Dear reviewer:

Thank you for your thoughtful and constructive feedback. Here we provide a complete documentation of the changes made in the manuscript in response to each of your comments. Reviewers' comments are shown in plain text, while author responses are shown in bold green text.

Summary: Past global surface temperatures over the past few centuries can be estimated from present borehole temperature profiles applying inversion techniques based on the solution of the heat transfer equation. Form a set of sites where temperature profiles have been measured, large-scale temperature reconstructions can be derived by averaging the local retrievals. The manuscript presents a method to estimate the uncertainties in the large-scale average estimations based on a bootstrap approach. the authors conclude that this new method provides better and more realistic uncertainty estimates that previous methods. Those previous methods simply calculated the average of the high and low ends of the local uncertainty ranges.

Recommendation: I think that in general the manuscript is valuable and should be published after some revisions. However, I am afraid that one of the motivations of the present study, namely that the previous estimations of uncertainness was unrealistic, contains a conceptual misconception, although it has been previously published. Therefore, the motivation of the present manuscript should be amended to present a correct statistical case. I explain below in more detail my main concern.

1) The manuscript presents a base method to estimate global or large-scale uncertainties that has been published previously. This method just constructs the high-end (and low-end) uncertain range of the global average by calculating the average of the high-end (or low-end) range of the local estimations. This is, however, not correct, as it can be illustrated in a short counter-example. The interpretation of a 5-95% uncertain range in a frequentist approach is that the range covers the true value with 90% probability (technically, it means that a putative infinite number of realizations of the measurements and their corresponding uncertain estimations will contain the true value 90% of the time). For the sake of this reasoning, we can a bit sloppily say that that the probability that the true value is within the estimated uncertainty range is 90%. However, if the uncertainty ranges are constructed by averaging the 5% and the 95% local ranges, this probability is much much larger than 90%. Let us focus on the high end (95%). The probability for the average to be outside that 95% range is not 0.05, but actually 0.05 to the Nth power , where N is the number of profiles (sites).This results because each profile from which that average is constructed, has a probability of 0.05. If N=100, this number is very small, much smaller than 0.05.

The authors realize in the discussion that indeed this estimation is not correct. There, they apply a much

more correct estimation assuming that the global profile is the average of N random variables, and therefore, assuming that these N random variables are independent, the error in the average amounts to the sqrt of the average squared error. If all individual errors are equal, this amounts to that individual error divided by the sqrt(N).

There is one important underlying assumption: the errors should be independent across space. But even if this assumption is not completely fulfilled, this estimation is much more realistic that simply the average of the upper and lower local percentiles, which is clearly incorrect.

Thus, to some extent, the manuscript corrects a previous statistical misconception. In this sense, it is useful, but the motivation of the manuscript should be cast differently,as the reader will be really surprised to see, without any caveat, a clearly wrong method as a benchmark.

As suggested by the reviewer, we have indicated on the text that the main caveat of the SVD and PPI techniques used in our manuscript is the lack of a correct statistical method to provide with confidence intervals for the average of inversions from several subsurface temperature profiles. We have also changed the aim of the manuscript indicating that we provide a new, revised and improved method to aggregate inversions from different subsurface temperature profiles.

On the other hand, the question of the spatial correlation of uncertainties, which is critical for the validity of both methods (bootstrap and error propagation) is not mentioned at all.

The bootstrap approach is definitively better - and I could not see any clear error in this application of bootstrapping. However, this approach does not take into account the possible spatial correlation of the local errors. I do not know how significant these correlations might be, but if they are, then the bootstrap estimation of the uncertainty will be too narrow –in the same as the error propagation proposed by the authors in the discussion– since the effective number of degrees of freedom will not be N, but smaller.

If these correlations are relevant, the bootstrap should take it into account, e.g. by block-bootstrap, in which correlated regions are first averaged together, and then bootstrapped. I think that this problem is technically very difficult to solve satisfactorily, but again, I believe that the presented bootstrap approach is indeed useful.

Indeed, spatial correlation may be important to determine the confidence interval of the global estimates of temperature and heat flux change from subsurface temperature profiles. To account for the effect of spatial correlation on global averages, we have estimated the effective degrees of freedom of surface air temperatures at each grid cell of the CRU TS 4.05 product (Harris et al., 2020).

The degrees of freedom (dof) of two temporal series depends on the correlation coefficient (c) between them (Fraedrich et al., 1995) as

$$\mathrm{dof} = \frac{2}{1 + c^2}. \tag{1}$$

In order to estimate the spatial variation of the degrees of freedom of CRU temperatures, we apply Equation (1) to the temperature series in a given cell and the four closest neighbours, obtaining the effective degrees of freedom as the average of the four different estimates. Figure 1 in this document shows the spatial degrees of freedom for annual temperature series and 30-yr running means generated by repeating this process for all grid cells in the CRU product. Results considering the eight and twelve closest neighbours are also displayed. Orography seems to be the leading factor in local variability, with the small number of observations included in the product for several areas, like the Arctic and Africa, also displaying an effect.

We include the different effective degrees of freedom in the bootstrap estimates by estimating the weighted mean of the inversions in the Sampling ensemble to retrieve the corresponding member of the Bootstrapping ensemble. That is, the inversions within the Sampling ensemble are weighted by the corresponding degrees of freedom at the location of the profile, thus inversions from temperature profiles within zones with high degrees of freedom weight more than inversions from profiles in other zones. Concretely, we consider the degrees of freedom obtained using the twelve closest neighbours and 30-yr running means, as this is the case showing higher zonal differences in Figure 1. However, the retrieved global averages and 95% confidence intervals from bootstrap inversions including the different effective degrees of freedom and without considering them present very similar results (Figure 2). Additionally, similar results are obtained when considering annual temperatures, and four and eight neighbours (not shown). Therefore, the effect of the different degrees of freedom at borehole locations is not particularly large for global temperatures retrieved from temperature profiles. We have added a Supplementary information document to the manuscript including these points, as well as Figures 1 and 2 in this document.

Particular points:

2) A definition of the quasi-equilibrium temperature will help some readers.

We have added few lines describing the quasi-equilibrium temperature profile and its main characteristics on the new version of the manuscript.

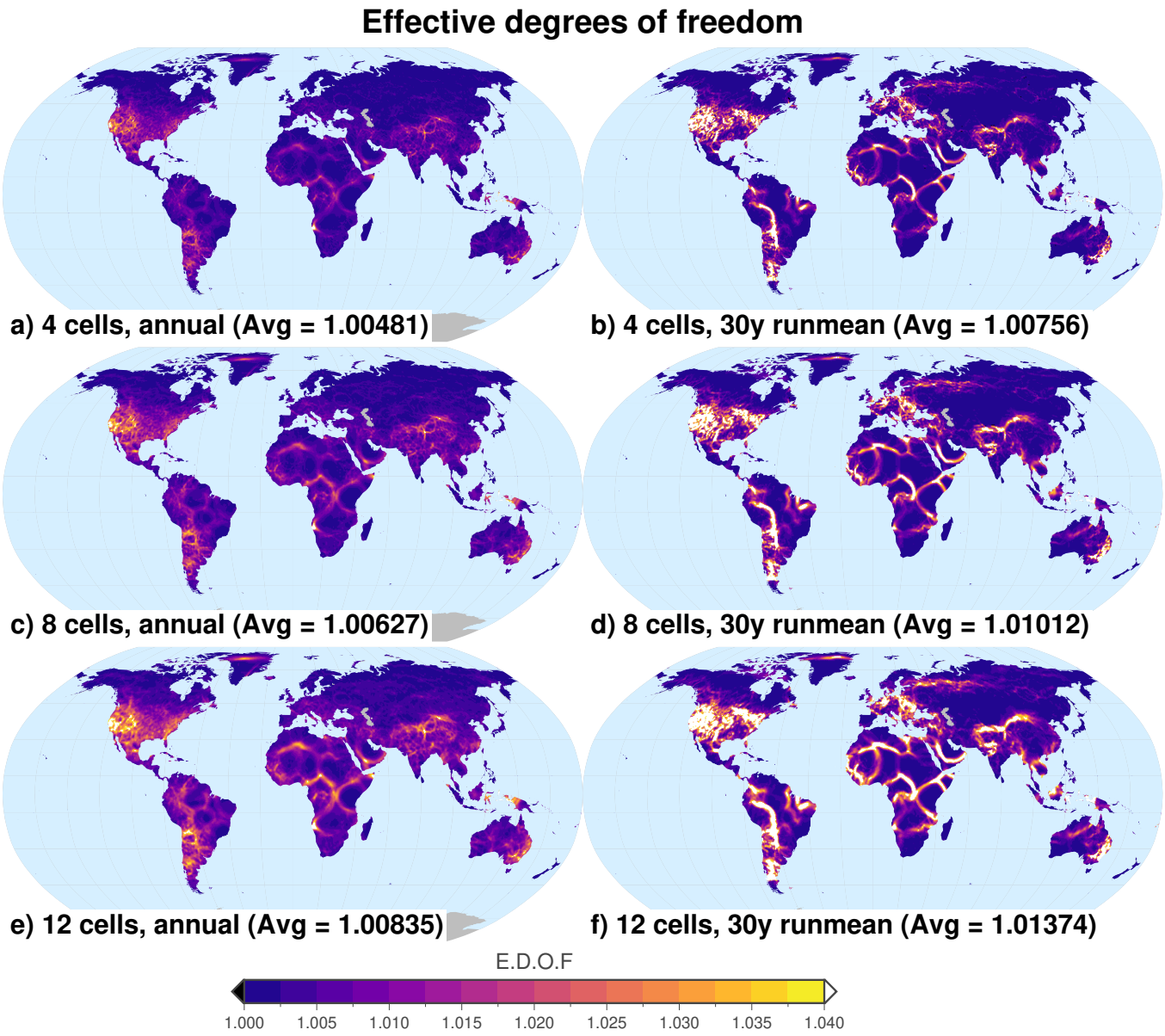

Figure 1: Effective degrees of freedom for CRU TS 4.05 temperatures from annual (left column) and long-term (30-yr running means, right column) series. Results considering the four (first row), eight (second row), and twelve (third row) closest grid cells are also displayed.

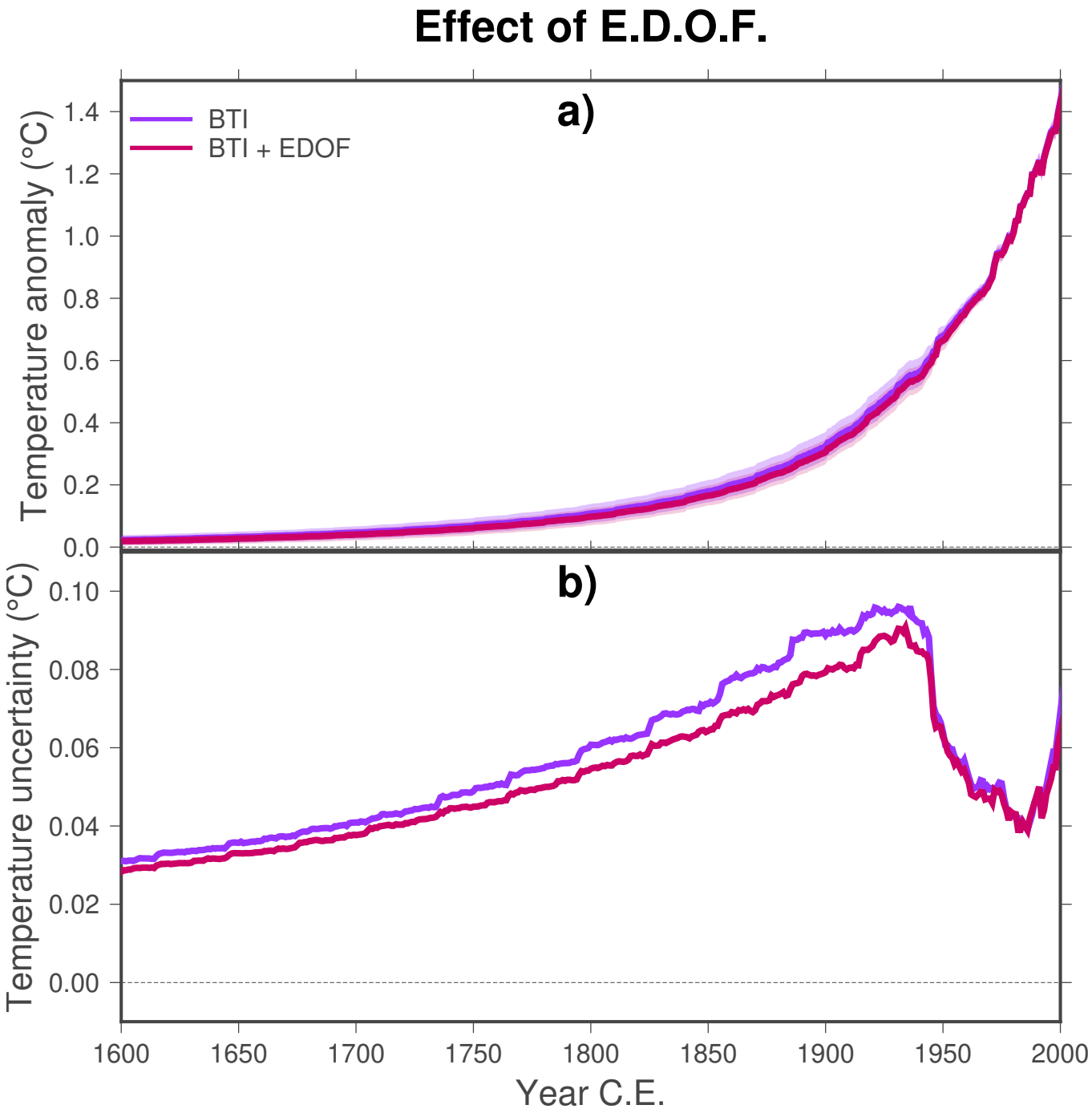

Figure 2: Estimated temperature evolution from subsurface temperature profiles. (a) Global averaged surface temperature histories considering the different effective degrees of freedom at the location of each profile (red line), and weighting all profiles equally (purple line). (b) Range of the 95% confidence interval for bootstrap inversions considering the effective degrees of freedom at the location of each profile (red line), and weighting all profiles equally (purple line).

3) Section 3.5, perhaps the most important section, is not very clearly written (I needed to read it several times). For instance , line 257: “named Sampling and Bootstrapping ensembles (S and B ensembles in Figure 2). The Sampling ensemble consists ...”. and the reader expect the following sentence to explain what the Bootstrapping ensemble is. However, the text goes on with “The BTI method considers the uncertainty arising from ...”. This is quite confusing. Actually, the bootstrapping ensemble is a typical bootstrap sampling from the set of individual local profiles, where each profile has been derived from one value of the uncertain parameters (T0, Gamma0, and thermal conductivity). The only restriction is that each sites contributes with one member to the ensemble.

All in all, I found the technical description unnecessarily too cumbersome.

We have modified Section 3.5 in order to improve the clarity of the text. Please, see this section on the new version of the paper.

References

Fraedrich, K., Ziehmann, C., and Sielmann, F. (1995). Estimates of Spatial Degrees of Freedom. *Journal of Climate*, **8**(2), 361 –369. DOI: 10.1175/1520-0442(1995)008<0361:EOSDOF>2.0.CO;2.

Harris, I., Osborn, T. J., Jones, P., and Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, **7**(1), 109. DOI: 10.1038/s41597-020-0453-3.