In the following responses, we reproduce the reviewers' comments in italics and include our detailed responses in bold text thereafter.

Response to Reviewer 2

1. *The biggest issue I have with the paper is trying to figure what was done in the methods. Obviously, we need this kind of comparison in climatology and it is extremely important. I didn't have any issues with the scientific prose in the paper, but I would like the authors to be much more pedagogical in their methodological exposition so that folks can discern what has been done. I did not get anything out of the brief description of the four methods in Section 2. Hence, I was hoping that Section 3 would alleviate these concerns; alas, it did not.*

   The details of the four different CFR methods employed in our manuscript are widely reported across many different publications, which we summarize in Section 2. We nevertheless will be happy to further explain the methods that underlie the employed reconstructions to better situate the reader

2. *In Section 3, some things are not stated that need to be: are the X and Y processes independent (I think so)? It seems you are assuming a constant mean in time t (which is not likely true) and that the covariance function of the spatial fields at each time have the same structure (I can buy this). I also want to know if you are assuming that the fields are Gaussian.*

   We will further elucidate the following points in the manuscript.

   - **Our method does not require X and Y to be independent. This is one of the advantages of our functional data analysis methods.**
   - **We do not assume a constant mean in t. We removed a common trend that is a constant at t, and then we allow the mean of the detrended data to be spatially varying. We indeed assume the spatial covariance function follows the same structure at each time.**
   - **Our method does not require the random fields to be Gaussian, though the climate model data can be approximated by a Gaussian random field. In the case of Gaussian random fields, our test for mean and**

covariance is equivalent to the test for distribution because the mean and covariance determine the Gaussian distribution, but this equivalence cannot be generalized to other distributions.

3. *In testing for whether the means of the two processes are the same, why would we not just look at the average (overall spatial locations and times) and use asymptotic normality to test whether these X minus Y averages have a zero mean? This just works with differences $\Delta(t,s) = X(t,s) - Y(t,s)$. Then you don't have to assume the mean is constant....it subtracts to zero under the null. You can easily estimate the variances of the average Δ value assuming a null that the two fields have the same covariance structure. This seems to be the fundamental way to handle the two sample equality issue in general abstract spaces. I'm guessing that what you've done can be justified, but it would seem that I have to go to your past papers to dig this up. I just have this uneasy feeling that the EOF approach is needlessly complicated.*

   We would like to clarify that we did not assume constant mean for the whole spatial domain. Instead, we allow the mean to be spatially varying and we aim to test whether $\mu_X(s) = \mu_Y(s)$. We removed the common trend from X and Y, which is calculated as the global average at each year based on both data sets. This detrending procedure will not affect our spatial mean test.

   Regarding the reviewer's suggestion, we agree it is the most natural approach. We indeed carried out a very similar idea in Li and Smerdon (2012) by centering, scaling and decorrelating X and Y using their common mean and common covariance matrix, and then evaluating whether the two post-processed data sets followed the same distribution. However, working on that project made us realize several drawbacks of this seemingly rigorous method, and motivated us to seek a more robust and flexible method resulting in our follow-up work Li et al. (2016). The drawbacks can be summarized as: 1) temporal dependence is not taken into account; 2) X and Y are assumed to be independent; and 3) the result is sensitive to misspecification of the spatial covariance function. We later turned to the functional data method in Li et al. (2016) which was used in our current manuscript for the following reasons: 1) it allows dependence between X

and Y and temporal correlation within each data; 2) it is nonparametric, thus there is no concern for model misspecification; and 3) it enables an assessment of the discrepancies between X and Y at different directions or subspaces.

4. *I also can't rationalize why I need to use the data from times 1 to k in various places. Seems I should use all N times once.*

   We apologize for the confusion. We used the data from times 1 to k because we employed a self-normalization approach to generate samples for the variance estimation. Self normalization is analogous to bootstrap or subsampling, but bootstrap tends to break the temporal correlation and subsampling needs to choose the optimal subsample size and the subsample form. Self normalization for temporally correlated data was developed by Zhang and Shao (2015) for generating recursive samples for the parameter estimates in which the variance is the variable of interest. The recursive samples are obtained by drawing samples from time 1 to k, $k = 2, ...N$, meaning each time the new sample is formed by expanding the previous sample by adding the current observation. Self normalization is also tuning-parameter free and allows for temporal correlation. More details can be found in Zhang and Shao (2015).

5. *It would seem to me that we want to test whether the means are the same and the covariances are the same in tandem. Not either the mean is the same or the covariance is the same separately, but to test both in tandem. So why not set $\Delta(t,s) = X(t,s) - Y(t,s)$ and work with these differences as above. If the means are the same, the mean of the Δ process is identically zero at all times and spatial locations. Then we could stack the $\Delta(t,s)$ in a giant vector — call it V — over all spatial locations and time points. Now, if we could get the covariance matrix of all components in this giant vector — call it Σ — we would just look at $\Sigma^{-1/2}V$. This quantity would be composed of IID N(0,1) variates if the original fields are Gaussian. And it is easy to test whether data is $IID \;\; N(0,1)$ by a plethora of methods (QQ plots, Kolmogorov-Smirnov, chi-squared tests, etc). To estimate Σ, grab your favorite space-time covariance estimators to estimate both the X covariance and the Y covariance structures in time and space. Call these estimates*

Σ_X and Σ_Y, respectively. Let $\Sigma^* = (\Sigma_X + \Sigma_Y)/2$ be the common estimate under the null that the two processes have the same covariances and are independent. Now just use $Cov(\Delta(t,s), \Delta(t',s')) = 2$ times the corresponding entry in the matrix Σ^*. Then I think it's game over: you've tested both hypotheses at the same time.

We once again fully agree with the reviewer that their suggestion is most natural. As mentioned in our response for bullet point 3, we have tried a similar idea in Li and Smerdon (2012). The method works to some degree, but it is unsatisfactory in several ways as we listed in point 3. The temporal correlation in climate data, the possible dependence between the synthetic (X) and CFR (Y) data, and the complex correlation structure of the global climate data that can challenge the validity of any stationary and parametric covariance function, gave us impetus to seek more flexible and robust methods. Hence, we have employed the functional data method in our manuscript. Another benefit for principal component based methods is that noise will be filtered out in the analysis. This is very important for our hypothesis testing, because we are analyzing very high dimensional data for which noise can dominate the result and lead to misleading conclusions.

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), 267–279.

Zhang, X. and X. Shao (2015). Two sample inference for the second-order property of temporally dependent functional data. *Bernoulli 21*(2), 909–929.