separating the description of what is actually observed from its interpretation will help the clarity.

We have moved some sentences from the Results to Discussion. We would be glad to move additional material if some sections still seem like too much interpretation.

In summary, there are two interesting approaches taken but the manuscript is not clear and the interpretation overcooked. I would advise that a clearer separation of the explanation and interpretation of each is required and that the authors think more carefully about how they present the underlying rationale, limitations and results.

Some additional comments

Ln 97 It is unclear how many individuals per population per treatment are used, which would be a useful a statistic to include?

We have added this information on lines 143: "Each population had 8–12 individuals per treatment level (mean = 11, median = 11)."

Ln 126 More effort could be put into detailing why the unavoidably high humidity would not affect these results, as this does not represent a drought.

We have clarified on lines 140-1 that lower humidity would have made the drought more severe, but low soil moisture is stressful in and of itself.

Ln 187 Why are the hypotheses (1) 1 or 2 climatic variables, (2) Multiple climate variables – what is special about 2, should it not be just (1)1 and (2) Multiple?

We wanted to delineate between 'simple' (1-2) and 'complex' (3+). Obviously, this is a subjective cutoff, but it seems reasonable to us.