We want to thank the editor for his guidance in addressing each specific comment. In the following responses, we reproduce the reviewers' comments in italics and include our detailed responses in bold text thereafter.

Overall Comments: *As much as I like the topic of this paper, I don't find the revision to be much more informative than the original. The typos are now gone, but new issues arise and the authors still delegate many important matters to citations. I am a statistician, trying to learn how to conduct these tests, and this paper is incredibly vague. I could not reproduce any of the results from what is given in the paper. I do not want to see another version of this. Maybe some of my troubles lie with the Copernicus' journal's editorial process, but it is frustrating to see virtually the same paper again.*

We address the specific comments of the reviewer in the sections below. We nevertheless wish to push back on the contention that our last revision was "was virtually the same paper," relative to what we originally submitted. To the contrary, we submitted a comprehensive response to the reviewers and a heavily edited manuscript. We therefore respectfully disagree with the reviewer's assessment, which is easily refuted by a quick scan of the track-changes version of the previous submission.

Specific Comments:

1. *I am still unclear how the hypotheses are tested. I believe that the authors are getting p-values from some sort of self normalization procedure, but this is never made clear. I previously thought that they were coming from the chi-squared distribution, but now I doubt this (lines 259 and circa 275). But if true, doesn't self normalization need to be discussed? Since I am unfamiliar with this technique, what hope do climate scientists have? This is my frustration with the paper: I do not know why we are doing what is being done.*

   Because the self normalization test is a well developed technique, we did not include many details in the paper to avoid redundancy. For the convenience of readers, however, we now provide the details of the self-normalization procedure in the new Appendix of the revised manuscript.

2. *At the centre of the tests, I still don't understand why we are projecting onto eigenfunctions. If I want to test whether the mean is the same from two samples at a fixed site, I look at the difference between the univariate averages – univariate asymptotic normality comes up. If we want to examine all sites in $\mathcal{D}$ simultaneously, a vector of mean differences arise and multivariate asymptotic normality arises. The authors provide some words on this, but hardly anything swaying.*

As we noted in our previous response, if we just want to test whether the mean is the same from two samples at a fixed site, the two sample t test (or the asymptotical normality test) would be sufficient. To examine the difference at all sites in $\mathcal{D}$ simultaneously, however, the chi-square test based on the multivariate normality of observations at all sites has several practical and technical issues. The chi-square test treats the mean at each site equally such that noise can dominate the testing results, whereas our current test aims to examine the difference at major modes of the climate fields. The projection onto EOFs in our method helps filter out the noise and thus allows us to focus on the subspace containing the leading dynamical climate pattern. Additionally, our method can test both the mean and covariance, while chi-square only tests the mean. One technical challenge for the multivariate normality test for high dimensional data is to estimate the variance and covariance function of the multivariate process over the globe. Climate data measure a highly heterogeneous process and any parametric models will be limited in their ability to capture the dependency structure of a global process. For example, our previous paper (Li and Smerdon, 2012) demonstrated that the chi-square test is sensitive to the misspecification of the covariance function. Our current method is purely non-parametric and thus free of the risk of model misspecification. Another technical issue for the multivariate normality test is that the method requires the two multivariate data to be independent, a condition that obviously does not hold for our data. In contrast, the method that we have applied allows for dependence between the two data sets. All of these points relating to the comparison between our method and the more traditional multivariate normality (chi-square test) method are now even more plainly addressed in

Lines 82-87 of the revised manuscript.

3. *There still isn't anything that I see in the paper that tests for both equality of means and autocovariance simultaneously.*

   We agreed that tests for both equality of means and autocovariance simultaneously are a natural way to evaluate the discrepancies between two climate fields, and that is exactly what Li and Smerdon (2012) and Li et al. (2016) did in their papers using either a parametric method or a functional data method. Because our current method aligns with Li et al. (2016), we can easily implement their joint test on our data, but obtaining a single p-value is not the focus of this paper. Instead, we are interested in dissecting each segment of the testing results to understand the mechanism behind the differences across the CFRs. A joint test for both the equality of means and autocovariance therefore does not add value to these interests. To explicitly explain why we did not perform the joint test, we have added the below text to the revised manuscript at Lines 263-266:

   "Another available test for evaluating the difference between two climate fields is to combine hypotheses (i) and (ii) into one single test, as in Li and Smerdon (2012) and Li et al. (2016). We omit this joint test because the focus of this paper is to understand why the mean and covariance in a reconstructed field behave differently. Thus, each individual test is sufficient and more pertinent for such a purpose."

   This explanation is also reiterated at Lines 596-601 in the Discussion and Conclusion Section.

4. *Section 2.2 seems new, but its notation is bad! First, you are denoting variants of quantities with a prime, and mixing this in equations where T denotes transpose. Compounding this, you have a variable named T! Matrices and vectors are not bolded. There are quantities like P^r related to P' (why suppress the prime?). It took me an hour to deconvolve this simple section!*

   We have revised this section with the following changes:

- **Standardized matrices of P and T are now denoted as P_{std} and T_{std}, respectively.**

- **Reduced rank representations of P_{std} and T_{std} are now denoted as P^r_{std} and T^r_{std}, respectively.**

We use a superscript T to note the transpose of a matrix, which is a common notation in the statistical literature and is clearly distinguishable from where the temperature matrix T is used.

In our experience, the guidelines for matrix notation vary with different journals. Some journals require bold text for matrix variables, while some specifically require that they are not bold. Regardless of the journal, however, all notation should be consistent throughout a given paper. The notation of matrices is consistently not bold throughout our manuscript and for now we maintain this convention. If 'Climate of the Past' suggests that we use bold matrix notation throughout our manuscript, we will be happy to do so.

5. *Grammatically, the paper is pretty good. Nonetheless, there are a few spots where articles are abused or there is awkwardness. For one such example, the first line in the abstract should probably start with "This paper derives". And spatiotemporal really still needs a dash to be Oxford compliant.*

   We leave this to the proofreading and copy editing, as suggested by the editor.

6. *I apologize for being so picky, but your paper seems like a black-box for climate scientists to follow rather than something informative.*

   This characterization is off base. Our paper is comprehensively cited, allowing any reader to discover the specifics of the method that was applied. The purpose of our paper, however, is to use existing methods to elucidate why and how CFR methods perform the way they do. It is therefore entirely appropriate to provide citations as the principal background on the employed methods, while complementing the discussion with methodological

summaries as we have done in our paper. While we believe this approach is sufficient and concise, we now provide even more details on the statistical methodology in a new Appendix within the revised manuscript. We also will publish access to the codes used to perform our analyses on the GitHub site now listed in the revised manuscript.

References

Li, B. and J. E. Smerdon (2012). Defining spatial comparison metrics for evaluation of paleoclimatic field reconstructions of the common era. *Environmetrics 23(5)*, 394–406.

Li, B., X. Zhang, and J. E. Smerdon (2016). Comparison between spatio-temporal random processes and application to climate model data. *Environmetrics, 27(5) 27*, 267–279.