Recommender’s comments:

1. I liked very much the analysis of the climate clusters, I reckon it is a very good way to address the issue of unbalanced sampling design. Only concern here is that I don’t understand what are the data points on Fig 2. There are 30 points for each meteorological station, so it seems there is one point for each year, but what is it? TN, TX, their average? Please clarify in the text and/or figure legend.
2. Once the clusters were identified, I don’t understand why clusters 2 and 4 are kept in the analysis although they don’t include any sampling point from the “early” period. There seems to be a problem in testing a cluster x period interaction with 2 clusters containing data from a single period only. Alternatively, if you are interested mostly in inter-cluster differences, then you should not analyze longitudinal data from the 2 periods. Choose your question here, and select the data that help you answering it. I suspect there is enough data within cluster 1 and 3 to address the issue of temporal variation in reproductive success across clusters. At present, you are trying to analyze temporal and cluster effects independently, but you end up describing partial interactions which makes your argumentation very case-specific (e.g. L. 295-303, L. 310-320 you mention a significant effect of period, but you describe only effects within clusters and never the shared trend across clusters).
3. I do have a remaining concern about including single samples from a given year within each Cluster (i.e. 1992, 2017 and 2019 in Cluster 1 ; 1993 in Cluster 3). By including these points, you make the assumption that annual variation within each period is negligible, but then you do mention extreme events in 2017 in the discussion to help explaining your results. I would like to see how the inclusion or exclusion of these points affect the results of the study. On the whole, there is quite a lot of stress on these particular events in the manuscript (e.g in the abstract and the discussion), but they might not be reflecting any wider trend, and they are not directly supported by the climatic data you provide.
4. I am far from convinced that the non-parametric approach chosen here is the most appropriate for your data. I did not understand if you analyzed each egg mass as a replicate point, or if data were aggregated by sampling site. There seems to be an issue with pseudo-replication (uneven number of egg masses and/or year replicates for each sampling site), so a mixed model including sampling site as a random effect would be more appropriate here. Cluster and period should be included as fixed effects. The fact that the variables are not Gaussian should not be a problem, as Generalized Linear Mixed Models can be used (e.g. Poisson or Negative Binomial distribution for the number of eggs ; Gamma distribution for egg length; Binomial distribution for hatching or parasitism rate).
5. I have strong doubts about the methods for the analysis of change in temperatures : (i) first, the 30 years of monthly averaged TX between 1990 and 2019 are analyzed with linear regression, instead of specific methods that exist for the analyses of time series (e.g., spectral analyses, auto-regressive models, etc) that are more powerful to detect underlying trends ; (ii) for the analysis of TX averaged over summer, the time period is split into two groups (one for 1990-2004, and one for 2005-2019 if I understood correctly) and period is modelled as a factor in an ANOVA. I really don’t see the logic in this analysis, the summer-averaged TX could be analysis as a continuous time series, just like the monthly-averaged TX. It might make sense to split time into two discrete groups if it were to analyze the differences between the two periods of sampling (1992-1995 vs 2010-2014) that are indeed separated by a few years. The analysis chosen here has very low power and I am not surprised it fails to detect any effect of time.

In addition, the abstract mentions an increase in climatic variability, but I did not find this analysis in the text (I might have missed it)? Time series analyses can also be used to reveal temporal changes in the amount of stochastic noise in the data, which could support your argument here.

1. Some of the figures are poorly informative. The axis on Fig 4 is redundant with the legend box that already assigns colours to clusters, and it is misleading as it does not reflect the continuous nature of the temporal data. Fig 6 should be redone: bar plots are not helpful here to visualize small quantitative differences and the temporal dimension is not apparent on the graphs. I would rather suggest scatter plots of average values for each sampling point (with confidence intervals or quantiles) plotted against year on the x-axis, and if possible the different clusters plotted on the same graph if it not too crowded.
2. Tables are difficult to read and provide information that is not entirely necessary to understand the results. I would suggest to modify Figure 1 to include information about sampling periods (with a color code for instance) and to remove Table 1 to the Supplementary Information. Tables 2 and 4 are quite redundant with the text, the values of adjusted mean are difficult to interpret, I would also suggest to remove both Tables to the Supplementary Information and to focus on improving Fig 4 and 6 to make any trends more visible.
3. I had the feeling that the discussion was a bit lengthy and it was sometimes hard to keep up with the main message. I reckon it could benefit from a bit of shortening, with a tighter focus on the main questions asked in the study (e.g. inter-cluster comparisons might not provide major insight here, the part on local adaptation seems a bit far from your results, detailed description of isolated data points make it difficult to see the higher trend).