Response to the reviewers' comments. cp-2019-120 Melo-Aguilar et al.

Camilo Melo-Aguilar [1,2], J. Fidel González-Rouco [1,2], Elena García-Bustamante [3], Norman Steinert [1,2], Jorge Navarro [3], Johann H. Jungclaus [4], and Pedro J. Roldan-Gómez [1]

[1]Universidad Complutense de Madrid, 28040 Madrid, Spain
[2]Instituto de Geociencias, Consejo Superior de Investigaciones Cientificas-Universidad Complutense de Madrid, 28040 Madrid, Spain
[3]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), 28040 Madrid, Spain
[4]Max Planck Institut für Meteorologie, Hamburg, Germany

Correspondence: Camilo Melo-Aguilar (camelo@ucm.es)

The authors would like to thank the reviewers for their constructive suggestions and the time they devoted in reading and proof-reading the manuscript. We have tried to integrate all suggestions and think that the manuscript has improved with them. We do appreciate their contribution. The original manuscript with track changes is included at the end of this document so the specific modifications can be found in the text.

The next sections contain a detailed point by point response to the reviewers comments. Comments are labeled by reviewers and order of appearance, i.e. R2C3 is the third comment of reviewer 2. The original number by the reviewer is also preserved.

1 Anonymous Referee 1

GENERAL COMMENTS:

R1C0: *REVIEWER'S COMMENT:*

This manuscript assesses the spatial and temporal limitations of ground surface temperature histories reconstructed from borehole temperature profiles, as well as the effect of different forcings on the interpretation of those temperature histories. Authors used a large ensemble of millennial simulations, which includes experiments to evaluate the role of individual forcings in the simulated climate evolution, for this assessment. I find this work relevant for the study of the Earth's system dynamics, paleoclimate, and climate modeling. The evaluation of limitations on the borehole database is of particular relevance. Nevertheless, I think some issues regarding the inversions and trend analysis should be addressed before the manuscript is ready for publication.

AUTHORS' RESPONSE:

The authors welcome the positive perspective of the reviewer on the paper. We are grateful for the reviewer's

comments.

Please find below the comprehensive point-to-point response to your review. The original manuscript with track changes (MSTC hereafter) is included so the specific modifications can be found in the text.

R1C1: *REVIEWER'S COMMENT:*

Inversions: 1- Authors limit the depth of the synthetic boreholes that they create mimicking the depth distribution of the measured borehole database. This limitation seems to be only applied to the bottom depth in the synthetic profiles, as the upper depth is not mentioned on the text. I assume, therefore, that the upper depth in the synthetic profiles is the surface. However, measured borehole temperature profiles rarely include data at the surface, and I wonder if authors have studied the case in which the upper depth of the synthetic profiles is configured as the upper measured depth in the corresponding borehole temperature profile. Would this have an effect on the results?

AUTHORS' RESPONSE:

We appreciate that the reviewer noticed this issue that was not explained in the original text. The upper depth of the borehole temperature profiles (BTPs) was uniformly set to 20 m. We have included an explanation for this issue in the current version of the manuscript. Please see lines 202-203 of the MSTC.

R1C2: *REVIEWER'S COMMENT:*

2- The authors aggregate inversions performed using profiles with different bottom depths. Previous studies (Beltrami et al. , 2011, 2015) analyzing measured temperature profiles have shown that this practice biases the retrieved surface temperature histories, since the period of reference for each subsurface anomaly profile is different. I realize that authors are not using measured temperature profiles and that the synthetic profiles may not share this bias with the real case, but I wonder if authors have assessed the case in which all synthetic boreholes are truncated to a common bottom depth. Additionally, authors should include a brief note in the conclusions stating that although the depth masking do not affect their results, this is not the case when analyzing real borehole profiles

AUTHORS' ANSWER:

Figure 1 of this document illustrates the case in which all synthetic borehole temperature-anomaly profiles are truncated to a common bottom depth, following the reviewer suggestion, in the ALL-F_2 member of the ALL-F ensemble as an example. The global average profiles in the ideal borehole scenario using the standard configuration (600 m depth), as well as an alternative configuration in which all synthetic profiles are truncated at a "shallow" depth of 300 m are presented. In both cases, the upper 300 m show almost identical results. This is because in the surrogate reality of the model world, synthetic borehole temperature-anomaly profiles are directly created. Thus, the upper part of both profiles contains the same surface temperature signal of the las few hundred years. The shallow profile misses only the information of the earlier times that propagates deeper into the subsurface. The inversion of both profiles yields very similar results. Note that in both cases, the target temperature signal is

accurately retrieved.

In the real-world cases, the anomaly profiles are obtained by subtracting the quasi-steady state parameters (i.e the geothermal gradient and equilibrium surface temperature, Beltrami et al. , 2011). The latter is usually estimated from the bottom part of the borehole temperature profiles (BTPs) by linear fitting to the data and extrapolation to the surface. Therefore, truncating the BTPs at different depths may yield different values of the quasi-steady state parameters, and thus, of the temperature-anomaly profile impacting the results of inverted temperature histories. This source of uncertainty is present in experimental cases and has not yet been reproduced in pseudo-proxy experiments (PPEs). Therefore, in PPEs, the only source of uncertainty introduced by having borehole temperature-anomaly profiles of different depths is that related to the fact that shallower boreholes miss part of the past climate variability that deeper boreholes do not. This is as far as the approach in the text can reach and the differences are minor. Most likely, the loss of information of the earlier times in the shallower profile has a limited influence on the results because the CESM-LME simulations show a relatively low multi-centennial variability in the MCA-LIA transition.

We have included a mention stressing the fact that even though the synthetic borehole temperature-anomaly profiles are not strongly biased by the difference in the depth of the profiles, this may not be the case in real-world cases. See lines 340-343 of the MSTC.

R1C3: *REVIEWER'S COMMENT:*

3- Related to the previous comment, I have noticed that each simulated GST anomaly used to generate synthetic profiles have a different period of reference (lines 240-243). Therefore, the inversions of those synthetic profiles are relative to different climatologies. An easy solution to that would be to define a common period to compute all GST anomalies before generating the synthetic boreholes, and maintain such reference in the comparison with temperatures from the model.

AUTHORS' ANSWER:

We have implemented the approach of using a common period to compute all GST anomalies following the reviewer's suggestion. This is done for the ALL-F_2 member of the ALL-*F* ensemble as an example. All GST anomalies have been computed with respect to the 850-1960 CE mean. This period is selected since all borehole sites considered in this study are dated after 1960 CE. Once the anomalies are calculated with respect to a common period, each grid point anomaly (with the presence of borehole temperature profile) is trimmed at the actual logging date according to the real date distribution. Subsequently, the synthetic borehole temperature anomalies profile is created, and finally, the inversion using singular value decomposition is calculated. The period of reference is also used in the comparison with simulated GST as well as in the ideal scenario.

Figure 2 herein show the results as it was presented in the manuscript (Fig 2a) as well the results from this alternative approach (Fig 2b). Note that the differences between these two approaches are hardly noticeable. In both

cases, the B-mask inversion underestimate the ideal scenario (IBS). Other ensemble members yield comparable results. Therefore, the use of a different period of reference to estimate the GST anomalies, and subsequently, the borehole temperature-anomaly profiles cannot account for the differences between the IBS and the B-mask cases. We think that both are approaches correct. Using one or the other represents an appropriate adaptation of the real-world case. Since there are not important differences of using one or the other, we will keep the original approach in the document.

R1C4: *REVIEWER'S COMMENT:*

4- It is very surprising that global mean temperatures from inversions computed using the B-mask configuration (green lines in Figs. 2, 3, 6, 7 and 8) are perfectly aligned to start in the year 2000 of the common era. Nonetheless, authors clearly state at several points on the text that the B-mask configuration considers the logging dates of the real borehole database. Does this mean that such different dates were not considered when aggregating the inversions? The green lines should display some kind of shift relative the ideal borehole scenario (IBS, red lines in Figs. 2, 3, 6, 7 and 8) configurations due to the different logging dates.

AUTHORS' ANSWER:

This is an interesting issue raised by the reviewer. Indeed, as the reviewer points out, the B-mask pseudo-reconstruction should be shifted relative to the ideal borehole scenario. The latter is because the most recent borehole temperature profiles considered in our study are dated in 2002 while the ESM simulations go up to 2005 CE.

We have changed the representation of the B-mask pseudo-reconstruction in all the figures along the document. Now, the shift of the B-mask relative to the IBS case is evident. Note, however, that the differences with the figures presented in the original manuscript for the global case are very small. In addition, this has no influence in the estimation of the 20th century trends, which in the case of the B-mask inversion, have been estimated considering only the period 1900-last-available-date. We have included a mention on this issue in the text. See lines 306-307 of the MSTC.

R1C5: *REVIEWER'S COMMENT:*

5- Which is the difference between GST-mask and B-mask? The authors state on line 306 that GST-mask was built by sampling GST in time, space and depth following the real borehole distribution. But the maximum simulated depth is 42m (35m is the last model node). I find this statement misleading, since borehole depths are much deeper than the simulated depth and it is not indicated which temperature is GST (although I suppose it is GST-L12, see Minor Concerns below).

AUTHORS' ANSWER:

We have corrected the description of GST-mask in this part of the document. Actually, the GST-mask is the

masked version of GST with the actual borehole distribution only in space and time. As the reviewer pointed out, our explanation was misleading in the sense it is not possible to mask GST in depth because the maximum simulated depth is 42 m. The main message here is the analysis of the effect that decreasing spatial sampling with time would have on the representation of the global GST. We hope this correction makes the interpretation of this part clear. Please see lines 305 and 321 of the MSTC. In addition, we have included a table containing a detailed description of the different acronyms employed in the document following a suggestion of Reviewer #2. See R2C3 and Table 4 of the MSTC.

R1C6: *REVIEWER'S COMMENT:*

Trend Analysis: 6- The comparison of trends under different masking configurations constitutes the core of the work, and yet I believe more details are needed regarding the trend analysis. There is no reference to the test applied to determine the significance of the trends. A common t-test would not be suitable for climate series due to the displayed autocorrelation, thus another test should be applied.

AUTHORS' ANSWER:

Indeed, an explanation of the test we applied to the significance of trends was not included in the original manuscript. As the reviewer points out, in the case of temperature time series the regression residuals are not statistically independent and often depict strong autocorrelation. The effect of autocorrelation can be handled by considering an effective sample size in order to account for the non-independence of the residuals. This allows for estimating the significance of the trends considering both an adjusted standard error and adjusted degrees of freedom (Santer et al. , 2000, 2008; Hartmann et al. , 2013). In our case, we accounted for the temporal autocorrelation using a lag-1 autoregressive statistical model. Then, an effective sample size was calculated which is subsequently employed in the estimation of the standard error and degrees of freedom. Finally, the statistical significance of individual trends is obtained assuming that the statistic is distributed as Student's t. Santer et al. (2000, 2008) showed that the significant level computed from adjusted estimates of the standard deviation of regression residuals, the standard error and the t statistic, yield a more conservative estimation than without considering autocorrelation. Furthermore, the use of the effective sample size in the estimation of the critical t value result in an even more conservative approach.

For the statistical significance of trend differences we apply a "paired trends" test following Santer et al. (2008). The test statistic is of the form:

$$d = \frac{(b_1 - b_2)}{\sqrt{s(b_1)^2 + s(b_2)^2}} \tag{1}$$

Where $b_1 - b_2$ represents the trend difference and $s(b_1) + s(b_2)$ are the standard errors of b_1 and b_2, respectively. The latter have been calculated using the effective sample size considering autocorrelation. For the statistically significance we applied a two-tailed test assuming that d is distributed as Student's t. The effective sample size is

also considering to account for the reduced degrees of freedom (Storch and Zwiers, , 1999).

We have included an explanation in the MSTC. See lines 309-313 and 323-324

R1C7: *REVIEWER'S COMMENT:*

7- Also, it is not clear which is the method followed to generate the whisker plots comparing IBS_L12-GST and IBS_SAT-SAT trends in the linear fit case. SAT and GST temperatures are annual series, while the length of the time steps in IBS_L12 and IBS-SAT is 15 years. I guess that the trend of SAT (GST) and IBS-SAT (IBS_L12) were estimated independently and then subtracted, but if so, I do not know if this is the correct approach. Should the annual temperatures be averaged in 15yr-periods, then subtract the reconstructions and estimate the trend of the resulting series? Maybe both approaches yield similar results, but a clarification here would be very helpful.

AUTHORS' ANSWER:

The box-and-whisker plots comparing the pseudo reconstructed IBS_L12(IBS_SAT) with the simulated GST(SAT) were generated by subtracting the individual trends following Santer et al. (2000, 2008). We have performed the analysis following the reviewer's suggestion in order to compare the results of both analyses. The annual time series have been averaged in 15 yr-periods and then subtracted from the pseudo-reconstructions. Then, the linear trend is estimated from the resulting series. Figure 3 of this document shows the results of such analysis for the IBS_L12-GST case (methodological issues); similar results can be expected for the other cases. As the reviewer pointed out, both approaches yield similar results. Note that, even though there are small differences is some of the boxes compared to Fig. 2b of the original manuscript, the overall picture is in essence the same. The differences between IBS_L12(B_mask) and GST(GST_mask) are distributed around 0 and GST(IBS_L12) minus GST_mask(B_mask) are distributed around ~0.2 K century^{-1}. The linear trends for GST and GST_mask in Fig. 3 of this document have been estimated from the 15 yr average series. The resulting trends are slightly lower than the trends estimated from the annual time series, as in Fig. 2b of the original manuscript. Nonetheless, this has no impacts on the results. On the contrary, it shows that the trend estimates are robust regardless the differences between annual and 15 yr averaged time series.

Since the results from both approaches are similar, we keep the trend difference analysis by subtracting the individual trends, as it was originally presented. We have included an explanation of this in the text. See the caption of Fig. 2 in the MSTC.

R1C8: *REVIEWER'S COMMENT:*

8- Crosses in Figs. 6 and 8 should represent the median of the GHG-only and LULConly ensembles, since the authors are comparing against the median of the rest of ensembles (horizontal lines in the boxes). Note that the median is not always equal to the mean of the distribution.

AUTHORS' ANSWER:

We are aware that the median and the mean does not necessarily coincide. However, due to the small sample size in the GHG- and LULC-only (only 3 ensembles members), we initially used the mean over the median in order to account for the information of the 3 ensemble member. The median on the other hand, would be representative of only one of them. Figures 4 and 5 of this document illustrate the effect of including the median instead of the mean for the single-f ensemble. Note that the differences with respect to Figs. 6 and 8 of the original document are small. Thus, using either the mean or the median yield identical results in terms of the dominant influence of one specific external forcing factor on the overall SAT-GST decoupling effect. Therefore, we have included the median of the GHG- and LULC-only ensembles in Figs. 6 and 8 following the reviewer's suggestion. Additionally, we have included some changes in the text related to this issue. See Figs. 6 and 8 and lines 497-501, 505, 508 and 528 of the MSTC.

R1C9: *REVIEWER'S COMMENT:*

9- In line 313, it is stated that "borehole reconstructions are able to retrieve the masked or unmasked GST" based on the results of Figure 2. However, there is no analysis of the trend of the GST-B_mask case. There is an analysis of the IBS_L12-B_mask case, but the IBS_L12 configuration is not the same as the unmasked case. Why not to include the trend analysis of the GST-B-mask case? Something similar can be said about the SAT-B_mask case in Figures 6 and 8.

AUTHORS' ANSWER:

Figure 2 of the original manuscript shows that "borehole reconstructions are able to retrieve the masked or unmasked GST" since either IBS_L12-GST or B_mask-GST_mask differences accurately retrieve the target signal evolution as they distributed around 0. Therefore, we use the IBS_L12 pseudo-reconstruction as a reference GST for the comparison to the masked case, instead of the direct comparison with the simulated GSTs, in order to compare time series of the same type (i.e. 15-yr discretized pseudo-reconstructed time series). This was stated in Section 3.2 (lines 248-254 of the original manuscript; 253-259 in the MSTC). However, the GST-B_mask comparison would yield similar results. Figure 6 of this document includes the box-and-whisker plot for the GST-B_mask differences. Note that it shows a similar picture than the IBS_L12-B_mask column with some relatively small differences in the median (0.15 and 0.17 K century^{-1}, respectively). Similar results can be expected in the SAT-B_mask case.

We think that, for the sake of consistency, it is better to keep the comparison between pseudo-reconstructed time series IBS_L12(IBS_SAT)- B_mask. In addition, this has no effects on the overall results.

SPECIFIC COMMENTS:

Other minor points:

R1C10 : **10- Equation 2 is incorrect. Right term of the equation should display the second order partial derivative of temperature. Check Carslaw and Jaeger (1959).**

Answer: **We have corrected equation 2 accordingly. Line 189 of the MSTC**

R1C11 : **11- Section 3.2: does ST_L12 means GST_L12 as in the rest of the text? If so, please be consistent through the text.**

Answer: **ST_L12 stands for the soil temperature at model layer 12 which is the soil layer we used as reference to create the synthetic BTP's as it is explained in Section 3.2. GST on the other hand, is defined as the temperature directly as the ground surface which in this case that would be the ST at the first model layer (L1; 0.007 m depth). Indeed, this ground surface temperature is the target signal of the borehole temperature reconstructions. Thus, in our study the pseudo-reconstructed GST, obtained from the IBS_{L12}, is evaluated against the model GST defined here as the ST_L1. We have included a definition of GST in order to make this issue more clear. See lines 229 and 259-260 of the MSTC.**

R1C12 : **12- Related to the previous comment, the definition of GST is not clear on the text since Section 4. Is it GST_L12?**

Answer: **We have included a definition of GST. See response to R1C11 and also the response to R2C3.**

R1C13 : **13- Line 301: Which are the trends in Hartmann et al. (2013)? You could add those numbers to the text for an easier comparison with your results.**

Answer: **The trends of the observational databases in Hartmann et al. (2013) have been included. Please note that the trends presented in Hartmann et al. (2013) represent global air surface temperature (not GST) for the 1901-2012 period since there is not available information for the GST at the global scale over this period of time. Please see line 314-315 of the MSTC.**

R1C14 : **14- Line 381: "poor sampling enhances the influence of local behavior". What do authors mean here by local behavior? Please, expand this sentence.**

Answer: **Local behavior refers to the fact that obtaining the regional temperature average from only a few grid points is not totally representative of the whole region. Therefore, the regional mean may be biased by the small sample size (local behavior) relative to the whole region. We have included a short sentence clarifying this issue. See lines 402-403 of the MSTC.**

R1C15 : **15- Line 451: I believe authors mean "50% of the simulated trend".**

Answer: **The reviewer is right, we actually refer to the 50% of the simulated SAT trend. We have included the "missing word" trend in the document. Please see line 474 of the MSTC.**

R1C16 : **16- Lines 308 and 461: which is the level of confidence in the statistical test applied to this trends?**

Answer: **We have included the significance level in the statistical test of both individual trends and trend differences ($p<0.05$). See MSTC lines 309 and 323-324. However, it is worth noting that in line 461, we are not expressing statistically significance but only the fact that SAT minus GST trends over the mentioned areas are evidently larger in the GHG-only ensemble than in the All-F ensemble.**

R1C17 : **17- Line 506: change "bu" by "by"**

Answer: **It has been changed. See lines 532-533 of the MSTC.**

R1C18 : **18- It is not clear which metrics are affected by the different masking configuration is Figs. 2 and 4. Is B_mask affected by those limitations or B_mask is always defined as indicated in Section 4.1? This should be easy to clarify on the captions and on the text.**

Answer: **We agree that there may some confusion about the different masking configuration included in Figs. 2 and 4 since we tagged all of the masking cases as $GST_{mask}(B_{mask})$, even though, they include different masking configurations (i.e. spatial-only, spatial+depth and full spatial+depth+time). We have included a proper explanation of this issue. See lines 336-338 of the MSTC and also the caption of Fig. 4. We hope that this explanation makes the different masking configuration more clear.**

2 Anonymous Referee 2

GENERAL COMMENTS:

The authors would like to thank the reviewers for their constructive suggestions and the time they devoted in reading and proof-reading the manuscript. We have tried to integrate all suggestions and think that the manuscript has improved with them. We do appreciate their contribution. The original manuscript with track changes is included at the end of this document so the specific modifications can be found in the text.

The next sections contain a detailed point by point response to the reviewers comments. Comments are labeled by reviewers and order of appearance, i.e. R2C3 is the third comment of reviewer 2. The original number by the reviewer is also preserved.

R2C0: *REVIEWER'S COMMENT:*

The paper is a meticulous set of synthetic experiments within a climate model space that uses the model reality to generate the data on which the experiments are carried out. That is, the input and output of the signals to be analyzed are know thus results are expected to be self-consistent and provide for the perfect pseudo reality in which experiments can be performed under controlled conditions.

The methodological approach for assessment of the impulse/response analysis are sound and well tested by the borehole reconstruction community and this paper represent a valuable contribution to the subject and to Climate of the Past.

I have several comments that I hope are useful. Some of my suggestions are outside the scope of the paper and I only mention them as suggestions for future work.

AUTHORS' RESPONSE:

The authors welcome the positive perspective of the reviewer on the paper. We are grateful for the reviewer's comments.

Please find below the comprehensive point-to-point response to your review. The original manuscript with track changes (MSTC) is included so the specific modifications can be found in the text.

R2C1: *REVIEWER'S COMMENT:*

- The main issue for me, as the authors mention in the last part of the conclusions, is that all experiments are done in a noise-free environment. Data noise in borehole climatology is extremely important as it has an important effect on the maximum resolution that real data can provide. I assume that the authors are planning a second paper where the methodologies are explored in data and noise environment. I would encourage such paper.

AUTHORS' RESPONSE:

We agree with the reviewer, our experiment represents an ideal noise-free environment that does not fully represent the real-world cases. As the reviewer points out, data noise in real borehole measurements is an important

issue that required a carefully treatment of the data as well as a correct set up of the inversion. A follow up paper that additional consider noise in the experimental set up, as well as other factors not included in the this work, would be desirable. Up to date, the set up described in this paper is the most realistic adaptation of the method to pseudo-proxy experiments. Nevertheless, it would be desirable to continue improving it in the future. We will explore the possibilities to develop such a work.

We have included a note in the conclusions stressing the fact that our experiment is developed in a noise free environment and that further consideration of this issue would be desirable in future works to fully represent the main features in which real-world borehole temperature reconstructions are developed. See MSTC lines 660-667

R2C2: *REVIEWER'S COMMENT:*

-The authors examined the reconstructions based on (noise-free) subsurface temperature anomalies and not from complete borehole temperature profiles as these data are acquired.

AUTHORS' RESPONSE

We appreciate that the reviewer pointed out this issue. Indeed, the synthetic temperature profiles created from the surrogate reality of the model world represent only the transient perturbation induced by the past surface temperature variations. This transient component is superimposed on the background quasi-steady geothermal state. In real borehole temperature profiles (BTPs) the quasi-steady state component (geothermal gradient and equilibrium surface temperature) is usually estimated from the bottom part of the BTPs by linear fitting to the data and extrapolation to the surface. Then, the background components are subtracted from the BTPs to generate the temperature anomalies associated with the downwelling climatic components (Beltrami et al. , 2011). In our case, we directly create temperature anomalies profile in a noise-free environment. We have included an explanation regarding this issue in the text. See lines 180-181, 184-186 and 199-200 of the MSCT and also the response to R2C1 and R2C9.

SPECIFIC COMMENTS:

Minor issues:

R2C3 : **The paper is very dense requiring a lot of effort to keep up with the acronyms.**

Answer: **We have included a table containing a detailed description of the different acronyms employed in the document. See Table 1 of this document and Table 4 of the MSTC. Also, lines 270-271 of the MSTC. We hope this helps the reader to easily keep up with the different acronyms used in the paper.**

R2C4 *Line 36:* **A reference is needed for this claim**

Answer: **We have included the following references: Jansen et al. (2007); Fernández-Donado et al. (2013). See line 37 of the MSTC.**

R2C5 *Line 56-57:* **Depth of borehole temperature profiles was examined in Beltrami et al., 2011. A correction for logging time differential was used in Jaume-Santero et al., 2016.**

Answer: **The point regarding the effect of the depth differences of borehole temperature profiles was also raised by Reviewer #1. Thus, we have included a mention regarding this issue. Please see the response to R1C2. For the logging time differential, actually, we used a similar approach to Jaume-Santero et al. (2016). In our case, global and regional averaging was also done on a yearly basis in order to account for the differences in logging dates.**

R2C6 *Line 69:* **Delete word "global"**

Answer: **The word "global" has been deleted. See line 69 of the MSTC.**

R2C7 Line 206: **Convenient is misspelled**

Answer: **It has been corrected. See line 210 of the MSTS**

R2C8 *Line 210:* **In the noise-free case presented here, the set up of the inversion depends on the choice of model, the geometry of the problem (i.e. the depth of the temperature profile and the depth sampling rate). The sampling rate is not given in the paper. I have assumed it was 1 m.**
In practice, the number of eigenvalues retained in the inversion are also -besides the above mentioned factors - heavily determined by the noise level in the measurements. The tests with four and five eigenvectors retained thus do not have a straight forward meaning as in practical term the noise level would determining the number of principal components retained in the ground surface temperature reconstruction.

Answer: **The sampling rate is indeed 1 m, as it was stated in Section 3.1, line 198 of the original manuscript (lines 200-201 fo the MSTC). Our PPE is indeed developed in a noise-free environment as in the case of previous PPEs (e.g. González-Rouco et al. , 2009; García-García et al. , 2016). In real-world cases the level of noise plays a relevant role on the retained number of eigenvectors. Higher noise limits the number of eigenvectors retained, and thus, the resolution of the retrieved climatic signal (Beltrami and Bourlon , 2004). We have opted for a configuration with a conservative low number of eigenvalues (3) even if this is a noise-free PPE, and actually find consistence between the inverted and the simulated trends. We have additionally shown results for 4 and 5 eigenvectors, as these numbers have been used in previous experimental (Beltrami and Bourlon , 2004; Jaume-Santero et al. , 2016) and PPEs (González-Rouco et al. , 2009) works. We have included a mention regarding this issue in the text. See lines 354-356 of the MSTC.**

R2C9 *Section 3.2:* **Section 3.2 (Pseudo) pseudo-proxy. The pseudo-proxy data generated here consist of subsurface temperature anomalies, not borehole temperature profiles (BTP). The BTP is a superposition of the subsurface temperature anomaly on the quasi-steady state temperate gradient. Perhaps authors should use BT anomaly (BTA).**

Answer: **The synthetic temperature profiles we create are indeed only the temperature perturbation induced by the surface temperature variations. We have included an explanation regarding this issue in the text. See lines 180-181, 184-186 and 199-200 of the MSCT as well as the response to R2C2.**

R2C10 *Line 240:* **Line 240 on: I am confused regarding the generation of the subsurface temperature anomaly field: "Once the spatial distribution of the borehole network is represented in the CESM-LME grid, the LM STL12 series at each of these grid points is trimmed at the actual logging date according to the date distribution (Fig. 1a). Then, the temperature anomalies are calculated with respect to the trimmed period mean".**

Does this mean that each "trimming period mean" or the "trimmed period" is used as a reference to estimate the anomalies? If so, in either case, each anomaly would have a different reference which could complicate the interpretation. In fact, they should take a single common period to estimate all anomalies for the comparison with the IBS case. Then verified that the differences on trimming period means may have something to do with the differences between the red and green inversion results in Figure 2a.

In addition, how is the varying number of boreholes accounted for in the last 30-40 years for the green curve in Figure 2a?

Is the number of BT anomalies for IBS-L12 and B-mask the same in the most recent two inversion steps?

Do subsurface temperature anomalies in the IBS extend to the surface or to 7.8 m (node 12)? Real borehole temperature measurement standard analysis use data below 15-20 m?

Answer: **Reviewer #1 expressed a similar concern. We have performed the test taking a common period to estimated the anomalies. Figure 2 of this document shows the results of such test. Note that it yields almost identical results to the estimation of trends using the trimmed period as we presented in the manuscript. Thus, the differences between the ideal scenario (red line) and the B-mask (green line) cannot be explained by the differences on trimming periods. For further details please see the response to R1C3 for details.**

About the varying number of boreholes accounted for in the last 30-40 years, the global and regional averaging was done on a yearly basis in order to account for the differences in logging dates. Therefore, the number of BT anomalies for IBS-L12 and B-mask is not the same in the most recent two inversion steps. A similar approach has been used in other works that make use of actual borehole temperature profiles measurements (e.g. Jaume-Santero et al. , 2016). For further details, see the response to R1C4.

Finally, the subsurface temperature anomalies in the IBS cases start at 20 m depth. Actually, this was not originally explained in the document. Thus, we have included a proper explanation in the text. See lines 202-203 and also the response to R1C1.

R2C11 *Line 273:* **Line 273: From Figure 2a, it is not clear to me that the temperature anomaly reconstructions capture the MCA-LIA transition. It seems to me that the resolution was lost by 1850 CE. The next sentence in line 273-274 seems to mention this but it seems contradictory.**

Answer: **The pseudo-reconstructed GST from the IBS shows, indeed, a nearly flat transition from the MCA to the LIA. This is in part due to the low multi-centennial variability depicted by the simulated GST itself. The simulated GST shows non-significant negative trends for the period 850-1850 CE. We have modified this statement in the text accordingly. See lines 280-282 of the MSTC.**

R2C12 *Line 276:* **Line 276: "Nevertheless, in model experiments that simulate larger MCA LIA changes the borehole reconstruction is able to recover somewhat warmer temperatures during the MCA (González-Rouco et al., 2006, 2009)." This would depend on the depth of the anomalies and also the number of principal components retained.**

Answer: **As the reviewer points out, the number of principal components retained in the solution play also an important role in the resolution of the retrieve surface temperature histories. Indeed, this is evident in the results of González-Rouco et al. (2009). We have included a mention on this issue in the text. See lines 286-287 of the MSTC.**

R2C13 *Line 346:* **Line 346 and Lines 573-574 "The variability of the depth of the borehole records..." This is so only because the work is based on the analysis of subsurface temperature anomalies that contain little signals below 200 m because of the character of the ESM output used here.**

Answer: **We have included a sentence stressing this fact. See lines 600-602 of the MSTC.**

R2C14 *Line 370:* **Line 370 on, including Figure 3: The number of boreholes in Africa is small, and the area is huge. Much larger that the European slice in Fig 3b. Perhaps, giving the number of sites per unit area may help assess the discrepancies. The red and green lines in Fig 3c, seem contradictory for the cases shown. Are these differences arising from the different initial conditions of each of the 13 simulations? I wonder again whether the referencing over the trimmed period may have something to do with this (see comment on line 240)**

Answer: **There is indeed a relatively low coverage of borehole grid points with respect to the total grid point of the regions we considered (N. America=106/488, Europe=33/191 and Africa=37/215), although the ratio for Africa is comparable to that of Europe. The effect related to this issue was stated in lines 381-384 of the original document (lines 400-404 of the MSTC): "poor spatial sampling enhances the influence of local behavior". Besides the ratio of borehole grid points relative to the total of grid points within each of the regions, the spatial distribution of the borehole sampling is also an important factor. The latter is specially relevant in the most recent decades when the number of available borehole-grid points decreases significantly. Furthermore, these recent borehole logs tend to be cluster over some specific areas, which enhance the local emphasis of "poor spatial sampling". This is particularly the case of the African region where there are only five borehole sites dated after 1990 CE; four of them located in the southern part of the region. In the manuscript, there is a sentence stressing this fact (lines 381-384 in the original document; 403-406 in the MSTC). The spatial sampling, and the decrease of available borehole sites with time, can be inferred from the maps in Fig. 3. We have included a note in the text to draw the**

attention on this issue. See lines 402-406 and also the response to R1C14.

Regarding the case example shown in Fig 3c, we chose a case for each of the regions that represents the median of the distribution in the boxplots. Note that for the African case, the 20th century trends of the red and the green lines are in agreement with the median value shown in the boxplots. An increasing trend of ~0.2 K century^{-1} depicted by the red line and a decreasing trend of ~-0.1 K century^{-1} depicted by the green one. Such difference in the 20th century trends arises from the poor sampling in the last decades to calculate the regional average in the B_{mask} case (green line) as explained above.

R2C15 *:* **I wonder what are the SAT-SAT mask) differences from the ESM simulations?**

Answer: Figure 7 of this document includes the SAT-SAT_{mask} differences. We have included the SAT masked using both the spatial-only mask and the spatial+temporal mask to allow comparison of the effects from the different masking configurations. Note that in both cases the masking leads to an overall underestimation of the global SAT. This is comparable to the effect of masking GST with the spatial-only and the spatial+temporal masks as shown in Fig. 2 of the manuscript, although in the case of SAT the effect is slightly smaller. The SAT-SAT_{mask} trend differences considering spatial-only(spatial+temporal) masking are centered around 0.075(0.077) K century^{-1} (Fig 7 of this document). However, in the spatial+temporal masked case there is a larger spread of SAT-SAT_{mask} differences across the 13 member of the ALL-*F* ensemble, ranging from -0.143 to 0.185 K century^{-1}. The latter is due to the enhanced effect of internal variability produced by temporal sampling. Even the spatial sampling alone can produce differences of almost 0.16 K century^{-1} over a trend of 0.5 K century^{-1}.

R2C16 *Line 632-633:* **Line 632-633 : I would like the authors to expand in this issue. Perhaps, there is a need systematically collect additional borehole temperature profiles.**

Answer: We have included a mention on this issue following the reviewer's suggestion, stressing the fact that borehole temperature reconstructions would be benefited from systematically collecting additional borehole temperature profiles in the future. We understand the logistic and funding challenges, but it would be really useful. See lines 625-626 and 666-667 of the MSTC.

R2C17 **Summary and suggestions:**

This is a good paper worthy of publication in COP.

Although out of the scope of this paper:

- It would be worth examining this problem for other ESM's simulations.

- I would also suggest that the authors consider writing a follow up paper with an identical analysis as in this paper, but based on a set of artificially generated full temperature logs, including simulated data noise. It may be that many of the differences that they observed in the noise-free set of experiments may change; and some differences could potentially be blurred significantly.

Answer: **We appreciate the reviewer considers our work is worth for publication in COP. We agree that examining both the methodological and physical aspects on borehole temperature reconstructions we addressed in this study, using other ESM's simulation would yield valuable additional information on this topic. In fact, The CMIP6/PMIP4 (Eyring et al. , 2017; Jungclaus et al. , 2017) experiments offer an opportunity to address this issue with state-of-the-art ESM. The latter would be specially interesting if new ensembles of simulations with ESMs including both All- and single-forcing experiment were developed, thus allowing to explore the influence of different external forcing factors on the physical-related processes as we did in the present work. Up to date, this is only possible by using the Community Earth System Model-Last Millennium Ensemble (Otto-Bliesner et al. , 2016). Extending this approach to other ESM would yield additional information of the influence of different external forcing factors and different model physics on, for instance, the long-term SAT-GST relationship. It would also help to consider cases of model having a larger temperature response (i.e. climate sensitivity), thus gaining more confidence on the overall effects.**

A follow up paper that considers additional factors in the experimental set up, would also be desirable.We will explore the possibilities to develop such a work. We have included some of these considerations in the text. See lines 660-666 of the MSTC.

Table 1. Abbreviations and acronyms used in this paper.

Acronym	Meaning
GST	Simulated ground surface temperature defined as the first soil layer temperature (ST_{L1}; 0.007 m depth)
SAT	Simulated 2 m air temperature. Original model output TREFHT
IBS_{L12}	Ideal borehole scenario created from ST_{L12} as the boundary condition for the forward model
IBS_{SAT}	Ideal borehole scenario created from SAT as the boundary condition for the forward model
GST_{mask}	GST masked with the realistic representation of the variability of the spatio-temporal distribution of the global borehole network. In some cases GST_{mask} refers to the full spatial+date sampling whereas in other cases it refers only to spatial sampling (i.e. Fig 4)
B_{mask}	Realistic scenario of the borehole temperature inversions including sampling in space, time and depth. It may also refer to the cases in which the sampling is only in space or space+depth (i.e. Figs 2 and 4)

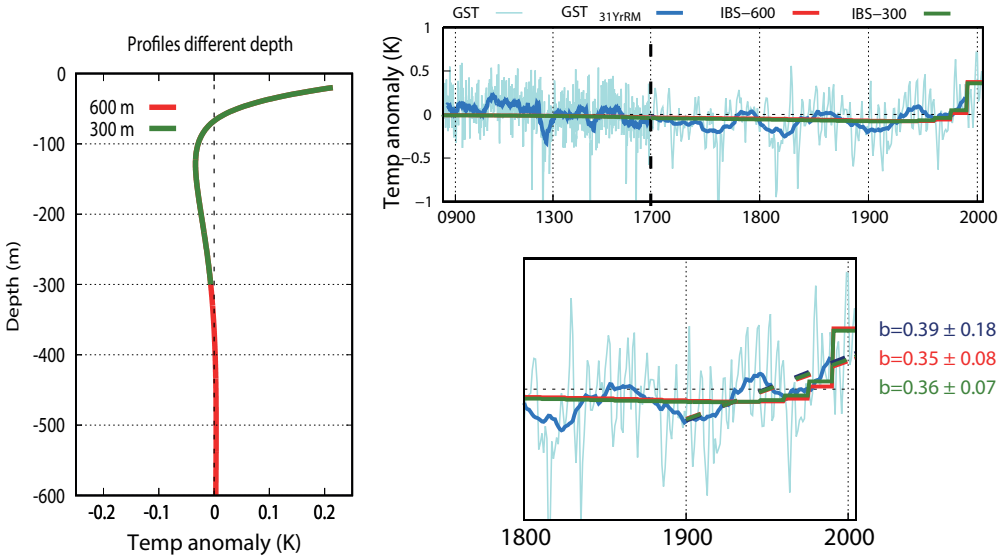

Figure 1. Left: global average of the synthetic borehole temperature anomalies profiles from the ideal borehole scenario (IBS) using a bottom truncating depth of 600 and 300 m for comparison (IBS_{600} and IBS_{300}, respectively). Results are shown for the ALL-F_2 member of the ALL-F ensemble as an example. Right: LM global GST annual anomalies and the corresponding 31-yr filtered outputs, the global IBS_{600} and the IBS_{300} pseudo-reconstructions for the ALL-F_2 realization. Note the different discretization in the x axis after 1700 CE. A zoom of the last 205 year is shown to allow visualization. The dashed lines depict the linear trends for the 1900-2005 CE period. The values are indicated on the right-hand side in K century^{-1}.

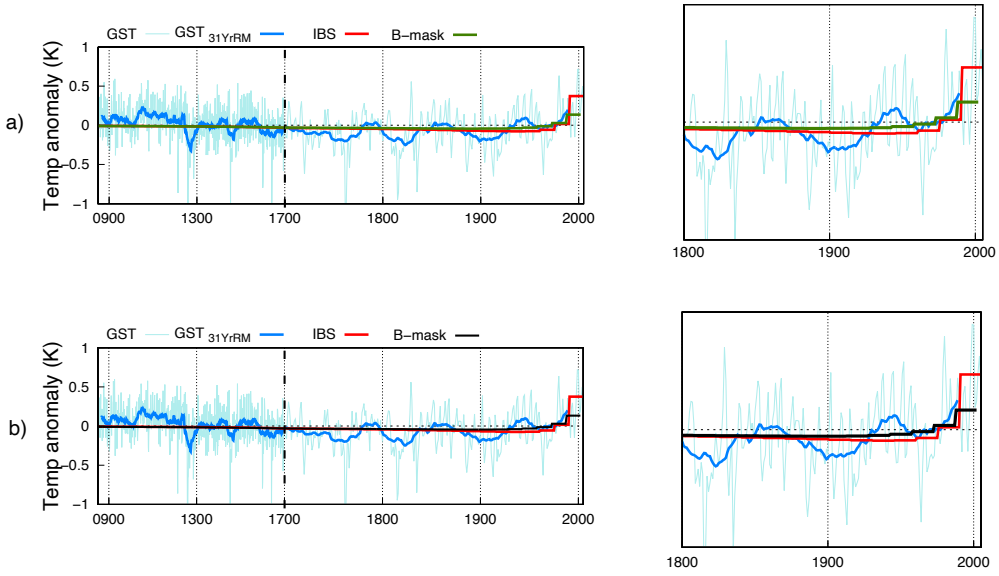

Figure 2. LM global GST annual anomalies and the corresponding 31-yr filtered outputs, the global IBS_{L12}, and the B_{mask} pseudo-reconstructions as it was presented originally in the paper (a), as well as an alternative approach in which a common period (850-1960 CE) has been used to compute the GST anomalies (b). The GST anomalies, the IBS_{L12} and the B_{mask} cases in b) are presented as anomalies with respect to the 850-1960 CE period. Results are shown for the ALL-F_2 member of the ALL-*F* ensemble. Note the different discretization in the x axis after 1700 CE. A zoom of the last 205 year is shown to allow visualization.

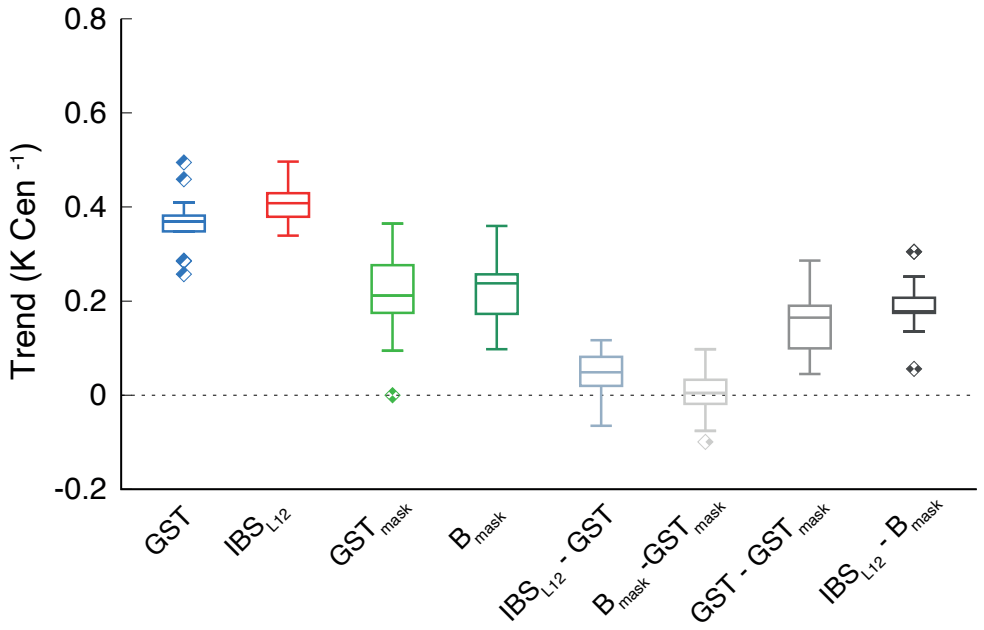

Figure 3. Boxplots describing the centennial trends over the period 1900-2005 CE calculated for each of the 13 ensemble members within the ALL-*F*. The GST, IBS_{L12}, GST_{mask}, B_{mask} and the differences between IBS_{L12} - GST, B_{mask} - GST_{mask}, GST - GST_{mask} and IBS_{L12} - B_{mask} cases are represented. Individual trends are presented as the linear fit to the data after averaged in 15 yr-periods of the annual time series (GST and GST_{mask}). The trends of the differences between IBS_{L12} - GST, B_{mask} - GST_{mask}, GST - GST_{mask} and IBS_{L12} - B_{mask} have been obtained by subtracting the 15 yr averaged series and then a linear fit to the resulting series has been applied. The 25th, 50th and 75th percentiles are indicated and the whiskers represent the lowest/highest value within 1.5 interquartile range (IQR) of the 25th/75th percentile. Outliers are indicated as diamonds in the same color of the box. They represent the values lower/higher than the lower/upper whisker

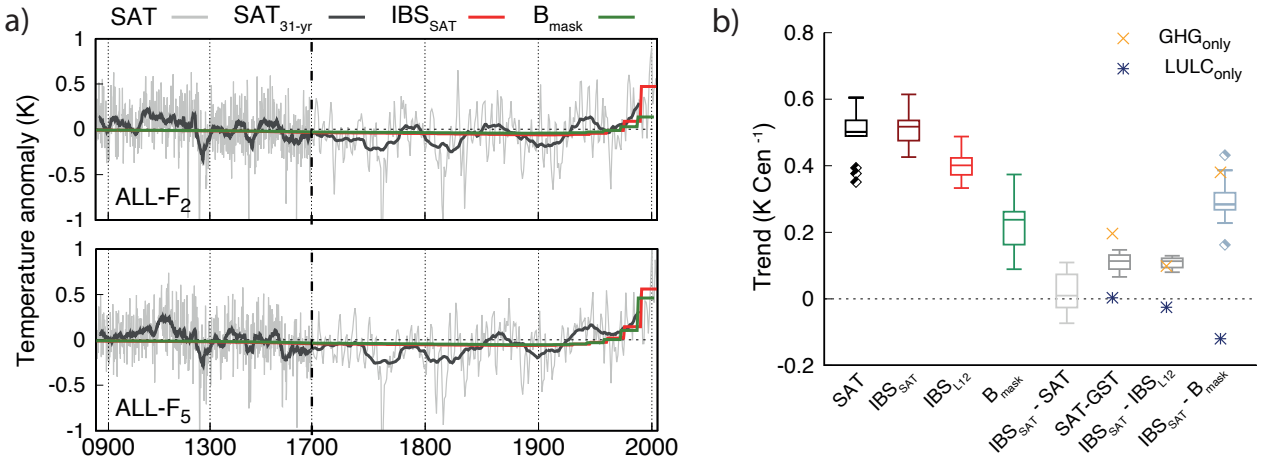

Figure 4. a) LM global SAT annual anomalies and the corresponding 31-yr filtered outputs, the global IBS_{SAT} and the B_{mask} pseudo-reconstructions for the ALL-F_2 and ALL-F_5 members of the ALL-F ensemble. Note the different discretization in the x axis after 1700 CE. b) Estimated linear fit as Fig. 2b, but the SAT, IBS_{SAT}, IBS_{L12}, B_{mask} and the differences between IBS_{SAT} - SAT, SAT - GST, IBS_{SAT} - IBS_{L12} and IBS_{SAT} - B_{mask} cases are represented. Additionally, the median of the GHG- and LULC-only ensembles is indicated by the crosses and asterisks, respectively, in the IBS_{SAT} - IBS_{L12} and the IBS_{SAT} - B_{mask} columns.

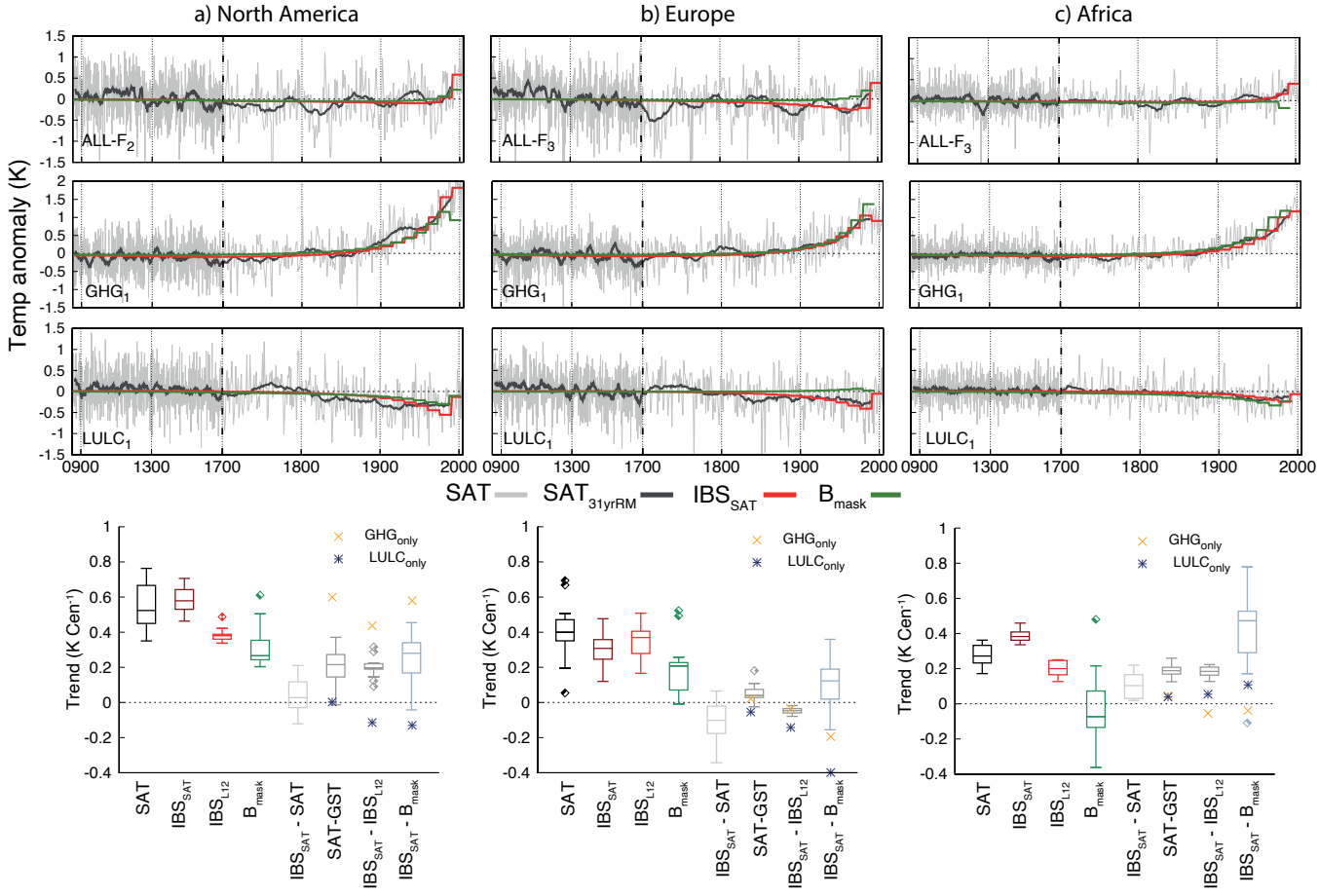

Figure 5. LM regional SAT annual anomalies and the corresponding 31-yr filtered outputs, the global IBS_{SAT} and the B_{mask} pseudo-reconstructions for the ALL-F_2, GHG_1 and $LULC_1$ (from top to bottom) members of the ALL-*F*, GHG and LULC-only ensembles, respectively: North America (a), Europe (b) and Africa (c). Note the different discretization in the x axis after 1700 CE. Bottom panels: as Fig. 6b but for each of the regions presented.

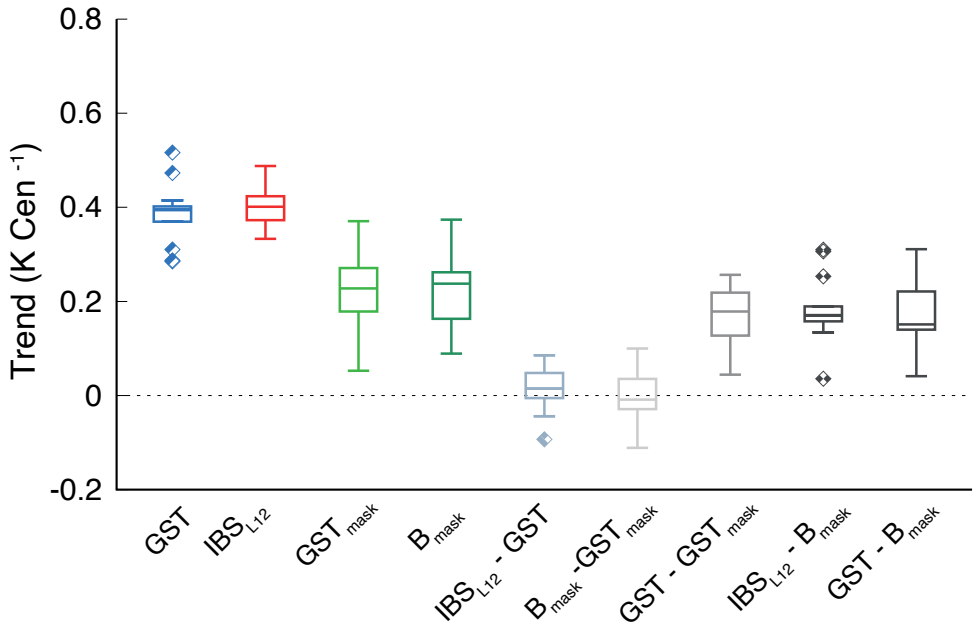

Figure 6. Boxplots describing the centennial trends over the period 1900-2005 CE calculated for each of the 13 ensemble members within the ALL-*F*. The GST, IBS_{L12}, GST_{mask}, B_{mask} and the differences between IBS_{L12} - GST, B_{mask} - GST_{mask}, GST - GST_{mask}, IBS_{L12} - B_{mask} and GST - B_{mask} cases are represented. Individual trends are presented as the linear fit to the data. The trends of the differences between IBS_{L12} - GST, B_{mask} - GST_{mask}, GST - GST_{mask} IBS_{L12} - B_{mask} and GST - B_{mask} have been obtained by subtracting the individual trends. The 25th, 50th and 75th percentiles are indicated and the whiskers represent the lowest/highest value within 1.5 interquartile range (IQR) of the 25th/75th percentile. Outliers are indicated as diamonds in the same color of the box. They represent the values lower/higher than the lower/upper whisker.

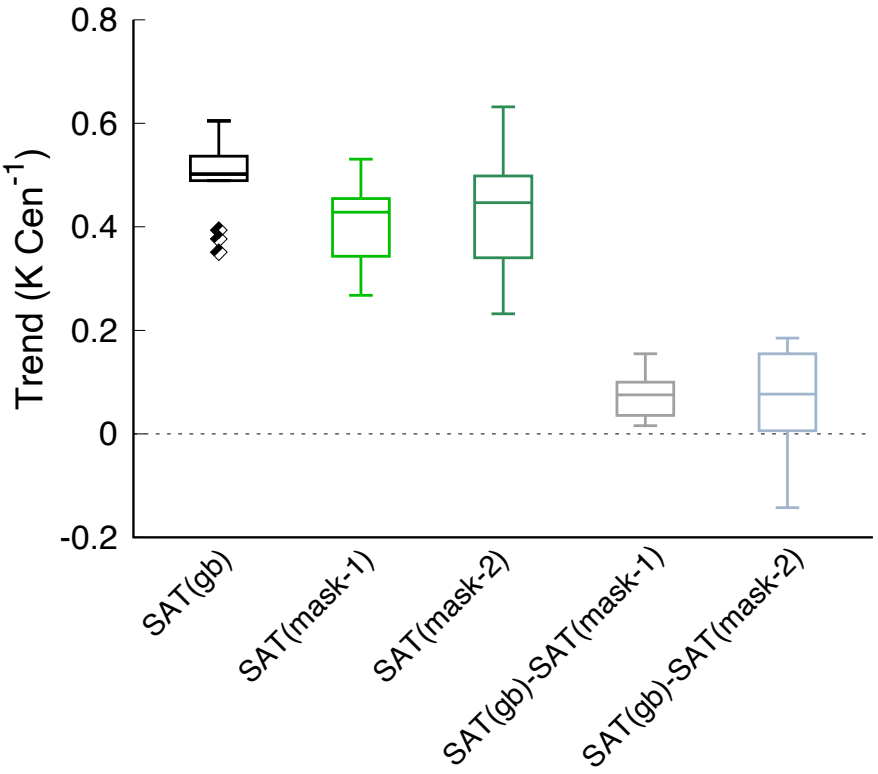

Figure 7. Boxplots describing the centennial trends over the period 1900-2005 CE calculated for each of the 13 ensemble members within the ALL-*F*. The surface air temperature (SAT) global anomalies, SAT masked in space with the real borehole distribution SAT(mask-1), SAT masked in space and time with the real borehole distribution SAT(mask-2), and the differences between SAT(gb)-SAT(mask-1) and SAT(gb)-SAT(mask-2). The 25th, 50th and 75th percentiles are indicated and the whiskers represent the lowest/highest value within 1.5 interquartile range (IQR) of the 25th/75th percentile. Outliers are indicated as diamonds in the same color of the box. They represent the values lower/higher than the lower/upper whisker

References

Beltrami H., and Bourlon E.,: Ground warming patterns in the Northern Hemisphere during the last five centuries, Earth Planet. Sci. Lett, 227, 169-177, 2004

Beltrami, H., Smerdon, J. E., Matharoo, G. S., Nickerson, N.,: Impact of maximum borehole depths on inverted temperature histories in borehole paleoclimatology, Climate of the Past, 745–756, 10.5194/cp-7-745-2011, 2011

Beltrami, H., Matharoo, G. S., Smerdon, J. E.,: Impact of borehole depths on reconstructed estimates of ground surface temperature histories and energy storage, Journal of Geophysical Research: Earth Surface, 120, 5, 763-778, 10.1002/2014JF003382, 2015

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., Taylor, K. E.,: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 5,1937–1958, 10.5194/gmd-9-1937-2016, 2016

Fernández-Donado, L., González-Rouco, J. F., Raible, C. C., Ammann, C. M., Barriopedro, D., García-Bustamante, E., Jungclaus, J. H., Lorenz, S. J., Luterbacher, J., Phipps, S. J., Servonnat, J., Swingedouw, D., Tett, S. F. B., Wagner, S., Yiou, P., Zorita, E.,: Large-scale temperature response to external forcing in simulations and reconstructions of the last millennium, Climate of the Past, 393–421, 10.5194/cp-9-393-2013, 2013

García-García A., Cuesta-Valero F. J., Beltrami H., Smerdon J. E.: Simulation of air and ground temperatures in PMIP3/CMIP5 last millennium simulations: implications for climate reconstructions from borehole temperature profiles, Environmental Research Letters, 11, 4, 044022, http://stacks.iop.org/1748-9326/11/i=4/a=044022, 2016

González-Rouco J. F., Beltrami H., Zorita E., and Stevens B.,: Borehole climatology: a discussion based on contributions from climate modelling, Clim. Past, 5, 99-127, 2009

Hartmann, D.L., A.M.G. Klein Tank, M. Rusticucci, L. Alexander, S. Brfönnimann, Y. Charabi, F. Dentener, E. Dlugokencky, D. Easterling, A. Kaplan, B. Soden, P. Thorne, M. Wild, P.M. Zhai,. 2013.: Observations: Atmosphere and Surface Supplementary Material. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Available from www.climatechange2013.org and www.ipcc.ch.

Jansen E., Overpeck J., Briffa K. R, Duplessy J. C., Masson-Delmontte V., Olago D., Otto-Bliesner B., Peltier W. R., Rahmstorf S., Ramesh R., Raynaud D., Rind D., Solomina O., Villalba R., Zhang D.,: Paleoclimate. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007

Jaume-Santero F., Pickler C., Beltrami H., and Mareschal J.C.,: North American regional climate reconstruction from ground surface temperature histories, Climate of the Past, 12, 2181–2194, 10.5194/cp-12-2181-2016, 2016

Jungclaus, J. H., Bard, E., Baroni, M., Braconnot, P., Cao, J., Chini, L. P., Egorova, T., Evans, M., González-Rouco, J. F., Goosse, H., Hurtt, G. C., Joos, F., Kaplan, J. O., Khodri, M., Klein Goldewijk, K., Krivova, N., LeGrande, A. N., Lorenz, S. J., Luterbacher, J., Man, W., Maycock, A. C., Meinshausen, M., Moberg, A., Muscheler, R., Nehrbass-Ahles, C., Otto-Bliesner, B. I., Phipps, S. J., Pongratz, J., Rozanov, E., Schmidt, G. A., Schmidt, H., Schmutz, W., Schurer, A., Shapiro, A. I., Sigl, M., Smerdon, J. E., Solanki, S. K., Timmreck, C., Toohey, M., Usoskin, I. G., Wagner, S., Wu, C.-J., Yeo, K. L., Zanchettin, D., Zhang, Q., Zorita, E.,: The PMIP4 contribution to CMIP6 – Part 3: The last millennium, scientific objective, and experimental design for the PMIP4 *past1000* simulations, Geoscientific Model Development, 10-11, 4005–4033, 10.5194/gmd-10-4005-2017, 2017

Otto-Bliesner B. L., Brady E.S., Fasullo J., Jahn A., Landrum L., Stevenson S., Rosenbloom N., Mai A., Strand G.,: Climate Variability and Change since 850 CE: An Ensemble Approach with the Community Earth System Model, Bulletin of the American Meteorological Society, 97, 5, 735-754, 10.1175/BAMS-D-14-00233.1, 2016

Pollack H. N., and Smerdon J. E.,: Borehole climate reconstructions: spatial structure and hemispheric averages, J. Geophys. Res., D11106, 109, doi: 10.1029/2003JD004163, 2004

Santer, B. D., Wigley, T. M. L., Boyle, J. S., Gaffen, D. J., Hnilo, J. J., Nychka, D., Parker, D. E., Taylor, K. E.,: Statistical significance of trends and trend differences in layer-average atmospheric temperature time series, Journal of Geophysical Research: Atmospheres, 106, D6, 7337-7356, 10.1029/1999JD901105, 2000

Santer, B. D., Thorne, P. W., Haimberger, L., Taylor, K. E., Wigley, T. M. L., Lanzante, J. R., Solomon, S., Free, M., Gleckler, P. J., Jones, P. D., Karl, T. R., Klein, S. A., Mears, C., Nychka, D., Schmidt, G. A., Sherwood, S. C., Wentz, F. J.,: Consistency of modelled and observed temperature trends in the tropical troposphere, International Journal of Climatology, 28, 13, 1703-1722, doi:10.1002/joc.1756, 2008

Storch, H., Zwiers FW.,:Statistical Analysis in Climate Research, Cambridge University Press, https://doi.org/10.1017/CBO9780511612336, 1999

Methodological and physical biases in global to sub-continental borehole temperature reconstructions: an assessment from a pseudo-proxy perspective

Camilo Melo-Aguilar [1,2], J. Fidel González-Rouco [1,2], Elena García-Bustamante [3], Norman Steinert [1,2], Johann H. Jungclaus [4], Jorge Navarro [3], and Pedro J. Roldán-Gómez [1]

[1]Universidad Complutense de Madrid, 28040 Madrid, Spain
[2]Instituto de Geociencias, Consejo Superior de Investigaciones Cientificas-Universidad Complutense de Madrid, 28040 Madrid, Spain
[3]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), 28040 Madrid, Spain
[4]Max Planck Institut für Meteorologie, Hamburg, Germany

Correspondence: Camilo Melo-Aguilar (camelo@ucm.es)

Abstract. Borehole-based reconstruction is a well-established technique to recover information of the past climate variability based on two main hypothesis: first, that past ground surface temperature (GST) histories can be recovered from borehole temperature profiles (BTPs); and second, that the past GST evolution is coupled to surface air temperature (SAT) changes and thus, past SAT changes can be recovered from BTPs. Compared to some of the last millennium (LM) proxy-based reconstructions, previous studies based on the borehole technique indicate a larger temperature increase during the last centuries. The nature of these differences has fostered the assessments of this reconstruction approach searching for potential causes of bias. Here, we expand previous works to explore potential methodological and physical bias using pseudo-proxy experiments with the Community Earth System Model-Last Millennium Ensemble (CESM-LME). A heat-conduction forward model driven by simulated surface temperature is used to generate synthetic BTPs that are then inverted using singular value decomposition. This procedure is applied to the set of simulations that incorporate all the LM external forcing factors as well as those that consider the concentration of the green house gases (GHG) and the land use land cover (LULC) changes forcings separately. The results indicate that methodological issues may impact the representation of the simulated GST at different spatial scales, with the temporal logging of the BTPs as the main sampling issue that may lead to an underestimation of the simulated GST 20th century trends. Our analysis also shows that in the surrogate reality of the CESM-LME the GST does not fully capture the SAT warming during the industrial period and thus, there may be a further underestimation of the past SAT changes due to physical processes. Globally, this effect is mainly influenced by the GHG forcing whereas regionally, LULC changes and other forcings factors also contribute. These findings suggest that despite the larger temperature increase suggested by the borehole estimations during the last centuries of the LM relative to some other proxy reconstructions, both the methodological and physical biases would result in a underestimation of the 20th century warming.

1 Introduction

Over pre-instrumental times, the climate evolution is estimated from a variety of indirect sources used as proxy indicators of climate variations (Houghton et al., 2001). During the last decades, there have been advances in expanding and improving proxy-temperature reconstructions targeting the common era (CE; Masson-Delmotte et al., 2013). This information offers a frame for addressing the 20th century warming in a broader temporal context, contributing to a more suitable analysis of the forced climate system response with respect to the natural background. Overall, different hemispheric to global scale proxy-reconstructions depict similar climate variability during the last 2000 years. However, the amplitude and timing of past climate states are still not fully constrained. For instance, the magnitude of the temperature variation from the Little Ice Age (LIA; ~1450-1750 CE) to present day displays a range of uncertainty in the estimations of different proxy-sources. While some of them estimate a temperature increase of ~0.6°C (e.g. Ammann and Wahl, 2007), other sources suggest a larger warming of ~1 to ~1.4°C (e.g. Pollack and Smerdon, 2004; Frank et al., 2007). In general, there are uncertainties regarding the amplitude of low-frequency changes within the Last Millennium (LM) as in the transition from the Medieval Climate Anomaly (MCA; ~0950-1250 CE). The magnitude of such changes depends on the climate sensitivity of the system and it is important that reconstructions and climate models are consistent on its estimation (Fernández-Donado et al., 2013). Therefore, assessing the origin and levels of these uncertainties in the amplitude and time of the temperature changes is pivotal in the context of LM climate variations.

One of the proxy reconstructions yielding a larger warming from the LIA to present is that arising from the borehole technique (Jansen et al., 2007; Fernández-Donado et al., 2013). This method is based on the assumption that the soil captures the surface climatic signal due to the conductive heat transport of the ground surface temperature (GST) variations into the subsurface. Therefore, the inversion of borehole temperature profiles (BTPs) provides an estimation of the past GST variations (Pollack and Huang, 2000). In addition, it is also assumed that the surface air temperature (SAT) is strongly coupled to GST (Smerdon et al., 2004), and thus, borehole reconstructions stand as a good proxy of the past SAT variation. Due to the conductive propagation of the temperature signal with depth, only the low-frequency variations are recorded in the subsurface (Pollack and Huang, 2000) and consequently, the inversion of BTPs can offer an estimation of centennial long term trends.

The borehole reconstructed temperatures at hemispheric scales have been subject to assessment and debate during the last decades due to their estimated multi-centennial trend of ca. 1°C over the 1500-2000 CE period, a relatively large value in comparison with other proxy sources (Jones et al., 2009; Fernández-Donado et al., 2013; Masson-Delmotte et al., 2013). It has been discussed whether some methodological limitations and environmental or physical factors may somewhat hinder achieving robust borehole based estimations of past temperature trends (e.g. Rutherford and Mann, 2004; Pollack and Smerdon, 2004; González-Rouco et al., 2009). Methodological aspects refer to a variety of issues that encompass (Pollack and Huang, 2000; Bodri and Cermak, 2007): site specific processes contributing to noise in BTPs, like interactions with orography and hydrology leading to horizontal (vertical) advection (convection) that render the conductive regime assumption invalid; and sampling irregularities such as low density and/or irregular spatial distribution of boreholes at regional and larger spatial scales, regional differences and variability in logging dates and depth of profiles, vertical resolution, errors in measurements,

etc. All these methodological issues may have an impact on the recovery of past GSTs from BTPs. In turn, changes in surface environmental or boundary conditions may have an impact on the recovery of past SAT variations from reconstructed GSTs, i.e. on SAT-GST coupling. Long term variability in surface climate parameters like changes in snow cover, freezing, land use and land cover (LULC) or evaporation changes may therefore play a role in the surface energy balance and thus in SAT-GST interactions (Mann and Schmidt, 2003; Smerdon et al., 2004; González-Rouco et al., 2003, 2009).

The impact of methodological issues on hemispherical scale reconstructions, like the uneven geographical distribution of the borehole sites and the spatial gridding of the data where discussed by Rutherford and Mann (2004) and Pollack and Smerdon (2004) using the observational borehole dataset (Huang and Pollack, 1998). They showed that the spatial aggregation and weighting scheme, used in Huang et al. (2000) and Harris and Chapman (2001) has minor implications for the warming trends indicated by these reconstructions. They also argued that the predominantly mid-latitude distribution of boreholes should be able to capture the northern hemisphere (NH) temperature evolution. Likewise, Beltrami and Bourlon (2004) addressed the spatial aggregation of the borehole sites to obtain NH averages, using gridding instead of geographic aggregation and various grid cell sizes. They reported a 1°C NH GST increase for the 1500-1980 CE period consistent with the Huang et al. (2000) and Harris and Chapman (2001) estimations. Additional sampling issues like differences in timing and depth of borehole logs have received less attention. Most of the logs were done before 1980 CE. Thus, the aggregation of this information should be done with caution as the post 1980 years, when ~~global~~ warming has been larger, are underrepresented. Their influence is thus difficult to assess in studies with real BTPs. Huang et al. (2000) used fixed century-long ramp trends for each BTP to estimate hemispheric long-term temperature variations from 1500-2000 CE. The estimated trend does not fully capture the warming within the last decades of the 20th century, as the bulk of the borehole sites were logged before 1990 CE. Harris and Chapman (2001) reconstructed the 1500-2000 CE NH mid-latitude temperatures from a set of BTPs that were taken to a common time frame. They forward propagated each reduced temperature profile by considering a constant temperature evolution from the logging date up to 1995 CE. Both of these analyses represent a partially muted estimation of the industrial warming since only an small portion of BTPs data include information of the last decades of the 20th century (Pollack and Smerdon, 2004; Jones et al., 2009). Both works yield a similar NH GST increase, and thus an equivalent SAT increase considering SAT-GST coupling, of about ~1°C for the 1500-2000 CE interval.

The sensitivity of the SAT-GST coupling to some physical processes at the surface has also been addressed. The discussion has been focused on whether variations in the long-term behavior of surface properties may result in a biased representation of the SAT histories by the reconstructed GST (Jansen et al., 2007). This thread of analysis has benefited considerably from introducing contributions from modeling. For instance, the influence of snow cover was analyzed by various studies under the hypothesis that changes in snow cover may influence the SAT-GST offset and produce long term drifts in air-ground coupling. Bartlett et al. (2005) studied this issue in multidecadal observational records and several studies (Mann and Schmidt, 2003; Chapman et al., 2004; Schmidt and Mann, 2004) used 50-yr long simulations with the GISS ModelE general circulation model (GCM). Their results suggest that hemispheric scale reconstructions are unlikely to suffer from snow cover biases. Nevertheless, long term changes in snow cover may alter SAT-GST tracking by introducing transient and persistent long term signatures in the coupling, for example by virtue of changes in external forcings. González-Rouco et al. (2003, 2009)

evaluated this at multi-centennial timescale by considering millennial-long simulations of the coupled atmosphere-ocean GCM ECHO-G (Legutke and Voss, 1999) driven by both natural (solar and volcanic) and anthropogenic (greenhouse gases; GHG) forcing factors. They found that in spite of existing decadal variability SAT-GST coupling was reasonably stable at global and hemispheric scales, thus supporting the use of BTPs at those scales. At regional scales, snow cover trends may develop that compromise BTP based reconstructions.

Other issues that may play a role in disrupting SAT-GST coupling through changes in the energy balance and that have received some attention are soil moisture variations (González-Rouco et al., 2009; Cermak and Bodri, 2018) and LULC changes (Cermak et al., 2017; MacDougall and Beltrami, 2017). Melo-Aguilar et al. (2018) addressed these issues by considering for the first time SAT-GST coupling in a large ensemble of simulations of the LM (CESM-LME; Otto-Bliesner et al., 2016) that include a sub-ensemble of experiments involving a realistic set up of all-forcing natural (solar, volcanic and orbital) and anthropogenic (GHG, anthropogenic aerosols and LULC changes) as well as setups of specific single-forcing sub-ensembles. This allowed for separating the effect of each forcing and understanding the integral effects when all forcings were used from the single forcing contributions and feedbacks. At global and hemispheric scales SAT-GST coupling still holds in the all-forcings simulations as a result of compensating effects of various forcings contributing to energy balance. At regional scales SAT-GST coupling is compromised by the influence of specific forcings, and often amplified by changes in snow cover. This is relevant for global and continental BTP based reconstructions because spatial sampling of BTPs should be representative of the target signal and avoid locations where SAT-GST become progressively decoupled.

An evaluation of the impacts of methodological and physical issues on borehole based reconstructions can be done by considering model simulations as a surrogate reality for the actual climate evolution (Smerdon, 2012). The simulations are considered as physically plausible climate realizations, compatible with the external forcings imposed and complex enough to allow for a credible implementation of the reconstruction method. The use of millennium-length GCM simulations has provided a long-term framework to analyze the physical background of the SAT-GST relationship and has allowed for developing such pseudo-proxy experiments (PPEs). For instance, González-Rouco et al. (2006) used LM simulations (1000-1990 CE) from the ECHO-G model in a PPE. They used simulated GSTs and a forward model to simulate BTPs of 600 m depth. These were subsequently inverted following standard procedures in borehole reconstruction strategies and obtained a low pass filtered version of SAT long term trends that could be afterwards compared to the simulated SATs for verification. Thus, they found that the method could appropriately retrieve past SAT long term trends. Additionally to this idealized case in which BTPs would be available for every model gridpoint, they also implemented the method in a more realistic case by limiting the sampling of BTPs to replicate their actual distribution in reality. This exercise rendered similar robust results. González-Rouco et al. (2009) extended this analysis in order to include the effects of the variability in the timing of BTP logging dates and their depths for a specific example over North America. Their results suggest that such variability does not prevent the borehole technique from retrieving the North American 20th century warming. García-García et al. (2016) used the LM all-forcing simulations from state-of-the-art Earth System Models (ESMs) within the frame of the CMIP5/PMIP3 (Coupled Model Intercomparison Project phase 5 / Paleoclimate Modeling Intercomparison Project phase 3; Taylor et al., 2012). They followed the idealized approach

in González-Rouco et al. (2006) sampling the full model grid over land. This allowed for demonstrating the performance of the borehole method over the current generation of ESMs, involving different land models and surface parameterizations.

The previous studies indicate that, globally, GST should be a good proxy of the past long-term SAT variations. Additionally, they support the overall performance of the borehole method at hemispheric and global scales under realistic scenarios involving a full setup of PMIP3 natural and anthropogenic forcings (Schmidt et al., 2011, 2012). In this work, we elaborate from previous analyses and present a set of PPEs in which we implement a borehole reconstruction strategy using LM simulations of a ESM and assess reconstructions performance for a range of spatial scales. We update analyses of methodological and physical influences on GST reconstructions and SAT-GST coupling, respectively, that had not been systematically addressed before. We build on the analysis of Melo-Aguilar et al. (2018) by using the same NCAR coupled ESM LM ensemble (CESM-LME hereafter) and design PPEs to evaluate the influence of sampling methodological issues at global and regional scales by considering the actual distribution of BTPs as well as their depths and logging times. Additionally, we use a sub-ensemble of simulations incorporating either full setup of natural and anthropogenic forcings, the so-called all-forcing experiments (ALL-*F* hereafter), and sub-ensembles of experiments incorporating specific individual forcings, the so-called single-forcing experiments (single-*F* hereafter). This allow for gaining understanding of how forcing factors and their influence on SAT-GST coupling can impact the results on different regions. Particularly, we focus on the GHG and LULC only (GHG-only and LULC-only hereafter) ensembles due their highest potential for impacting the SAT-GST relationship (Melo-Aguilar et al., 2018).

In the first part of the manuscript (Sect. 2), the general characteristics of the model and the simulations employed in this study are presented. Subsequently, Sect. 3 describes the pseudo-proxy configuration considered herein. Results are described in Sect. 4, including the specific effects due to the methodological (Sect. 4.1) and the physical issues (Sect. 4.2). The latter considers also independently the contribution of the LULC and GHG external forcings. Finally, Sect. 5 wraps up conclusions and discussion of the main results.

2 ESM simulations

CESM-LME simulations covering the period 850-2005 CE produced with the Community Earth System Model version 1.1 (CESM1; Hurrell et al., 2013), are used. The CESM1 includes the Atmosphere Model version 5 (Neale et al., 2012) and the Parallel Ocean Program version 2 (Smith et al., 2010) as well as the Los Alamos sea ice model (Hunke et al., 2015). The CESM-LME has a horizontal resolution of ~2° over the atmosphere and land, and ~1° in the ocean and sea ice areas.

The Community Land Model version 4 (CLM4; Lawrence et al., 2011) represents the land surface component in the CESM1. One of the main characteristics of the CLM4 is that the bottom boundary condition placement (BBCP), at 42.1 m depth, is the deepest among the current generation of land surface models with a soil column discretized into 15 layers (Table 1) including up to 5 additional layers in the overlying snowpack. In this respect, the CLM4 includes some improvements in the representation of some surface processes, relative to previous versions and arguably to other models (García-García et al., 2019). These include a better description of ground evaporation, thermal and hydrology properties of organic soils, snow albedo, snow cover fraction and burial fraction of vegetation by snow (Oleson et al., 2010). Such features make the CLM4 specifically suited for

the purpose of this work as they allow for a state-of.the-art representation of the energy transfer between the atmosphere and the soil, a key aspect in the SAT-GST relationship. Further, the relatively deep BBCP allows for a better representation of the downward propagation of the temperature signal at longer (e.g. decadal and centennial) timescales (Alexeev et al., 2007; Stevens et al., 2007; MacDougall et al., 2008).

Table 1. Soil layers and node depths in CLM4. The node depth, which indicates the depth where the thermal properties are defined for soil layers (Oleson et al., 2010), does not necessarily coincide with the center of the layer depth.

Layer	**Layer depth (m)**	**Node depth (m)**
L1	0.017	0.007
L2	0.045	0.027
L3	0.090	0.062
L4	0.165	0.118
L5	0.289	0.212
L6	0.492	0.366
L7	0.828	0.619
L8	1.382	1.038
L9	2.296	1.727
L10	3.801	2.864
L11	6.284	4.739
L12	10.377	7.829
L13	17.125	12.925
L14	28.252	21.326
L15	42.103	35.177

The CESM-LME includes an ensemble of ALL-*F* simulations that incorporates the complete set of agreed CMIP5/PMIP3 LM external forcings (Schmidt et al., 2011, 2012), including the natural and anthropogenic components. Up to date, a total of 13 ALL-*F* simulations are available. Additionally, smaller ensembles of single-forcing simulations, that consider each of the LM external forcings separately, are included (see Table 2 for details and references therein). This work makes use of the ALL-*F* ensemble as well as the GHG- and the LULC-only ensembles. Table 3 presents a detailed description of the original model output as it is described in Otto-Bliesner et al. (2016).

Table 2. External forcing reconstructions used in the CESM-LME, both in the ALL-*F* and the single-*F* sub-ensembles. Legend for external forcing: SOL, changes in total solar irradiance; VOLC, volcanic activity; GHG, concentrations of the well-mixed greenhouse gases CO_2, CH_2, and N_2O; LULC, land use land cover changes; ORB, orbital variations; and OZ/AER, anthropogenic ozone and aerosols.

Forcing	**Reference**
ALL-*F*	-
SOL	Vieira et al. (2011)
VOLC	Gao et al. (2008)
GHG	MacFarling Meure et al. (2006)
LULC	Pongratz et al. (2008) dataset, spliced to Hurtt et al. (2009) at 1500 CE. The only plant functional types (PFTs) that are changed are those for crops and pasture; all other PFTs remain at their 1850 control prescriptions
ORB	The CESM model adjusts yearly orbital position (eccentricity, obliquity and precession) following Berger et al. (1993)
OZ/AER	Fixed at the 1850 control simulation values until 1850 CE when the evolving anthropogenic changes up to 2005 CE. Stratospheric aerosols are prescribed in the model as a fixed single-size distribution in the three layers in the lower stratosphere above the tropopause. The ozone forcing is from the Whole Atmosphere Community Climate Model (WACCM)

Table 3. CESM-LME simulations used in this study. The first and second columns present the acronyms used in this manuscript for the ensembles and the ensemble members, respectively. The ID of the original experiment files is provided in column 3.

Ensemble acronym	**Ensemble member**	**Simulation id**
ALL-*F*	$ALL\text{-}F_i$	b.e11.BLMTRC5CN.f19_g16.0i
	i=1,...,13	i=01,...,13
GHG-only	GHG_i	b.e11.BLMTRC5CN.f19_g16.GHG.00i
	i=1,2,3	i=1,2,3
LULC-only	$LULC_i$	b.e11.BLMTRC5CN.f19_g16.LULC_HurttPogratz.00i
	i=1,2,3	i=1,2,3

3 Experimental design

The theoretical basis for the borehole temperature reconstruction states that the subsurface contains a thermal signature of the past surface temperature variations due to the superposition of the downward temperature signal propagating onto the background geothermal gradient. Therefore, the inversion of BTPs can yield information of the past surface temperature changes. The information of the last 500 to 1000 years is retained within the upper few hundred meters of the subsurface (Beltrami and Bourlon, 2004). Thus, BTPs of at least 300 m depth are required to retrieve the past 500 years (Jaume-Santero et al., 2016) whereas deeper profiles of at least 500 m are necessary to retrieve information of the complete LM (Pollack and Huang, 2000).

Within the model world, the depth of BTPs is limited to the land model depth. Although the CLM4 has the deepest BBCP among the current land surface models, it is still too shallow to directly provide such deep BTPs to account for LM temperature variations (see Table 1) . Consequently, these profiles must be synthetically generated from the surrogate reality of the simulated world as in González-Rouco et al. (2006). The following section provides an overall description of the models employed in this study to simulate the synthetic BTPs as well as for the inversion of the resulting profiles. Further details can be found in Mareschal and Beltrami (1992).

3.1 Forward and inverse models

~~The temperature at any depth z is~~ The BTPs are determined by the combination of the geothermal heat flux , a reference ground temperature and the temperature perturbation $T_t(z)$ induced by the surface temperature variations:

$$T(z) = T_0 + q_0 R(z) + T_t(z) \tag{1}$$

where q_0 represents the surface heat flow density, $R(z)$ is the thermal depth and T_0 is a reference ground temperature. In the forward model, the quasi-steady state component $(T_0 + q_0 R(z)t)$ ~~$T_0 + q_0 R(z)$~~ can be set equal to 0 because the aim is to derive $T(z)$ only as a function of the past surface temperature variations. The forward model, thus, determines the transient perturbation component $T_t(z)$, which can be thought as the anomaly with respect to the quasi-steady state thermal regime.

The propagation of the surface temperature signal within the subsurface is controlled by the one-dimensional time-dependent heat conduction equation (Carslaw and Jaeger, 1959):

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial z^2} \tag{2}$$

with κ as the thermal diffusivity and z and T as depth and temperature, respectively. Equation (2) is solved for an instantaneous temperature change at time t before present as:

$$T_t(z) = \Delta T erfc\left(\frac{z}{2\sqrt{\kappa t}}\right) \tag{3}$$

where $erfc$ is the complementary error function. $T_t(z)$ can be modeled by considering the temperature variations at the surface as a series of K time step temperature changes. In this way, each step imprints a thermal signature in the subsurface that is

merged to the signature of the previous step. Thus, $T_t(z)$ is given by (Mareschal and Beltrami, 1992):

$$T_t(z) = \sum_{k=1}^{K} \Delta T_k \left[erfc\left(\frac{z}{2\sqrt{\kappa t_k}}\right) - erfc\left(\frac{z}{2\sqrt{\kappa t_{k-1}}}\right)\right] \quad (4)$$

where ΔT_k are the surface temperature changes for K time intervals, each value representing an average over time $(t_k - t_{k-1})$.

Equation (4) is used to create synthetic BTPs, using LM surface temperature annual anomalies from the CESM-LME as the upper boundary condition. Although the synthetic temperature profile represent only the transient perturbation component, $T_t(z)$, of the BTP, it will be denoted as BTP thorough the document to avoid confusion. $T_t(z)$ is evaluated at every 1 m depth interval up to a depth of 600 m in order to accommodate for the propagation of the LM surface temperature variations. Subsequently, the upper 20 m of the resulting BTPs are removed in order to avoid the influence of the annual signal and reproducing realistic depths of the water table (Jaume-Santero et al., 2016). The thermal diffusivity used in the geothermal models is 1.5×10^{-6} m^2 s^{-1}, obtained from the values of the bedrock thermal conductivity (3.0 W m^{-1} K^{-1}) and volumetric heat capacity (2×10^6J m^{-3} K^{-1}) of the CLM4 (Lawrence et al., 2011).

For the inversion of the synthetic BTPs, singular value decomposition (SVD; Mareschal and Beltrami, 1992) is applied to retrieve the past long-term surface temperature variations. For the present work, the inverse model consists of a series of 15-yr step changes in surface temperature histories following García-García et al. (2016). Different parameterizations were also tested (e.g. 20-yr or 25-yr step changes; not shown) and showed consistent results with the 15-yr discretization. The latter was finally selected because it yielded a ~~convinient~~ convenient representation of the GTS histories. Likewise, the number of retained singular values is also important. The presence of small singular values leads to an unstable solution dominated by noise while the use of only a few principal components (PCs) result in a smoothed low resolution solution (Beltrami and Bourlon, 2004). The selection of the number of singular values is done by setting an eigenvalue cutoff, from which smaller values are eliminated. In this study we have used a cutoff value of 1.5×10^{-1} from which the solution is the linear combination of the three leading modes. We have additionally explored the effect on the solution of retaining four and five PCs.

3.2 Pseudo-proxy set up

The assessment of the impacts of the methodological and physical issues described above on borehole temperature reconstructions, is done by designing two pseudo-proxy configurations. First, a so-called ideal borehole scenario (IBS) is considered, in which a BTP is simulated at every land model grid point up to a depth of 600 m. This scenario is also characterized by a homogeneous logging date at the end of the CESM-LME simulation period (2005 CE) at every grid point. The latter is achieved by driving the forward model with the annual temperature anomalies calculated with respect to the 850-2005 CE mean. Subsequently, each of these BTPs is inverted and a latitude weighted average is used to obtain the global mean. A similar approach has been used in previous works showing that under such idealized configuration, the borehole technique is able to retrieve the simulated surface temperature variations at global scale (González-Rouco et al., 2006; García-García et al., 2016). Hence, this scenario provides a benchmark experiment that will allow for evaluating the impacts of having some methodological constraints that mimic those in reality. The IBS scenario embraces two cases. On the one hand, the model soil temperature (ST) at layer 12 (ST_{L12} ~7.8 m depth), is used as the upper boundary condition (IBS_{L12}). This case represents the ideal scenario to

generate the pseudo-reconstructed GSTs. On the other hand, the 2 m air temperature model output is employed (IBS_{SAT}) in order to obtain an ideal case of pseudo-reconstructed SATs. We use the ST_{L12} as the reference ~~GST~~ ST to force the IBS_{L12} forward model because in the upper 10 layers, the CLM4 is hydrologically active and it is interesting to keep this realism in the synthetic BTPs. Additionally, this depth is consistent with using annual resolution in the model data as at this level the annual wave is very damped in amplitude and layers below start to filter out decadal timescales. Nevertheless, this decision does not influence results significantly.

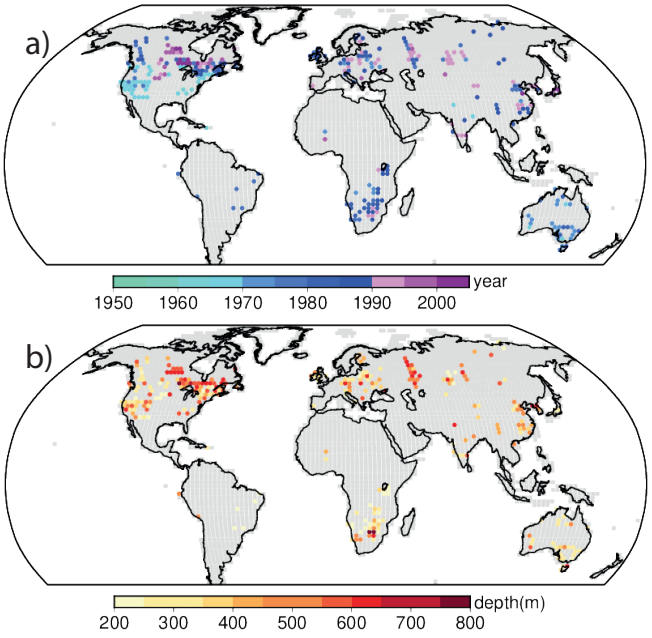

Figure 1. Spatial distribution of a) the logging date and b) the depths in the actual borehole network.

Second, a more realistic arrangement of the available global borehole network is implemented (B_{mask}), including the actual spatial distribution of the borehole sites as well as their real logging depths and dates. This information is obtained from the "Global Database of Borehole Temperatures and Climate Reconstructions" (Huang and Pollack, 1998). Only the BTPs deeper than 200 m are considered in this step since the focus is on multi-decadal to centennial timescales within the LM as in previous reconstructions works (e.g. Huang et al., 2000; Beltrami and Bourlon, 2004). After this selection, a total of 970 logs are retained. For this set up, each of the borehole sites is placed on a specific land model grid point according to their geographic information (Fig. 1). If more than one borehole corresponds to the same model grid point, only one of them is retained prioritizing the most recent one to allow for the climatic signal of the last decades of the 20th century to be represented. Additionally, for the bulk of the cases, the most recent borehole records coincide with the deepest ones. We finally kept 301 grid points that are nearest co-located to any of the real borehole locations. Once the spatial distribution of the borehole network is represented in the CESM-LME grid, the LM ST_{L12} series at each of these grid points is trimmed at the actual logging date

according to the date distribution (Fig. 1a). Then, the temperature anomalies are calculated with respect to the trimmed period mean. The resulting anomaly time series are subsequently used to force the forward model at each grid point, generating a pseudo BTP that is shortened to the actual borehole depth (Fig. 1b), in order to keep this configuration as realistic as possible. In this configuration, the inversion of the individual profiles yields information until the date when they were logged, therefore the reconstructed period is highly variable. The latter seemingly has an impact on the estimation of global averages since closer to present day the number of available sites decreases. Indeed ca. half of the borehole profiles were logged before 1980 CE, and only ~20% of the sites have been measured after 1990 CE. In order to account for such variability in the logging dates, global and regional averaging is done on a yearly basis (Jaume-Santero et al., 2016).

The results from the idealized IBS and the more realistic B_{mask} configuration are then compared in order to evaluate the methodological limitations as well as the potential physical bias. As the methodological aspects are related to the suitability of the spatio-temporal distribution of the borehole network to retrieve the past GST variations, the IBS_{L12} allows for assessing these type of limitations if compared to the B_{mask}; the differences between them informing about the effect of the methodological variants. This assessment is initially developed at global scale, and then, it is extended to smaller spatial domains in order to address the implications on areas with different spatio-temporal borehole distribution. In this part of the analysis only the ALL-*F* ensemble is used. In this work, GST, which is ultimately the target signal of the IBS_{L12}, is defined as the ST at the first soil layer (ST_{L1}; 0.007 m depth) following the same convention as in Melo-Aguilar et al. (2018).

The physical-related aspects and their influence on the estimates are subsequently explored via the comparison with the IBS_{SAT}, because any deviation from the pseudo-reconstructed SAT would be the result of the physical SAT-GST decoupling. Moreover, while the differences between the IBS_{SAT} and the IBS_{L12} cases would inform about the individual contribution of the physical aspects on the interpretation of the SAT variations from the pseudo-reconstructed GST, and the differences bewteen the IBS_{L12} and B_{mask} include the effect of the methodological sampling issues described above, the differences between IBS_{SAT} and the B_{mask} allow for the estimation of the combined effects of both the methodological constraints and the physical processes. Besides the ALL-*F* ensemble, in this analysis the GHG- and LULC-only ensembles are also considered in order to estimate their specific contribution to the potential physical biases. Therefore, the previous PPE setups (IBS_{SAT}, IBS_{L12}, B_{mask}) are run for each of the simulations in the ALL-*F*, GHG- and LULC-only ensembles.

In order to allow for an easy identification of the different abbreviations and acronyms referenced along the document and their precise meaning, Table 4 contains a detailed description of them.

4 Results

The results of generating a variety of PPEs are reported herein. First, those related to methodological sampling issues (Sect. 4.1) and second, addressing SAT-GST coupling (Sect. 4.2).

Table 4. Abbreviations and acronyms used in this paper.

Acronym	Meaning
GST	Simulated ground surface temperature defined as the first soil layer temperature (ST_{L1}; 0.007 m depth)
SAT	Simulated 2 m air temperature. Original model output TREFHT
IBS_{L12}	Ideal borehole scenario created from ST_{L12} as the boundary condition for the forward model
IBS_{SAT}	Ideal borehole scenario created from SAT as the boundary condition for the forward model
GST_{mask}	GST masked with the realistic representation of the variability of the spatio-temporal distribution of the global borehole network. In some cases GST_{mask} refers to the full spatial+date sampling whereas in other cases it refers only to spatial sampling (i.e. Fig 4)
B_{mask}	Realistic scenario of the borehole temperature inversions including sampling in space, time and depth. It may also refer to the cases in which the sampling is only in space or space+depth (i.e. Figs 2 and 4)

4.1 Methodological issues

The comparison of the simulated global GST anomalies within the IBS_{L12} and the B_{mask} pseudo-reconstructions in the ALL-F_2 and ALL-F_5 members of the ALL-F ensemble are represented as an example in Fig. 2a. These members were selected because they represent two possible results of the pseudo-reconstructed GST in the B_{mask} configuration that are discussed herein. Similar results are obtained if other members are selected. In general, the IBS_{L12} pseudo-reconstruction reasonably reproduces the gross features of the low-frequency GST variations over the LM in the CESM-LME. For instance, ~~the transition from the MCA to a colder LIA~~ a small non-significant multi-centennial cooling from the MCA to the LIA can be detected and the warming over industrial times ~~are~~ is successfully captured in both cases. Within the LIA and back to the MCA, the filtering effect produced by the heat diffusion averages over intervals of multidecadal and centennial warming and cooling. Changes in the past, like the simulated MCA warming get damped in BTPs because of this effect and a very smooth version of them is recovered by the reconstruction. Nevertheless, in model experiments that simulate larger MCA-LIA changes the borehole reconstruction is able to recover somewhat warmer temperatures during the MCA if the depth of the anomalies and the number of eigenvalues retained is adequate. (González-Rouco et al., 2006, 2009).

Therefore, the IBS_{L12} reproduces qualitatively the long term trends in both examples of Fig. 2a. However, the masked inversion (B_{mask}) reproduces identical results in the case of ALL-F_5 but diverges from ALL-F_2 in the last decades of the 20th century. This would suggest that the results of B_{mask} can be simulation dependent, i.e. dependent on the different initial conditions and therefore on internal variability. In order to evaluate this, a more quantitative estimation of trends is needed that

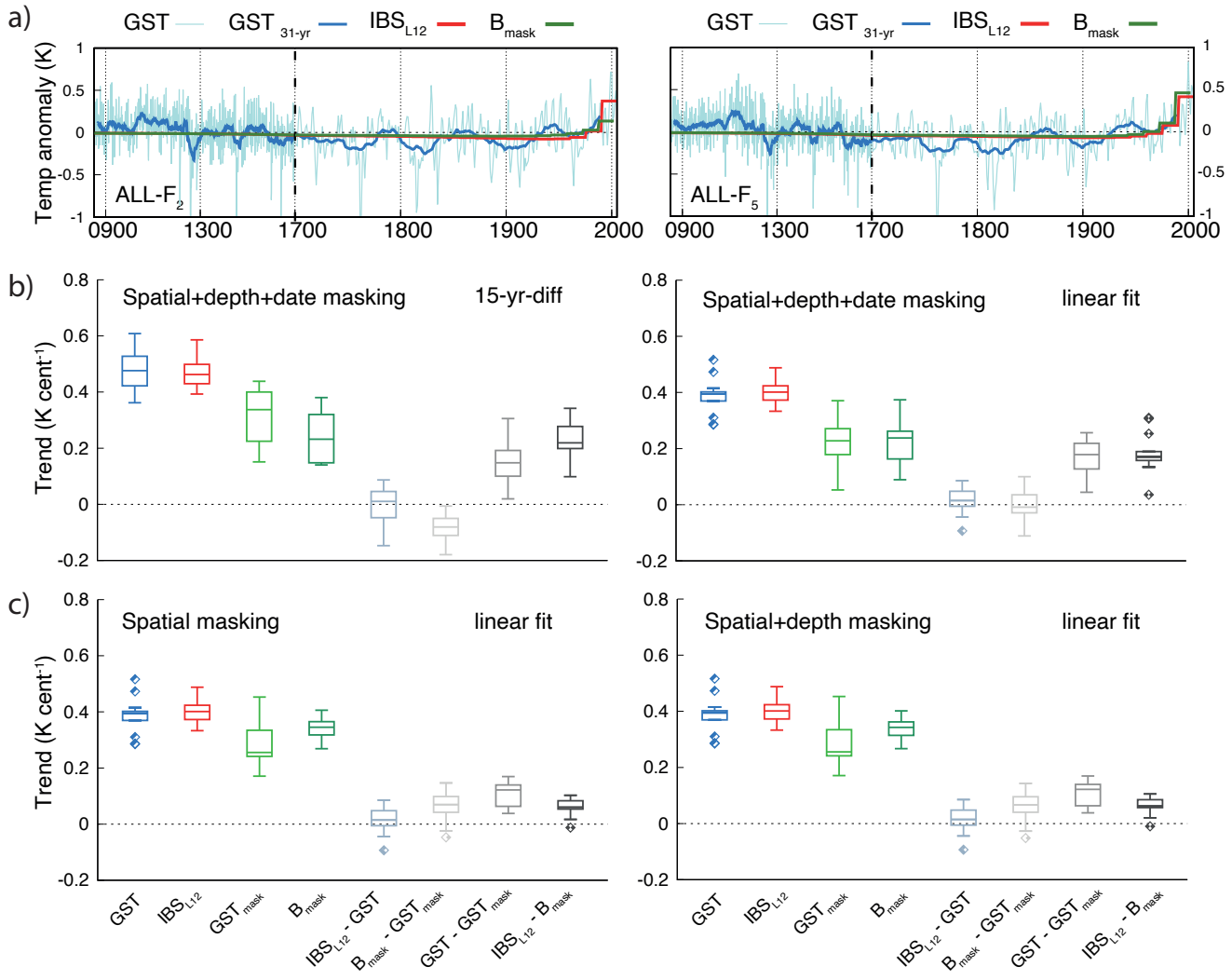

Figure 2. a) LM global GST annual anomalies and the corresponding 31-yr filtered output, the global IBS_{L12} and the B_{mask} pseudo-reconstructions for the ALL-F_2 and ALL-F_5 members of the ALL-F ensemble. Note the different discretization in the x axis after 1700 CE. b) Boxplots describing the centennial trends over the period 1900-2005 CE calculated for each of the 13 ensemble members within the ALL-F after applying space, depth and time masking in the model to mimic real BTP distribution. Boxplots of trend differences has been created by subtracting individual trends following (Santer et al., 2008). c) as in b) but applying spatial and depth masking (right) and spatial only masking (left). The GST, IBS_{L12}, GST_{mask}, B_{mask} and the differences between IBS_{L12} - GST, B_{mask} - GST_{mask}, GST - GST_{mask} and IBS_{L12} - B_{mask} cases are represented. Trends are presented as the 15-yr-diff (b left; see text) and the linear fit to the data calculated over an annual basis (b right). The 25th, 50th and 75th percentiles are indicated and the whiskers represent the lowest/highest value within 1.5 interquartile range (IQR) of the 25th/75th percentile. Outliers are indicated as diamonds in the same color of the box. They represent the values lower/higher than the lower/upper whisker. Note that colors in a) correspond with those in b) and c)

allow for comparing the idealized and masked pseudo-reconstructions and also both of them with the simulated time series. To facilitate this, two approaches are proposed. One of them is to make a simple linear fit of the temperatures, either simulated or pseudo-reconstructed over a reference period considered. As this approach can potentially be affected by the different discretization of the simulated (annual resolution) and pseudo-reconstructed (15-yr time steps), a second strategy is applied in which GST series are transformed to 15-yr averages coinciding with the time steps of the pseudo-reconstructed inversions over the time interval considered. Trends are then calculated by subtracting the mean values of the last and first 15-yr steps and dividing it by the total length of the reference time interval. This allows for verifying the robustness of results. We focus on the 20th century to evaluate trends because: the bulk of the warming takes place in this period; results are less sensitive to the selection of the 15-yr interval than if the 19th century is considered, due to the influence of natural (internal and forced) variability; and it offers the possibility to compare results with instrumental trends (Hartmann et al., 2013). Thus, Fig. 2b shows results for linear fit to the 1900-2005 CE period and for trends calculated from the rates of change between 1890-1905 and 1990-2005 CE (15-yr-diff in Fig. 2b). Figure 2b shows the frequency distribution of trends calculated from both approaches for all members within the ALL-*F* ensemble. Box-and-whisker plots are shown for all possible scenarios considered herein: GST, IBS_{L12}, GST masked with the realistic borehole configuration in space and time (GST_{mask}), B_{mask} as well as the differences among them (IBS_{L12} - GST, B_{mask} - GST_{mask}, GST - GST_{mask} and IBS_{L12} - B_{mask}). In the case of the B_{mask} series, the trend estimation considers the time interval from 1900 to the last avilable date.

Interestingly, the frequency distribution of trends within the ALL-*F* ensemble is similar for both strategies (15-yr-diff and linear fit). The estimated global trends (GST in Fig. 2b) show statistically significant values ($p < 0.05$) that range between 0.3 and 0.6 K century^{-1} across the 13 simulations. Thus, internal variability has an impact in these trends estimates. The significance of the trends is based on a *t* test and accounts for the temporal autocorrelation, using a lag-1 autoregressive statistical model, based on standard procedures for temperature time series (Santer et al., 2008; Hartmann et al., 2013). Likewise, autocorrelation is also accounted for the estimation of the reduced degrees of freedom (Storch and Zwiers, 1999). The trend values are somewhat smaller than those of the observational record that range between 0.73 and 0.83 K century^{-1} over the 1901-2012 period for the global mean surface temperature (Hartmann et al., 2013). In both cases, the IBS_{L12} pseudo-reconstruction yields a reasonable estimation of the global GST increase during the 20th century. Note that the GST trends are slightly larger for the 15-yr-diff relative to the linear fit. While the linear trend indicates a median GST increase of 0.39, the 15-yr-diff suggests a median of 0.47 K century^{-1}. Nevertheless, the IBS_{L12} - GST differences are distributed around zero in either case (Fig. 2b). The comparison of the two methods to estimate the 20th century temperature increase shows that in either case, the pseudo-reconstructed GSTs robustly represent the targeted temperature signal. It is remarkable that when GSTs are masked, i.e. sampled, in space ~~, depth~~ and time (GST_{mask}) following the real distribution, trends take a smaller range of values than those of GST. GST - GST_{mask} differences are above zero and can reach 0.3 K century^{-1}, thus significant and important in the context of the simulated warming. The significance test ($p < 0.05$) of trend differences is based on a "paired trends" test following Santer et al., (2008) and also accounts for temporal autocorrelation. It is remarkable how the level of impact may depend on the interplay of internal variability and BTP sampling. Additionally, the pseudo-reconstructions B_{mask} deliver correspondingly

a distribution of trends that agrees in range with those of GST_{mask}. In the case of 15-yr-diff B_{mask} underestimates slightly, with B_{mask} - GST_{mask} distributing slightly below zero, whereas in the linear case B_{mask} - GST_{mask} are centered over zero.

Therefore, the borehole reconstructions are able to retrieve the masked or unmasked GST. However, sampling plays a role and can produce an underestimation of GSTs. This underestimations depends on the interplay between sampling and internal variability and occurs in some simulations in which the selected points their depths and logging dates are not optimal to represent global warming. Note that in Fig. 2b B_{mask} actually includes the effects of masking, i.e. selecting, pseudo-BTPs co-located to the actual BTP network, trimmed to their actual depths in reality and generated using GST histories up to their logging dates. In turn, GST_{mask} includes the effects of masking, i.e. selecting in this case grid-point time series co-located to actual BTP locations and trimmed to their logging dates; depth masking does not play a role in the case of GST series.

At this point, the question remains on what is the relative role of each of the three masking effects. Figure 2c addresses this by showing similar plots as the linear fit in Fig. 2b but considering only spatial and spatial plus depth masking. Note that even if the masked version is referenced only as $GST_{mask}(B_{mask})$ in the x-axis, however, in Fig. 2c-left(right) the masking includes spatial(spatial+depth) masking. The results for GST, IBS_{L12}, ~~their masked versions~~ and IBS_{L12}-GST differences are identically shown as in Fig. 2b (right) for comparison. Results are virtually identical for both plots in Fig. 2c. This implies that the effects of depth masking are negligible. Nonetheless, this may not be the case in real-world BTPs of different depths because anomaly profiles and elimination of the the background geothermal gradient is done by linear fitting at the bottom of the profile (Beltrami et al., 2011, 2015), thus introducing a source of uncertainty that has not yet been considered in PPE approaches. Spatial masking does have an impact, albeit smaller than if the effects of logging dates are considered, with GST - GST_{mask} differences above zero and below 0.2 K century^{-1}. IBS_{L12} and B_{mask} are effective in retaining their targets.

The selection of the number of singular values to be retained in the SVD inversion may also exert some influence in the estimation of the GST recent trends. Therefore, the impact on reconstructed GST trends of including four and five PCs has been analyzed. The latter is illustrated in Fig. S1 of the supplementary material. On one hand, the solution considering the four leading modes yields similar estimations than the solution based on the three leading modes. Nonetheless, the 4 PCs IBS_{L12} solution is slightly biased to larger values (Fig. S1a). Such behavior is also present in the B_{mask} configuration relative to the GST_{mask}. This pattern is systematically observed in all members of the ALL-*F* ensemble as indicated in the Box-and-whisker plot in Fig. S1a. Note the positively biased IBS_{L12} - GST and B_{mask} - GST_{mask}. On the other hand, the results including the five leading modes yield a solution with some increase of the variance (Beltrami and Mareschal, 1995), noticeable within the 20th century. The simulated GST 20th century trends are biased to smaller values in comparison to those in Fig. 2b; systematically underestimated by the IBS_{L12} 5 PCs solution (Fig. S1b). In real-world data the level of noise in BTPs limits the number of retained principal components (Beltrami and Bourlon, 2004). We have used here a conservative low value that may be consistent with real-world applications including noise.

The hemispheric to global borehole reconstructions in the real-world conditions are far from the idealized scenario described by the IBS_{L12} since they are the result of an aggregation of a limited amount of BTPs that are sparsely distributed over the land surface. Further, there is a large variability in both depth and timing of the records since BTPs are obtained from different sources and they are rarely drilled for the development of climate studies (Jaume-Santero et al., 2016).

The global GST increase during industrial times is however underestimated by the B_{mask} scenario. This feature is common to all ensemble members within ALL-*F* pool of simulations, being the median IBS_{L12} - B_{mask} difference of 0.17 K century^{-1} (Fig. 2b), which accounts for about 43% of the simulated global GST 20th century warming. Most of this underestimation is due to temporal masking. Almost half of the grid points containing BTPs in the B_{mask} case are dated prior to 1982 CE and only ~5% of them present logging dates after 1995 CE (Fig. 1a). As a consequence, many of the synthetic BTPs do not include the information of the last two decades of the simulated period leading to a muted estimation of the global GST recent trend (Pollack and Smerdon, 2004). The variability of the depth of the borehole records, on the other hand, has no influence in the results of the B_{mask} configuration. This is because only BTPs deeper than 200 m have been retained, a depth that has been shown to be sufficient to capture the trend estimates from the LIA to present days and to even retain some features of the MCA to the LIA transition. Likewise, the spatial distribution of the borehole records has a smaller influence on the B_{mask} trends at hemispheric to global scales as indicated in both real-world borehole reconstructions (Pollack and Smerdon, 2004; Beltrami and Bourlon, 2004) and PPEs (González-Rouco et al., 2006).

In spite of the general pattern of discrepancies between B_{mask} and IBS_{L12} during the industrial period (e.g. ALL-F_2; Fig. 2a), there are also cases in which both estimates show a good agreement (e.g. ALL-F_5). This range of variability in the representation of the 20th century trends indicates that the internal climate variability within each realization exerts some influence in the simulated trends of the B_{mask} configuration. As the global average for the last two decades highly depends on the most recently logged BTPs, mostly concentrated in northern Canada, central Europe, Russia and Japan (Fig. 1a), the internal climatic response over these regions apparently play a significant role on the global 20th-century trend estimations. Nevertheless, it would be desirable to expand and update the current dataset of BTPs. In the meantime, it seems advisable to derive strategies that minimize the impact of aging in BTPs (Harris and Chapman, 2001), perhaps by blending BTP information and instrumental temperature observations in the last decades or other procedures that reduce underestimation of multidecadal trends during this period.

The variability in the spatiotemporal distribution of the borehole network may impact the representation of the past GST evolution from BTPs at smaller spatial scales. In order to explore this, three sub-continental domains with different levels of BTPs coverage have been selected. These domains include regions over North America, Europe and Africa. Their selection intents to be illustrative of sampling effects by broadly including existing BTPs in each domain although without necessarily representing optimized domain configurations in any sense. The GST annual anomalies for these regions as well as their corresponding 31-yr running mean low-pass filter outputs, the IBS_{L12} and the B_{mask} estimates cases are shown for one member of the ALL-*F* ensemble as an example for each of the regions (Fig. 3 center). The ALL-F_2 is presented for North America (Fig. 3a) while the ALL-F_3 is shown for both Europe (Fig. 3b) and Africa (Fig. 3c) as these members are representative of the mean behavior of the B_{mask} configuration within the ALL-*F* ensemble. The Box-and-whisker plots (Fig. 3 right) show the 20th century trends from the 13 members of the ensemble as in Fig. 2b using the linear regression. Interestingly, the results of the IBS_{L12} and the B_{mask} configurations show better agreement over the better sampled area of North America than over Europe and Africa. This is evident in the representation of the 20th century trends with a difference in their corresponding median value around 0.08 K century^{-1} in North America. Masking produces differences (GST - GST_{mask}, IBS_{L12} - B_{mask} in Fig. 3) that

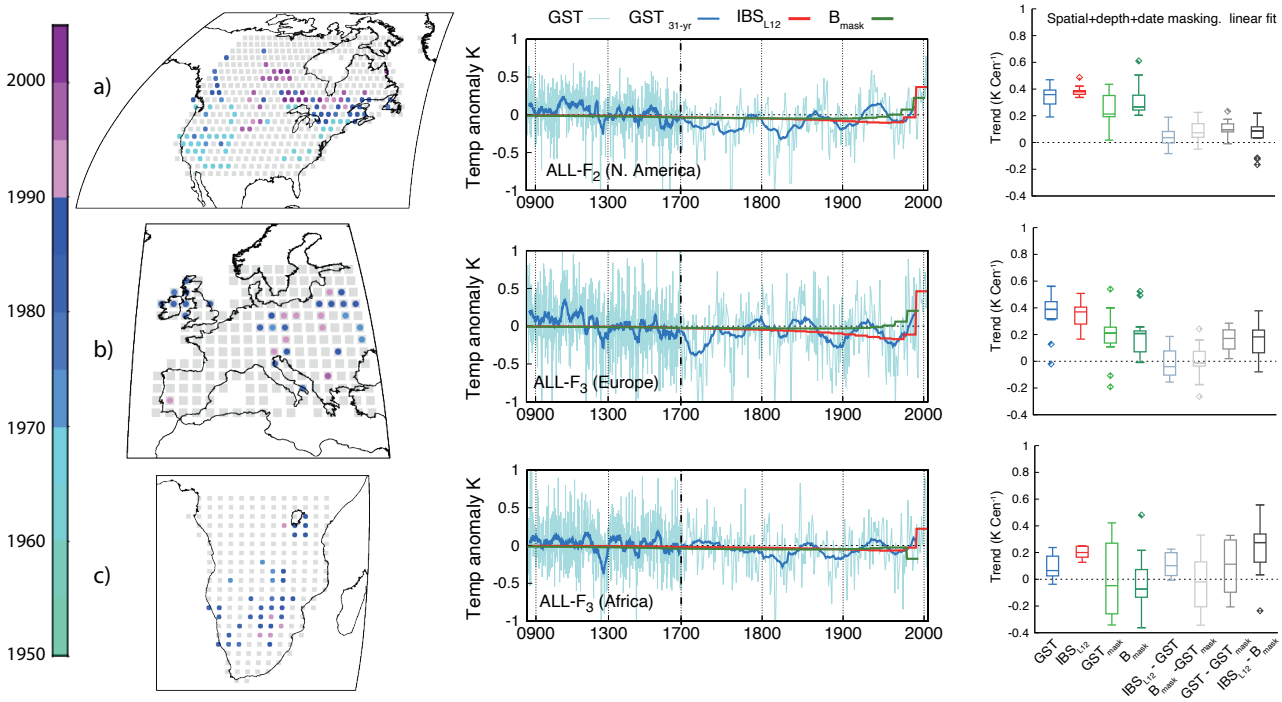

Figure 3. LM global GST annual anomalies and pseudo-reconstructions for three different sub-continental regions: a) North America (31.26-69.16° N, 135-55° W), b) Europe (36.59-59.68° N, 10° W - 30° E) and c) Africa (35.05 ° S - 2.28° N, 10-40° E). The ALL-F_2 member is displayed for the North American continent, whereas the ALL-F_3 member of the ALL-F ensemble is used for the European and African regions. Maps on the left show the spatial distribution of BTPs locations and dates of the actual borehole network for each of the regions. Grey dots show the model grid and the areas defining each of the regions. Right panels: as in Fig. 2a,b but with the 1900-2005 CE trends presented only for the linear fit method.

range from zero to 0.2 K century^{-1}, thus suggesting underestimation biases. The bias is reduced when only spatial masking is considered (Fig. 4), although some outliers still produce GST - GST_{mask} differences above 0.1 K century^{-1}. The size of those differences is not large but it represents between 20-30 % of the simulated trends.

The results for Europe and Africa are more variable. GST trends center around 0.4 K century^{-1} in North America and Europe and below 0.1 K century^{-1} in the African domain considered (Fig. 3). However, trends at these spatial scales can be highly dependent on internal variability and some simulations show no significant trends over the European and African domains. Additionally, poor spatial sampling enhances the influence of local behavior since only few grid points, often distributed within a relatively small area, determine the average of the whole region. Note that the representation of the 20th century trends in both the GST_{mask} and the B_{mask} spreads over a larger range than for North America, especially for the African region. This happens as a response to a decline in the availability of recent BTPs to calculate the regional averages during the last decades of the simulated period (see the maps in Fig 3.) In Europe, the B_{mask} configuration systematically underestimates the

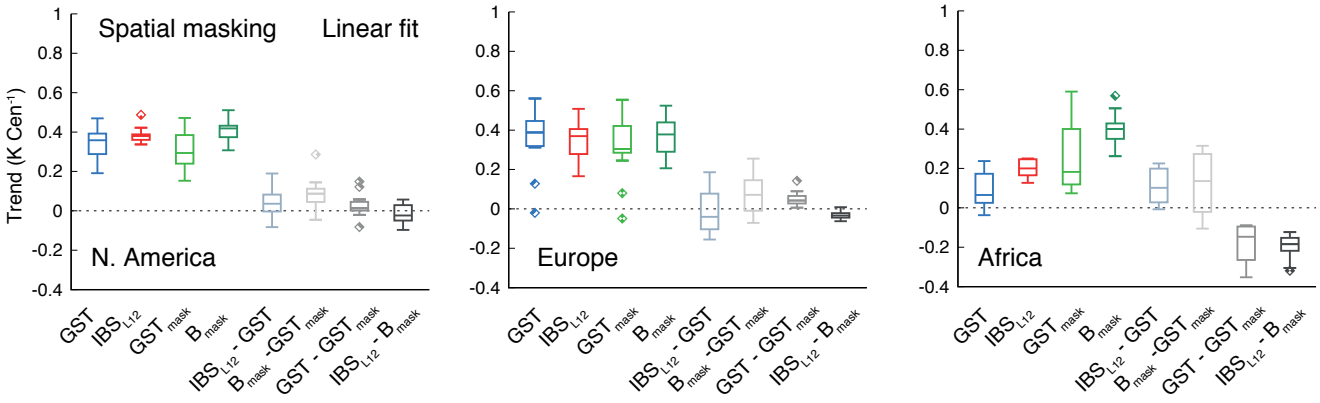

Figure 4. As in Fig. 3 (right) but with the masked cases (i.e. GST_{mask} or B_{mask}) including only spatial masking . Linear fit estimated 1900-2005 CE trends in the ALL-*F* ensemble for North America, European and African regions shown in Fig. 3. Acronyms for GST and PPE cases and their differences follow the same convention as in Figs. 2 and 3.

IBS_{L12} results, but there are cases (not shown) in which the results of the B_{mask} case closely matches the ideal scenario. Thus, these differences among estimates from the different ensemble members suggest an influence of the internal variability on the estimations of the recent temperature trends. Over the African continent, the estimates are more diverse. Therein, the difference between the IBS_{L12} and B_{mask} depicts a larger variability, with a general poor representation of the 20th century GST increase over this region.

When only spatial masking is considered the spread of results is reduced for all regions. For the European domain most of the solutions of GST - GST_{mask} (IBS_{L12} - B_{mask}) get confined below +0.1 (-0.1) K century^{-1} (Fig. 4). For the African domain, the dispersion in GST - GST_{mask} and IBS_{L12} - B_{mask} also shrinks, with both cases indicating overestimation than can be larger than 0.3 K century^{-1} when spatial masking only (Fig. 4) is considered, i.e. the selected grid points indicate larger warming than the rest of the regions. The selected case examples in Fig. 3 (center) illustrate how different the unmasked IBS_{L12} and the B_{mask} solutions can be both in representing the warming of the last decades of the 20th century and also de cooling during the 19th and early 20th century. The latter is most noticeable in the example shown for Europe where clear differences develop during the 19th century between IBS_{L12} and B_{mask}. Here, spatial sampling misses the cooling shown by IBS_{L12}. The reasons for that are discussed in Sect. 4.2.

Therefore, the B_{mask} 20th century trends are sensitive to the spatiotemporal distribution of the borehole network particularly at smaller scales. Whereas in the North American continent there is a better coverage of borehole logs, including most of those recently recorded, in Europe, and especially in Africa, there is a comparatively poorer representation and inhomogeneous spatial distribution of BTPs and particularly, of the recent ones, that ultimately impact the resulting temperature trend estimates. This suggests that the interpretation of the trends from borehole reconstruction estimations at the regional scale should be done with caution over areas with poor spatiotemporal coverage of BTPs (Huang et al., 2000).

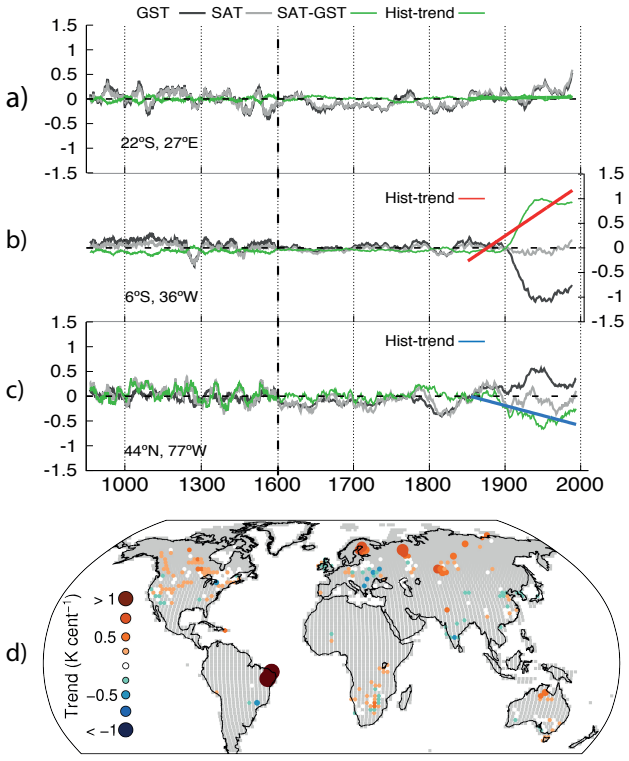

Figure 5. LM evolution of SAT and GST anomalies, SAT minus GST and the linear trends of SAT minus GST over the 1850-2005 CE period for three different grid points with borehole records in the ALL-F_2 as an example. Each grid point illustrates a possible case of the long-term SAT-GST relationship: a) strong coupling during the LM, b) decreasing of GST relative to SAT and c) increasing of GST relative to SAT during industrial times. d) Spatial distribution of linear trends in SAT minus GST anomalies over the 1850-2005 CE period evaluated at every grid point with the presence of a borehole site using the ALL-F ensemble mean. Grid points showing statistically significant trends ($p < 0.05$) are colored in red/blue depending on whether they are increasing/decreasing SAT-GST trends.

4.2 Biases of past SAT borehole-based reconstructions due to physical processes

In addition to the limitations imposed by the spatio-temporal borehole distribution, the reconstructed SAT can also be biased by changes in the long-term SAT-GST relationship. Specifically, changes in anthropogenic forcing after 1850 CE can contribute to SAT-GST decoupling (Melo-Aguilar et al., 2018). Figure 5a-c shows the LM evolution of SAT and GST anomalies relative to the 850-2005 CE mean as well as the difference between both and the linear trend of SAT-GST for the period 1850-2005 CE at selected grid points, illustrating three possible behaviors of the long-term SAT-GST relationship in the CESM-LME. First, a case in which this relationship remains stable during the whole LM (Fig. 5a). Note that there is no trend in the SAT-GST differences. This case represents the ideal strong SAT-GST coupling situation from which the GST would constitute a good proxy of the SAT. Second, two different cases in which the SAT-GST long-term relationship experiences significant variations

are depicted. The direction and magnitude of the SAT minus GST trend are suggestive of the type of impact on the SAT-GST coupling. In the first case (Fig. 5b), the sharp decrease of GST around 1900 CE results in a positive trend in the SAT-GST differences of 1.04 K century^{-1}, that is represented in red indicating a warmer SAT relative to the GST. The second one (Fig. 5c), depicts an example in which the SAT tends to be colder than GST during the industrial period thus, leading to a negative trend of about -0.4 K century^{-1} (blue). These two cases are suggestive of a physical interference affecting the temperature signal contained in the subsurface. Therefore, the inversion of BTPs under such characteristics would yield unreliable information of the past SAT variations. The ALL-F_2 simulation is used as example but similar results can be found in other ALL-F ensemble members.

To provide a spatial view of the borehole locations affected by changes in the long-term SAT-GST relationship in the CESM-LME world, we evaluate the SAT minus GST industrial (1850-2005 CE) trends for the ALL-F ensemble mean at every grid point with the presence of a borehole (Fig. 5d). The ensemble mean is used in order to identify the forced response, however, the internal climate variability in individual ensemble members may result in a different representation of both the magnitude and sign of the trends (Melo-Aguilar et al., 2018). Since the industrial period is affected by pronounced temperature trends due to the anthropogenic emissions of GHGs, the 1850-2005 CE interval is the most adecuate to evaluate the particularities of the SAT-GST linear trends. Additionally, the analysis in this part is intended to determine the reliability of the borehole technique to retrieve the SAT increase over the industrial period. Figure 5d illustrates that at a large number of grid points containing BTPs, the SAT minus GST linear trends are statistically significant ($p < 0.05$) identifying therefore some level of SAT-GST decoupling during the industrial period at such locations. These grid points are represented in red (blue) if the trend of the temperature differences is positive (negative). Additionally, the size of the circles is related to the magnitude of the trend. Therefore, both color and size indicate the direction and magnitude of the SAT-GST decoupling. Note that positive trends are dominant with some larger values at north-eastern Brazil, Fennoscandia, central Eurasia and the north of Siberia. Negative values are also evident, mostly distributed around central and eastern Europe and some more isolated cases distributed around the globe.

In order to assess the influence of the detected long-term SAT-GST decoupling on the representation of SAT from the borehole-based reconstructions at global scale, results of the SAT and GST trends and their IBS pseudo-reconstructions, as well as the B_{mask} case including the effects of sampling are shown in Fig. 6, together with the ALL-F_2 and ALL-F_5 ensemble members as an example. In both of them, the simulated SAT low-frequency variations over the LM are broadly reproduced by the IBS_{SAT}. Note that the simulated SAT 20th century trend is accurately captured by the IBS_{SAT} since the differences between ensemble member trends are distributed around zero (Fig. 6b). Therefore, the IBS_{SAT} can be considered a reasonable representation of the simulated SAT, specifically in the estimation of the 20th century warming. Thus, the comparison between IBS_{SAT} and IBS_{L12} will be informative of any deviation between the pseudo-reconstructed GST and the simulated SAT. Indeed, SAT-GST trend differences distributed over a median value of 0.1 K century^{-1} and consistently, the IBS_{SAT} - IBS_{L12} 20th century trends differences yield positive values, with a median of about 0.11 K century^{-1} and a high level of agreement, i.e. small dispersion, among the members of the ALL-F ensemble (Fig. 6b). This indicates that even under an ideal scenario, the pseudo-reconstructed GST does not fully capture the 20th century SAT warming, missing on average ~20% of the simulated

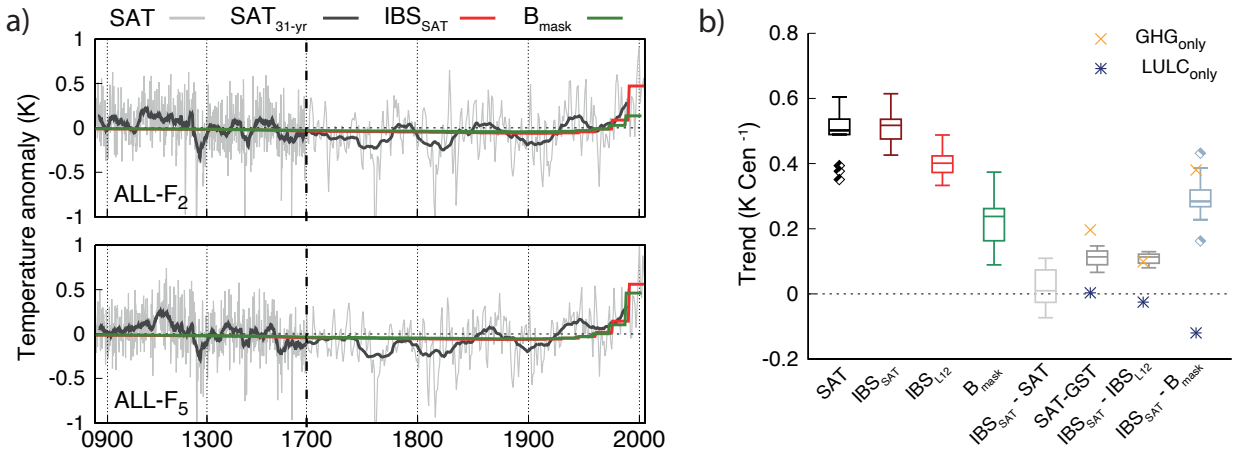

Figure 6. a) LM global SAT annual anomalies and the corresponding 31-yr filtered outputs, the global IBS_{SAT} and the B_{mask} pseudo-reconstructions for the ALL-F_2 and ALL-F_5 members of the ALL-*F* ensemble. Note the different discretization in the x axis after 1700 CE. b) Estimated linear fit as Fig. 2b, but the SAT, IBS_{SAT}, IBS_{L12}, B_{mask} and the differences between IBS_{SAT} - SAT, SAT - GST, IBS_{SAT} - IBS_{L12} and IBS_{SAT} - B_{mask} cases are represented. Additionally, the ensemble ~~mean~~ median of the GHG- and LULC-only ensembles is indicated by the crosses and asterisks, respectively, in the IBS_{SAT} - IBS_{L12} and the IBS_{SAT} - B_{mask} columns.

SAT increase in the CESM-LME as a response to the long-term SAT-GST decoupling. Furthermore, the effect of the physical processes is superimposed to the limitations due to the methodological aspects, leading to larger differences with respect to the more realistic pseudo-reconstructed GST (B_{mask}). In fact, the IBS_{SAT} - B_{mask} 20th century trends differences have a median of 0.28 K century^{-1} (Fig. 6b) that represents more than 50% of the simulated SAT trend. There is however a larger variability of trend differences between these two scenarios that ranges between 0.16 and 0.43 K century^{-1}. This spread results from the internal climate variability in each realization of the ensemble. Such range of variability is noticeable in Fig. 6a if IBS_{SAT} and B_{mask} trends for the two ensemble members are compared to each other.

Melo-Aguilar et al. (2018) showed that the long-term SAT-GST decoupling in the CESM-LME is mainly driven by LULC and GHG changes. LULC changes modify the radiative fluxes at the the surface leading to a different response of the SAT and GST. Likewise, the SAT increase during industrial times as a response to the increase in GHGs may not be fully transferred to the soil due to the insulating effect of snow cover feedbacks. To explore their effects on borehole reconstructions, the analysis described in Fig. 5d is extended to the GHG- and LULC-only ensembles (Fig. 7a). The GHG-only ensemble (Fig. 7a left) the dominant effect is represented by positive SAT minus GST trends, with a similar pattern to the ALL-*F* ensemble over North America, Fennoscandia, northern Russia and Siberia in Fig. 5, although the magnitude of trends is significantly larger. On the contrary, in the LULC-only ensemble negative SAT minus GST estimates are evident, indicating a similar response as in the ALL-*F* ensemble over central and eastern Europe and the Indian subcontinent (Fig. 5). Additionally, the strong positive trends over north-eastern Brazil resemble those found in the ALL-*F* ensemble.

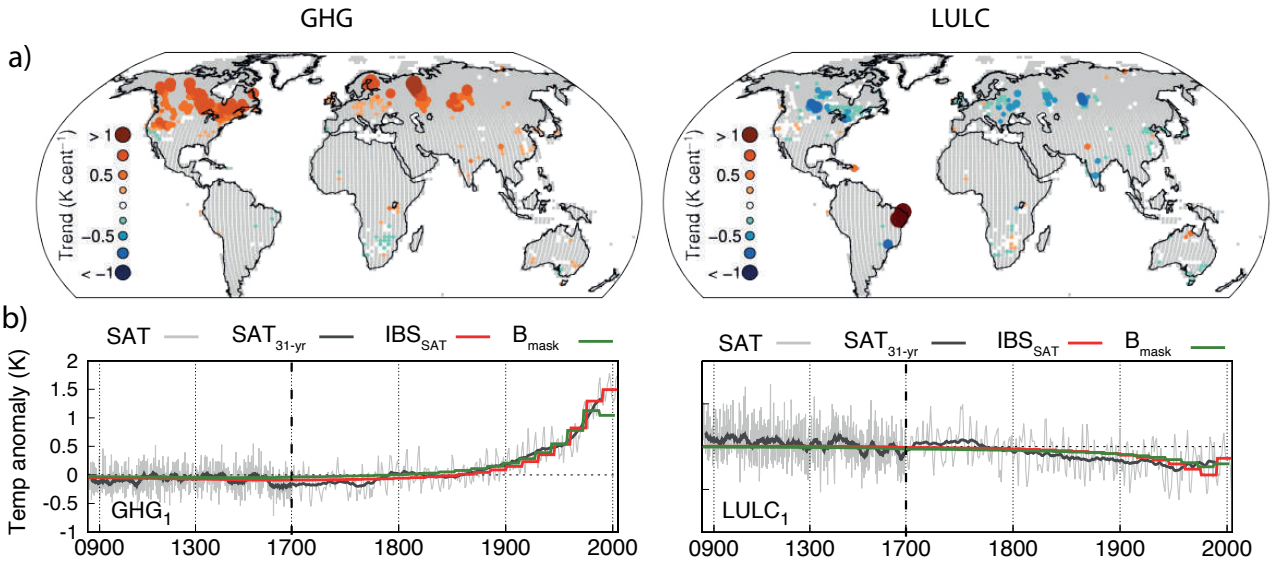

Figure 7. a) Spatial distribution of linear trends in SAT - GST anomalies over the 1850-2005 CE period evaluated at every grid point co-located to a borehole site in the GHG- and LULC-only ensemble mean (left and right, respectively). Only grid points delivering statistically significant trends ($p < 0.05$) are colored. b) LM global SAT annual anomalies and the corresponding 31-yr filtered outputs, the global IBS_{SAT} and the B_{mask} pseudo-reconstructions for the GHG_1 and $LULC_1$ ensemble members. Note the different discretization in the x axis after 1700 CE.

The IBS_{SAT}, the IBS_{L12} and the B_{mask} configurations are also implemented in the GHG- and LULC-only ensembles in order to evaluate their contribution to the physical SAT-GST decoupling at global scale. Figure 7b compares the simulated global SAT anomalies with the IBS_{SAT} and the B_{mask} pseudo-reconstruction in the GHG_1 (Fig. 7b left) and $LULC_1$ (Fig. 7b right) members of the GHG- and LULC-only ensembles, respectively, as an example. From a simple visual inspection it may be noticed that the GHG_1 simulation portrays a similar response than the ALL-*F* estimates (Fig. 6a) since the IBS_{SAT} and the B_{mask} cases diverge during the last decades of the simulated period. On the other hand, in the $LULC_1$ case (Fig. 7b right), there is not such similarity to the ALL-*F* cases. All series and pseudo-reconstructions suggest a consistent cooling in response to LULC changes throughout the last centuries of the millennium. This analysis suggests a larger contribution of the GHG forcing to the SAT-GST decoupling relative to the LULC forcing at the global scale.

To provide a quantitative support to this statement ~~the GHG- and LULC-only ensemble mean differences between the IBS_{SAT} - IBS_{L12} and IBS_{SAT} - B_{mask} 20th century trends are included in Fig 6b (crosses and asterisks, respectively).~~ the median of the differences between the IBS_{SAT} - IBS_{L12} and IBS_{SAT} - B_{mask} 20th century trends in the GHG- and LULC-only ensemble are included in Fig 6b (crosses and asterisks, respectively). Results are almost identical if the mean instead of the median of the single-*F* is included. Note there that in the IBS_{SAT} - IBS_{L12} case the GHG-only ensemble shows positive and very similar estimates than the median of the ALL-*F* ensemble (0.09 and 0.11 K century^{-1}, respectively), suggesting a large

contribution of the GHG forcing to the physical bias in the representation of SAT by the borehole pseudo-reconstructions in the CESM-LME at global scale. On the contrary, in the LULC-only ensemble the IBS_{SAT} - IBS_{L12} difference is negative and very small ~~(-0.03 K Century^{-1})~~ (-0.02 K Century^{-1}) indicating a negligible contribution at this spatial scale. These results are in agreement, as expected, with pure SAT-GST differences, that also receive a larger contribution from GHG at these spatial scales. The IBS_{SAT} - B_{mask} differences further highlight the larger influence of the GHG relative to the LULC forcing. Whereas in the GHG-only ensemble the ~~mean~~ median IBS_{SAT} - B_{mask} difference shows values in the same direction as the ALL-*F* ensemble, in the LULC-only ensemble the difference goes in the opposite direction (negative IBS_{SAT} - B_{mask} trend differences). Melo-Aguilar et al. (2018) reported that the contribution from the GHG forcing to the SAT-GST decoupling is controlled by the reduction in the NH winter snow cover as a response to higher temperatures during industrial times. This situation increases the exposure of the soil surface, previously insulated by snow cover, to the cold winter air, leading to an overall effect of warmer SAT relative to GST at a global scale.

Regionally, the influence of the SAT-GST long-term decoupling on the representation of simulated SAT from the pseudo-reconstructed GST deserves also to be considered since there are geographical variations of this effect (Fig. 5d). Figure 8 illustrates SAT annual anomalies with respect to the 850-2005 mean and the corresponding 31-yr low pass filter outputs as well as the IBS_{SAT} and the B_{mask} cases for the ALL-F_2, GHG_1 and $LULC_1$ simulations over the same areas described in Fig. 3. The spread provided by considering all members of each ensemble are depicted in the boxplots at the bottom of Fig. 8. Interestingly, the decoupling effect appears to be larger over North America and Africa than over Europe. As in the case of the global average, the effect over North America and Africa is shown by the underestimation of the SAT warming during industrial times, indicated by the positive SAT-GST median differences of ~0.2 K Century^{-1} consistent with comparable ~0.2 K Century^{-1} median change in IBS_{SAT} - IBS_{L12} that endorse the performance of the borehole method. On the contrary, in Europe, the SAT-GST decoupling leads to small differences, a negligible under- (over-) estimation of the SAT increase as shown by SAT-GST (IBS_{SAT} - IBS_{L12}).

The single-*F* experiments indicate variability in the influence of both the GHG and LULC forcings over the different areas considered herein. In North America there is a strong contribution of the GHG forcing, that is somewhat counteracted by the LULC influence. This is noticeable not only in the time series in Fig. 8a but also in the sign and magnitude of the SAT-GST and the IBS_{SAT} - IBS_{L12} ~~mean~~ median difference of both the GHG- and LULC-only ensembles (Fig. 8a bottom). Conversely, over the European region, the largest contribution comes from the LULC forcing (see the larger differences in SAT-GST and IBS_{SAT} - IBS_{L12} for LULC-only in Fig. 8b bottom). This is shown more clearly in the comparison of the selected examples in Fig. 8b for ALL-F_2, GHG_1 and $LULC_1$, that show that the LULC cooling dominates the long term trends over GHG warming in the ALL-*F* experiment. For Africa, the ALL-*F* experiment also evidences the damping of the GHG warming produced ~~bu~~ by LULC negative trends. However, other external forcings may also contribute to the response of the SAT-GST relationship at the continental scale in this region (Melo-Aguilar et al., 2018). It is also remarkable that the masked reconstruction (B_{mask}) in the ALL-*F* and LULC cases in Fig. 8b does not capture the long term cooling during the 19th and 20th centuries shown by the unmasked SAT averages (Fig. 8b) and GST averages (Fig. 3b center). Thus, the influence of LULC forcing also explains the different trends between IBS_{L12} and B_{mask} in Fig. 3b (center) during the 19th and early 20th centuries.

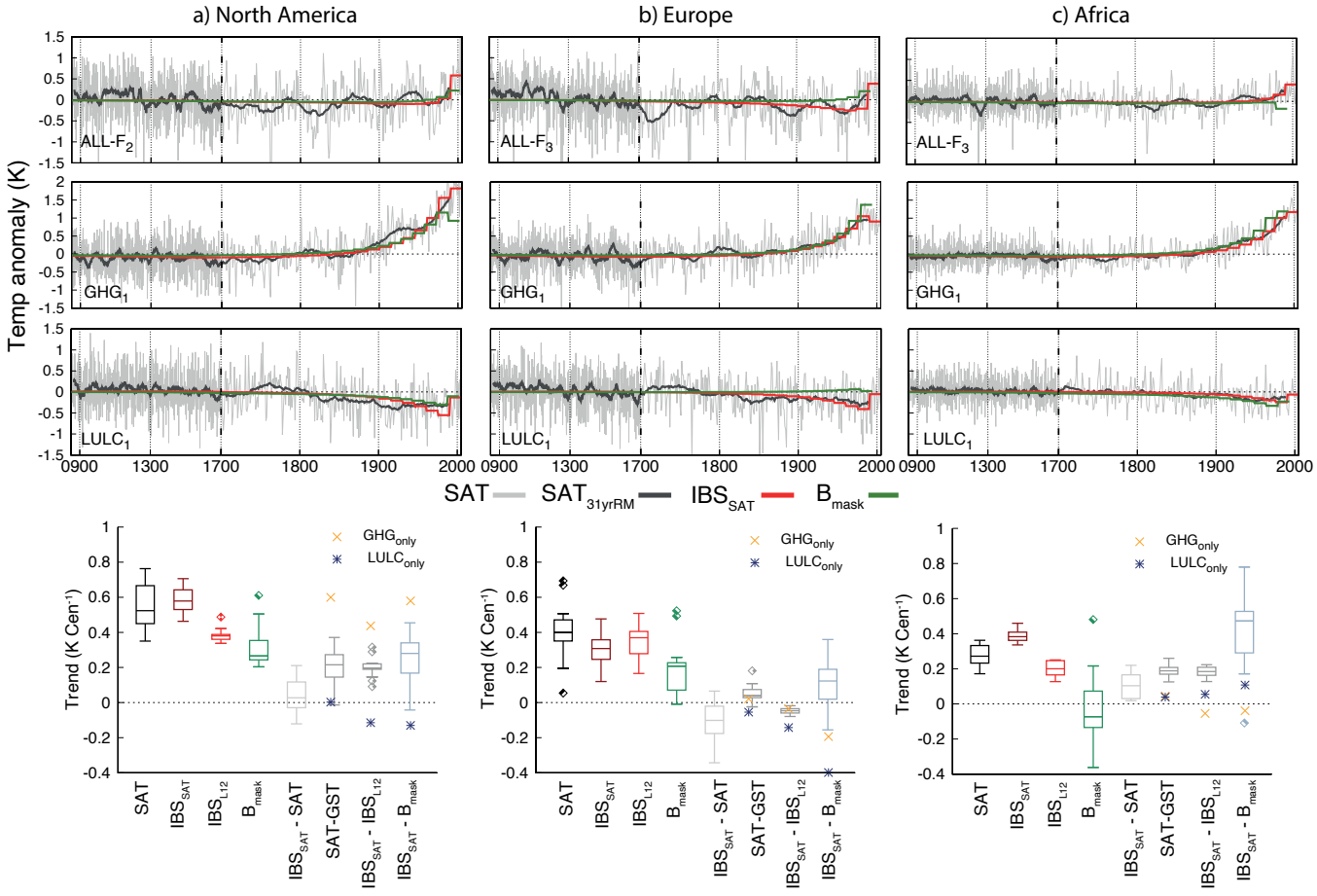

Figure 8. LM regional SAT annual anomalies and the corresponding 31-yr filtered outputs, the global IBS_{SAT} and the B_{mask} pseudo-reconstructions for the ALL-F_2, GHG_1 and $LULC_1$ (from top to bottom) members of the ALL-*F*, GHG and LULC-only ensembles, respectively: North America (a), Europe (b) and Africa (c). Note the different discretization in the x axis after 1700 CE. Bottom panels: as Fig. 6b but for each of the regions presented.

The superposition of the physical biases and the methodological aspects at the continental scales suggest further comments. For instance, for the North American region, while the methodological constraints have a relatively reduced impact on retrieving the simulated GST 20th century increase in the B_{mask} configuration (as discussed in Sect. 4.1), the effect of the physical processes result in a larger underestimation of the SAT 20th century evolution. This is evident in the IBS_{SAT} - B_{mask} difference, that has a median of 0.28 K century^{-1} (Fig. 8), representing ~50% of the simulated SAT 20th century warming. The increment can be compared with the IBS_{L12} - B_{mask} in Fig. 3right. The largest contribution comes from the GHG-only forcing, partially reduced by the LULC-only effect, as shown by the single forcing crosses in Fig. 8. Similarly, in African region, the effects of the physical processes contribute to enhance the underestimation of the SAT signal by the B_{mask}. IBS_{SAT} - B_{mask} differences

expand their spread in Fig. 3 reaching now values of 0.8 K Century^{-1}. On the contrary, for the European region the SAT-GST decoupling slightly counteracts the effects of the methodological aspects mostly via LULC negative biases. However, in both of these regions, the largest contribution to a biased estimation of the SAT 20th century trends comes from the methodological aspects (see SAT - GST and IBS_{L12} - B_{mask} differences in Fig. 3) that outweigh the bias from the physical processes.

5 Conclusions

Borehole based reconstructions lean on two hypothesis to derive the evolution of past temperature trends. One of them is that past GST histories can be recovered from BTPs where the conductive regime dominates; the second one is that the past GST evolution is coupled to SAT changes and thus, the past history of SAT changes can be recovered from BTPs. The first hypothesis can be affected by methodological issues that may distort the recovery of past GSTs from BTPs, whereas the second hypothesis can be affected by physical issues that may distort SAT-GST coupling and thus, the recovery of past SAT changes. This study analyses the performance of the borehole temperature reconstruction technique in a pseudo proxy framework that allows for addressing both methodological and physical issues.

Previous works using PPEs of borehole reconstructions have been implemented in simulations of the LM and focussed on the methodological performance at global (González-Rouco et al., 2006; García-García et al., 2016) and regional scales (González-Rouco et al., 2009), either using the output of a single climate model (González-Rouco et al., 2006, 2009) or an ensemble of PMIP3/CMIP5 LM experiments (García-García et al., 2016). We have extended the analysis of previous works by implementing PPE strategies within the ensemble of simulations of the LM produced with the CESM climate model, the so-called CESM-LME (Otto-Bliesner et al., 2016). We have updated past analyses by introducing a more realistic PPE setup to address both methodological and physical issues at global and regional scales. Additionally, the use of an ensemble of simulations with the same model and different forcing boundary conditions has allowed for considering the influence of internal variability and external forcings on the application of the borehole method.

The methodological implementation used herein adopts a standard SVD approach, as described in González-Rouco et al. (2006, 2009) and García-García et al. (2016). Similar to previous studies, the PPE has been developed in a so-called idealized scenario in which BTPs are assumed to exist at every land model grid point, and produced with a forward model using the complete simulated GSTs, 850-2005 CE. This is equivalent to consider that information from BTPs would be available everywhere in land and with logging dates updated to present times. Thus, in addition, more realistic scenarios considering distributions of BTPs that mimic their actual spatial, depth and logging date distributions have been constructed. The former is used as a benchmark from which the performance of the realistic case is evaluated.

Regarding physical influences on the SAT-GST relationship, we build on from the results of Melo-Aguilar et al. (2018) that demonstrate with this ensemble of CESM experiments that external forcings and related (e.g snow cover) feedbacks can have an impact on SAT-GST coupling, with implications for borehole reconstructions, particularly at regional scales. The work developed herein implements a PPE setup and analysis on the same model ensemble. This allows for considering SAT-GST changes in simulations including a full configuration of natural and anthropogenic forcings and some single forcing simulations

that can aid interpretation of the results. Differences between SAT and GST trends and between borehole reconstructions using SAT and GST generated BTPs allow for understanding about the limits of the second hypothesis stated above.

The CESM-LME has been reported to underestimate by 20% the 20th century trends (Otto-Bliesner et al., 2016). When considering the ensemble including all natural and anthropogenic forcings (ALL-F ensemble), linear fit GST (SAT) trends vary among ensemble members ranging between 0.28 and 0.51 (0.35 and 0.60) K century^{-1}. This inter-simulation variability is an expected result of changing the initial conditions to generate the ensemble and reflects internal variability. We estimate trends during the 20th century as a metric of comparison of simulated and pseudo-reconstructed trends by considering: differences between the final and initial 15-yr periods as well as linear trends; both approaches being consistent in delivering robust results. Results of the so-called idealized IBS PPE setup, in which sampling in time and space are not limited, produce a similar range of GST trends as the simulations, thus supporting the overall performance of the SVD technique. The method is able to retrieve the long term trends through the LM, and the warming during the industrial period. Thus, the SVD approach itself renders reliable results in terms of retrieving the boundary GST signal. The method shows some sensitivity to increasing the number of SVD modes. A number of modes consistent with previous modeling and experimental studies was selected here. Results are robust to small changes in this configuration.

A common result that affects all the subsequent tests in this work is that when methodological and physical constraints are imposed to the borehole reconstruction method, results depend on initial conditions and therefore on internal variability. This is a new although also arguably expected result, as imposing an specific spatial and temporal sampling setup at global and regional scales may produce different effects on 20th century trends depending on the particular trajectory of internal variability and how effective the distribution of BTPs is in grasping the global and regional warming signals embedded within the range of internal variability. Our findings indicate that sampling can introduce detectable biases in borehole reconstructions both at global and regional scales. In the specific setup included herein, considering a realistic distribution of depths does not produce any detectable impact. This may be, in part, due to the little signal contained in the synthetic BTPs below 200-300 m depth because the CESM-LME simulations depict relatively low multi-centennial surface temperature variability. ~~However~~ The distribution of logging dates and BTP locations, on the other hand, does have some impact. At global scales spatial and temporal masking introduces biases that can range between > 0 and 0.3 K century^{-1} (GST - GST_{mask} and IBS_{L12} - B_{mask} differences); spatial only masking reduces maximum differences to 0.1 K century^{-1}. This means that some simulations experience no significant change and in others masked PPE can experience underestimations of several tenths of a degree depending on the realization of internal variability. These are indeed small numbers but bear in mind that even a 0.1 K century^{-1} change represents about 20% of the total warming in these simulations.

At regional scales, the impacts of temporal and spatial sampling vary across regions with differences (e.g. IBS_{L12} - B_{mask}) that can range 0-0.2 K century^{-1} in a relatively well sampled region as North America, to 0.4 K century^{-1} in Europe, or 0.6 in the African domain. Most of it is due to the temporal sampling effect, particularly in the American and European regions where it gets reduced to values below 0.2 and 0.1 K century^{-1}, respectively, when considering spatial only masking; in the African domain spatial biases are larger and range between 0.1 and 0.3 K century^{-1}. At regional scales, some of these biases are larger

than their target temperature trend values. Thus, at regional scales spatial and temporal sampling is an issue. The effects are smaller over North America and larger over the other regions tested.

Temporal logging of the borehole records stands as the main sampling aspect contributing to the reduced skill in capturing the 20th century warming. This is because a large portion of the BTPs do not contain information of the warming during the last decades of the 20th century due to their relatively old logging dates. Such result suggests that the availability of recent measurements highly influences the results of this type of temperature reconstruction. The continental scale analysis has provided further insight on this issue since the pseudo-reconstructed GST yields a generally good estimation of the simulated GST evolution during industrial times for areas with a relatively good distribution of recent BTPs measurements, as the North American continent. On the contrary, this accuracy is lost as the availability of recent BTPs is reduced, as for instance, over the African region. The temporal effects should be relatively small if the reconstructions are considered only up to the dates of the oldest boreholes, at the expense of missing much of the warming developed during the last decades. Alternatively, strategies may be considered that would blend information from early borehole profiles with local instrumental data to mitigate the missing trend effect (Harris and Chapman, 2001). In addition, this type of analysis would benefit from re-logging boreholes whenever possible and logging additional BTPs in the future; thus updating the network. Regarding the spatial sampling effects, the definition of the domains considered herein has been somewhat subjective. Perhaps more ad hoc domain setups can be specifically considered that reduce the effects of spatial sampling like in Africa. Otherwise, cases like the one selected in the African domain, including very sparse sampling, should be avoided.

In the surrogate reality of the CESM-LME, the interpretation of the simulated SAT, derived from the reconstructed GST is additionally impacted by the physical processes that interrupt the long-term SAT-GST coupling. In the idealized scenario, the GST pseudo-reconstruction does not fully capture the global SAT increase during industrial times, missing about 20% of the simulated warming on average in the ALL-*F* ensemble. Globally, the larger increase of SAT relative to GST during industrial times arises from the dominant influence of the GHG external forcing. Nevertheless, the contribution of individual forcings varies geographically as indicated by the regional analysis. While over the North American continent the overall response is similar than at the global scale, with a large contribution of the GHG forcing, over Europe and Africa the SAT-GST decoupling is dominated by the LULC forcing.

The impact of the long-term SAT-GST decoupling is superimposed on the limitations due to the methodological aspects. At a global scale, the combined effect results in a biased representation of the simulated SAT 20th century trend by the realistic configuration, missing more than 50% of the simulated SAT (average values for the ensemble). Indeed, at this spatial scale, the largest contribution comes from the methodological aspects. Nonetheless, this may be different at smaller spatial scales. For instance, in North America, where the impact due to the methodological aspects is relatively small, the effect of the SAT-GST decoupling deteriorates the representation of the simulated SAT from the pseudo-reconstructed GST.

Overall, our results indicate that the combined effect of the methodological constraints and the physical SAT-GST decoupling leads to a systematic underestimation of the SAT 20th century trends at global scale; at regional scale overestimations can occur as in the African domain. Nonetheless, it is worth noting that despite this general pattern among the members of the ALL-*F* ensemble, there may be cases in which these impacts are relatively small or even disappear, consistent with previous studies

(González-Rouco et al., 2006; García-García et al., 2016). This influence of internal variability and comparison with the actual case in the real world can be better evaluated over reanalysis simulations of the 20th century (Hartmann et al., 2013), currently underway.

~~The results from PPEs herein are informative but~~ Even if the analysis included herein represents the most realistic implementation to date of the borehole method in PPEs, results are not meant to be directly translated to the real-world cases. Nevertheless, ~~they~~ PPEs provide valuable information about the uncertainties of paleo-reconstructions in a controlled experimental framework (Smerdon, 2012). Despite the results of this work indicate a clear pattern in the potential sources of bias that can be found in the borehole-based reconstruction, they should be interpreted with caution for real-world applications since several aspects may influence the level of impact. For instance, the use of other ESMs with different climate sensitivities, i.e. larger 20th century warming, and different representation of the influence of changes in land surface physics that could lead to a different representation of changes in the SAT-GST relationship. Additionally, the limitations in the local representation of sampling due to model resolution and more technical issues like the existence of local noise in BTPs have not been considered here. Exploring these issues in future works would be desirable in order to have a more complete evaluation of the method. The latter would be specially interesting if new ensembles of LM simulations with ESMs including both All- and single-forcing experiments were developed in the frame of the CMIP6/PMIP4 (Coupled Model Intercomparison Project phase 6 / Paleoclimate Modeling Intercomparison Project phase 4; Eyring et al. , 2017; Jungclaus et al. , 2017, respectively). This would allow exploring the influence of different external forcing factors and different model physics that have some influence on, for instance, SAT-GST decoupling. Up to date, this issue can only be addressed with the use of the CESM-LME, as we have done in this study. Additionally, this work clearly supports the need for updating and expanding the borehole network. More and, if possible, deeper and good quality BTPs are needed.

In light of the results of this work, both the methodological issues and the physical biases of the borehole-based reconstructions would lead to an underestimation of the temperature increase from the LIA to present day, specially over the industrial period. These findings are not able to explain the larger temperature increase suggested by the actual borehole estimations during the last centuries of the LM relative to those of some other proxy-based reconstructions.

Code and data availability. The code used in this analysis and results are available from the authors upon request.

Author contributions. This study is part of CMA's PhD. The experimental design was set up by CMA and JFGR. CMA carried out the data processing, analyzed the results and wrote the paper. All of the other authors contributed to the analysis and discussion of the results and to writing the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research was supported by an FPI grant (grant no. BES-2015-075019), from the Spanish Ministry of Economy, Industry and Competitiveness. We gratefully acknowledge the IlModels (CGL2014-59644-R) and GreatModelS (RTI2018-102305-B-C21) projects. We also thank the CESM1(CAM5) Last Millennium Ensemble Community Project and supercomputing resources provided by NSF/CISL/Yellowstone.

References

Alexeev, V. A., Nicolsky, D. J., Romanovsky, V. E., and Lawrence, D. M.: An evaluation of deep soil configurations in the CLM3 for improved representation of permafrost, Geophysical Research Letters, 34, n/a–n/a, https://doi.org/10.1029/2007GL029536, http://dx.doi.org/10.1029/2007GL029536, l09502, 2007.

Ammann, C. M. and Wahl, E.: The importance of the geophysical context in statistical evaluations of climate reconstruction procedures, Clim. Change, 85, 71–88, doi:10.1007/s10 584–007–9276–x, 2007.

Bartlett, M. G., Chapman, D. S., and Harris, R. N.: Snow effect on North American ground temperatures, 1950-2002, J. Geophys. Res., 110, F03 008, doi:10.1029/2005JF000 293, 2005.

Beltrami, H. and Bourlon, E.: Ground warming patterns in the Northern Hmeisphere during the last five centuries, Earth Planet. Sci. Lett., 227, 169–177, 2004.

Beltrami, H. and Mareschal, J. C.: Resolution of ground temperature histories inverted from borehole temperature data, Glob. Planet. Change, 11, 57–70, 1995.

Beltrami, H., Smerdon, J. E., Matharoo, G. S., and Nickerson, N.: Impact of maximum borehole depths on inverted temperature histories in borehole paleoclimatology, Climate of the Past, 7, 745–756, https://doi.org/10.5194/cp-7-745-2011, https://www.clim-past.net/7/745/2011/, 2011.

Beltrami, H., Matharoo, G. S., and Smerdon, J. E.: Impact of borehole depths on reconstructed estimates of ground surface temperature histories and energy storage, Journal of Geophysical Research: Earth Surface, 120, 763–778, https://doi.org/10.1002/2014JF003382, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JF003382, 2015.

Berger, A., Loutre, M.-F., and Tricot, C.: Insolation and Earth's orbital periods, Journal of Geophysical Research: Atmospheres, 98, 10 341–10 362, https://doi.org/10.1029/93JD00222, http://dx.doi.org/10.1029/93JD00222, 1993.

Bodri, L. and Cermak, V.: Borehole climatology: a new method how to reconstruct climate, Elsevier Science and Technology, Netherlands, 3rd. edn., 2007.

Carslaw, H. S. and Jaeger, J. C.: Conduction of heat in solids, Oxford Univ. Press. 2nd ed., New York, 3rd. edn., 1959.

Cermak, V. and Bodri, L.: Attribution of precipitation changes on ground–air temperature offset: Granger causality analysis, International Journal of Earth Sciences, 107, 145–152, https://doi.org/10.1007/s00531-016-1351-y, https://doi.org/10.1007/s00531-016-1351-y, 2018.

Cermak, V., Bodri, L., Kresl, M., Dedecek, P., and Safanda, J.: Eleven years of ground–air temperature tracking over different land cover types, International Journal of Climatology, 37, 1084–1099, https://doi.org/10.1002/joc.4764, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.4764, 2017.

Chapman, D. S., Bartlett, M. G., and Harris, R. N.: Comment on 'Ground vs. surface air temperature trends: implications for borehole surface temperature reconstructions' by M. E. Mann and G. Schmidt, Geophys. Res. Lett., 31, L07 205, doi:10.1029/2003GL019 054, 2004.

Fernández-Donado, L., González-Rouco, J. F., Raible, C. C., Ammann, C. M., Barriopedro, D., García-Bustamante, E., Jungclaus, J. H., Lorenz, S. J., Luterbacher, J., Phipps, S. J., Servonnat, J., Swingedouw, D., Tett, S. F. B., Wagner, S., Yiou, P., and Zorita, E.: Large-scale temperature response to external forcing in simulations and reconstructions of the last millennium, Climate of the Past, 9, 393–421, https://doi.org/10.5194/cp-9-393-2013, http://www.clim-past.net/9/393/2013/, 2013.

Frank, D., Esper, J., Fraedrich, K., Ziehmann, C., and Sielmann, F.: Adjustment for proxy number and coherence in a large-scale temperature reconstruction, Geophys. Res. Lett., 34, L16 709, Doi: 10.1029/2007GL030 571, 2007.

Gao, C. C., Robock, A., and Ammann, C.: Volcanic forcing of climate over the past 1500 years: an improved ice core-based index for climate models, J. Geophys. Res., 113, D23 111, doi: 10.1029/2008JD010 239, 2008.

García-García, A., Cuesta-Valero, F. J., Beltrami, H., and Smerdon, J. E.: Simulation of air and ground temperatures in PMIP3/CMIP5 last millennium simulations: implications for climate reconstructions from borehole temperature profiles, Environmental Research Letters, 11, 044 022, http://stacks.iop.org/1748-9326/11/i=4/a=044022, 2016.

García-García, A., Cuesta-Valero, F. J., Beltrami, H., and Smerdon, J. E.: Characterization of Air and Ground Temperature Relationships within the CMIP5 Historical and Future Climate Simulations, Journal of Geophysical Research: Atmospheres, 124, 3903–3929, https://doi.org/10.1029/2018JD030117, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JD030117, 2019.

González-Rouco, J. F., von Storch, H., and Zorita, E.: Deep soil temperature as proxy for surface air-temperature in a coupled model simulation of the last thousand years, Geophys. Res. Lett., 30, 2116–2119, 2003.

González-Rouco, J. F., Beltrami, H., Zorita, E., and von Storch, H.: Simulation and inversion of borehole temperature profiles in surrogate climates: Spatial distribution and surface coupling, Geophys. Res. Lett., 33, L01 703,doi:10.1029/2005GL024 693, 2006.

González-Rouco, J. F., Beltrami, H., Zorita, E., and Stevens, B.: Borehole climatology: a discussion based on contributions from climate modelling, Clim. Past, 5, 99–127, 2009.

Harris, R. N. and Chapman, D. S.: Mid-Latitude (30-60 N) climatic warming inferred by combining borehole temperatures with surface air temperatures, Geophys. Res. Lett., 28, 747–750, 2001.

Hartmann, D., Klein Tank, A., Rusticucci, M., Alexander, L., Brönnimann, S., Charabi, Y., Dentener, F., Dlugokencky, E., Easterling, D., Kaplan, A., Soden, B., Thorne, P., Wild, M., and Zhai, P.: Observations: Atmosphere and Surface, book section 2, p. 159–254, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/CBO9781107415324.008, www.climatechange2013.org, 2013.

Houghton, J., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., Dai, X., Maskell, K., and Johnson, C. A.: Climate Change, 2001: The scientific basis. Contribution of Working Group I to Third Assessment Report of the Intergovernmental Panel on Climate Change., Cambridge University Press, 2001.

Huang, S. and Pollack, H. N.: Global Borehole Temperature Database for Climate Reconstruction, IGBP PAGES/World Data Center-A for Paleoclimatology Data Contribution Series 1998-044. NOAA/NGDC Paleoclimatology Program, Boulder CO, USA. GR1, 1998.

Huang, S., Pollack, H. N., and Shen, P. Y.: Temperature trends over the past five centuries reconstructed from borehole temperatures, Nature, 403, 756–758, 2000.

Hunke, E. C., Lipscomb, W. H., Turner, A. K., Jeffery, N., and Elliott, S.: CICE: the Los Alamos Sea Ice Model Documentation and Software User's Manual Version 5.1, 2015.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model: A Framework for Collaborative Research, Bulletin of the American Meteorological Society, 94, 1339–1360, https://doi.org/10.1175/BAMS-D-12-00121.1, http://dx.doi.org/10.1175/BAMS-D-12-00121.1, 2013.

Hurtt, G. C., Chini, L. P., Frolking, S., Betts, R., Feedema, J., Fischer, G., Goldewijk, K. K., Hibbard, K., Janetos, A., Jones, C., Kindermann, G., Kinoshita, T., Riahi, K., andS. Smith, E. S., Stehfest, E., Thomson, A., Thorton, P., van Vuuren, D., and Wang, Y.: Harmonization of Global Land Use Scenarios for the Period 1500-2100 for IPCC-AR5, Integrated Land Ecosystem-Atmosphere Processes Study (iLEAPS) Newsletter, 7, 6–8, 2009.

Jansen, E., Overpeck, J., Briffa, K. R., Duplessy, J. C., Masson-Delmontte, V., Olago, D., Otto-Bliesner, B., Peltier, W. R., Rahmstorf, S., Ramesh, R., Raynaud, D., Rind, D., Solomina, O., Villalba, R., and Zhang, D.: Paleoclimate. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [S. Solomon and D. Qin and M. Manning and Z. Chen and M. Marquis and K. B. Averyt and M. Tignor and H. L. Miller (eds.)], Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Jaume-Santero, F., Pickler, C., Beltrami, H., and Mareschal, J.-C.: North American regional climate reconstruction from ground surface temperature histories, Climate of the Past, 12, 2181–2194, https://doi.org/10.5194/cp-12-2181-2016, https://www.clim-past.net/12/2181/2016/, 2016.

Jones, P. D., Briffa, K. R., Osborn, T. J., Lough, J. M., Van Ommen, T. D., Vinther, B. M., Luterbacher, J., Wahl, E. R., Zwiers, F. W., Mann, M. E., Schmidt, G. A., Ammann, C. M., Buckley, B. M., Cobb, K. M., Esper, J., Goose, H., Graham, N., Jansen, E., Kiefer, T., Kull, C., Küttel, M., Mosley-Thompson, E., Overpeck, J. T., Riedwyl, N., Schulz, M., Tudhope, A. W., Villalba, R., Wanner, H., Wilff, E., and Xoplake, E.: High-resolution palaeoclimatology of the last millennium: A review of current status and future prospects, Holocene, 19, 3–49, https://doi.org/10.1177/0959683608098952, https://doi.org/10.1177/0959683608098952, 2009.

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., Zeng, X., Yang, Z.-L., Levis, S., Sakaguchi, K., Bonan, G. B., and Slater, A. G.: Parameterization improvements and functional and structural advances in Version 4 of the Community Land Model, Journal of Advances in Modeling Earth Systems, 3, n/a–n/a, https://doi.org/10.1029/2011MS00045, http://dx.doi.org/10.1029/2011MS00045, m03001, 2011.

Legutke, S. and Voss, R.: The Hamburg Atmosphere-Ocean Coupled Circulation Model ECHO-G, Tech. Rep. 18, DKRZ, Hamburg, 1999.

MacDougall, A. H. and Beltrami, H.: Impact of deforestation on subsurface temperature profiles: implications for the borehole paleoclimate record, Environmental Research Letters, 12, 074 014, https://doi.org/10.1088/1748-9326/aa7394, 2017.

MacDougall, A. H., González-Rouco, J. F., Stevens, M. B., and Beltrami, H.: Quantification of subsurface heat storage in a GCM simulation, Geophys. Res. Lett., 35, L13 702, doi:10.1029/2008GL034 639, 2008.

MacFarling Meure, C., Etheridge, D., Trudinger, C., Steele, P., Langenfelds, R., Van Ommen, T., Smith, A., and Elkins, J.: Law Dome CO2, CH4 and N2O ice core records extended to 2000 years BP, Geophys. Res. Lett., 33, L14 810, 2006.

Mann, M. E. and Schmidt, G. A.: Ground vs. Surface air temperature trends: implications for borehole surface temperature reconstructions, Geophys. Res. Let., 1, 1–1, 2003.

Mareschal, J.-C. and Beltrami, H.: Evidence for recent warming from perturbed geothermal gradients: examples from eastern Canada, Climate Dynamics, 6, 135–143, https://doi.org/10.1007/BF00193525, http://dx.doi.org/10.1007/BF00193525, 1992.

Masson-Delmotte, V., Schulz, M., Abe-Ouchi, A., Beer, J., Ganopolski, A., González Rouco, J., Jansen, E., Lambeck, K., Luterbacher, J., Naish, T., Osborn, T., Otto-Bliesner, B., Quinn, T., Ramesh, R., Rojas, M., Shao, X., and Timmermann, A.: Information from Paleoclimate Archives, book section 5, p. 383–464, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/CBO9781107415324.013, www.climatechange2013.org, 2013.

Melo-Aguilar, C., González-Rouco, J. F., García-Bustamante, E., Navarro-Montesinos, J., and Steinert, N.: Influence of radiative forcing factors on ground–air temperature coupling during the last millennium: implications for borehole climatology, Climate of the Past, 14, 1583–1606, https://doi.org/10.5194/cp-14-1583-2018, https://www.clim-past.net/14/1583/2018/, 2018.

Neale, R. B., Gettelman, A., Park, S., Conley, A. J., Kinnison, D., Marsh, D., Smith, A. K., Vitt, F., Morrison, H., Cameron-smith, P., Collins, W. D., Iacono, M. J., Easter, R. C., Liu, X., Taylor, M. A., chieh Chen, C., Lauritzen, P. H., Williamson, D. L., Garcia, R., francois Lamarque, J., Mills, M., Tilmes, S., Ghan, S. J., Rasch, P. J., and Meteorology, M.: Description of the NCAR Community Atmosphere

Model (CAM 5.0), Tech. Note NCAR/TN-486+STR, Natl. Cent. for Atmos, pp. 1–289, http://www.cesm.ucar.edu/models/cesm1.0/cam/docs/description/cam5_desc.pdf, 2012.

Oleson, K. W., Lawrence, D. M., B, G., Flanner, M. G., Kluzek, E., J, P., Levis, S., Swenson, S. C., Thornton, E., Feddema, J., Heald, C. L., francois Lamarque, J., yue Niu, G., Qian, T., Running, S., Sakaguchi, K., Yang, L., Zeng, X., Zeng, X., and Decker, M.: Technical Description of version 4.0 of the Community Land Model (CLM), 2010.

Otto-Bliesner, B. L., Brady, E. C., Fasullo, J., Jahn, A., Landrum, L., Stevenson, S., Rosenbloom, N., Mai, A., and Strand, G.: Climate Variability and Change since 850 CE: An Ensemble Approach with the Community Earth System Model, Bulletin of the American Meteorological Society, 97, 735–754, https://doi.org/10.1175/BAMS-D-14-00233.1, http://dx.doi.org/10.1175/BAMS-D-14-00233.1, 2016.

Pollack, H. N. and Huang, S.: Climate reconstruction from subsurface temperatures, Annu. Rev. Earth. Planet. Sci., 28, 339–365, 2000.

Pollack, H. N. and Smerdon, J. E.: Borehole climate reconstructions: spatial structure and hemispheric averages, J. Geophys. Res., 109, D11 106, doi: 10.1029/2003JD004 163, 2004.

Pongratz, J., Reick, C., Raddatz, T., and Claussen, M.: A reconstruction of global agricultural areas and land cover for the last millennium, Global Biogeochemical Cycles, 22, n/a–n/a, https://doi.org/10.1029/2007GB003153, http://dx.doi.org/10.1029/2007GB003153, gB3018, 2008.

Rutherford, S. and Mann, M. E.: Correction to 'Optimal surface temperature reconstructions using terrestrial borehole data'., J. Geophys. Res., 109, D11 107, doi: 10.1029/2003JD004 290, 2004.

Santer, B. D., Thorne, P. W., Haimberger, L., Taylor, K. E., Wigley, T. M. L., Lanzante, J. R., Solomon, S., Free, M., Gleckler, P. J., Jones, P. D., Karl, T. R., Klein, S. A., Mears, C., Nychka, D., Schmidt, G. A., Sherwood, S. C., and Wentz, F. J.: Consistency of modelled and observed temperature trends in the tropical troposphere, International Journal of Climatology, 28, 1703–1722, https://doi.org/10.1002/joc.1756, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.1756, 2008.

Schmidt, G. A. and Mann, M. E.: Reply to Comment on 'Ground vs. surface air temperature trends: implications for borehole surface temperature reconstructions' by M. E. Mann and G. Schmidt, Geophys. Res. Lett., 31, L07 206, doi:10.1029/2003GL019 144, 2004.

Schmidt, G. A., Jungclaus, J. H., Ammann, C. M., Bard, E., Braconnot, P., Crowley, T. J., Delaygue, G., Joos, F., Krivova, N. A., Muscheler, R., Otto-Bliesner, B. L., Pongratz, J., Shindell, D. T., Solanki, S. K., Steinhilber, F., and Vieira, L. E. A.: Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.0), Geoscientific Model Development, 4, 33–45, https://doi.org/10.5194/gmd-4-33-2011, https://www.geosci-model-dev.net/4/33/2011/, 2011.

Schmidt, G. A., Jungclaus, J. H., Ammann, C. M., Bard, E., Braconnot, P., Crowley, T. J., Delaygue, G., Joos, F., Krivova, N. A., Muscheler, R., Otto-Bliesner, B. L., Pongratz, J., Shindell, D. T., Solanki, S. K., Steinhilber, F., and Vieira, L. E. A.: Climate forcing reconstructions for use in PMIP simulations of the Last Millennium (v1.1), Geoscientific Model Development, 5, 185–191, https://doi.org/10.5194/gmd-5-185-2012, https://www.geosci-model-dev.net/5/185/2012/, 2012.

Smerdon, J. E.: Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments., WIREs Clim. Change, 3, 63–77, doi: 10.1002/wcc.149, 2012.

Smerdon, J. E., Pollack, H. N., Cermak, V., Enz, J. W., Kresl, M., Safanda, J., and Wehmiller, J. F.: Air-ground temperature coupling and subsurface propagation of annual temperature signals, J. Geophys. Res., 109, D21 107, doi: 10.1029/2004JD0050 506, 2004.

Smith, R., Jones, P., Briegleb, B., Bryan, F., Danabasoglu, G., Dennis, J., Dukowicz, J., Eden, C., Fox-Kemper, B., Gent, P., Hecht, M., Jayne, S., M. Jochum, W. Large, K. L., Maltrud, M., Norton, N., Peacock, S., Vertenstein, M., and Yeager, S.: The Parallel Ocean Program (POP) Reference Manual, Tech. Rep. LAUR-10-01853, Los Alamos National Laboratory, http://tinyurl.com/SJBBD10/doc/sci/POPRefManual.pdf, 2010.

Stevens, M. B., Smerdon, J. E., González-Rouco, J. F., Stieglitz, M., and Beltrami, H.: Effects of bottom boundary condition placement on subsurface heat storage: Implications for climate model simulations, Geophysical Research Letters, 34, L02 702, doi:10.1029/2006GL028 546, 2007.

Storch, H. v. and Zwiers, F. W.: Statistical Analysis in Climate Research, Cambridge University Press, https://doi.org/10.1017/CBO9780511612336, 1999.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological Society, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

Vieira, A., Solanki, S. K., Krivova, N. A., and Usoskin, I.: Evolution of the solar irradiance during the Holocene, Astronomy & Astrophysics, 531, A6, https://doi.org/10.1051/0004-6361/201015843, 2011.