Answer to referee 1

The referee's comments are shown in black and our answers in blue:

The article provides the first systematic model-data comparison based on borehole temperature-depth profiles in Antarctica. They elaborate two techniques (depth and time domains) to compare these profiles and their reconstructions from four sites with climate model output. They conclude by outlining some useful metrics for future model data comparison and highlight the importance of internal variability on the observed tendencies.

We would like to thank the reviewer for the careful evaluation of our work and the very useful comments that will be addressed in the revised version.

Specific comments

1. L51: "Since the variable measured in the borehole is the temperature itself,. . ." In most cases, resistivity is measured, which is easily converted to temperature. Is this true of your measurement techniques?

   Yes, we agree with the referee that resistivity is generally measured in the field, using thermistor, and then it is converted to temperature by Steinhart-Hart equation. But this is true of every modern sensor (an electronic signal is converted to a meaningful variable). We propose to replace the following sentence:

   "Since the variable measured in the borehole is the temperature itself,. . ."

   by :

   "The most significant advantage of borehole paleothermometry is that temperature is directly measured with a thermistor calibrated in the laboratory. Thus, the calibration is independent of the climate at the measurement site."

2. L53: "the surface temperature history makes the reconstruction mathematically undetermined." The equation of temperature at depth usually results in a system of linear equations which is mathematically under and overdetermined. Varying mathematical inversion techniques are then utilized to reconstruct the ground surface temperature history. Please clarify. This should also be clarified on L.267.

   Yes, we propose to replace the following sentence:

   "the surface temperature history makes the reconstruction mathematically undetermined."

   by:

   "Nevertheless, the characteristics of heat conduction that blurs the surface temperature history make the reconstruction mathematically undetermined: several temperature histories can result in the same borehole temperature profile, because diffusion will smooth out high frequency temperature variations. Consequently, the temperature history cannot be determined unequivocally."

Line 267 reads "However, as stated above, borehole temperature reconstructions are "underdetermined", which means that there are many possible temperature histories that can fit the data." The word "underdetermined" will be between quote with an implicit reference to the explanation given above in line 53.

3. L.92: "Previous studies using forward models driven by climate model outputs were focused on ground temperature and not to borehole. . ." Please provide a couple examples (references) here.

   According to referee's suggestion, we will add some references, Beltrami et al., 2005; García-García et al., 2016; González-Rouco et al., 2003, 2006.

4. In Section 2.1, the borehole measurements and reconstructions are briefly explained. Since they come from four different publications, how can differences in inversion/reconstruction techniques affect the results presented in Figure 1?

   The reviewer raises an important point but comparing the different inversion/reconstruction techniques is out of the scope of our study. The temperature reconstructions are sensitive to the technique used. Notably, because the problem is underdetermined, several temperature histories are equally probable, and the final result will depend on some parameters used to calculate the inversion. We can illustrate this for instance by driving the borehole temperature model selected in this study by the published reconstructed temperature history and compare it to the observed borehole temperature. Difference have been found that are likely attributed to the different methodology and hypothesis but they are relatively small, suggesting that they do not have a major impact on our conclusions. Nevertheless, a more substantial analyses would be required to formally prove this.
   We will add in the revised version a cautionary note mentioning the potential influence of the application of those different techniques.

5. L. 136: "CESM1-CAM5 and MPI-ESM-P are not continuous in 1850." What is meant by this? Please clarify.

   This will be clarified:

   "CESM1-CAM5 and MPI-ESM-P simulations do not cover the entire millennium. Historical simulations covering 1851–2005 C.E. were launched independently of simulations covering 850–1850 C.E. (referred to as the past1000 experiment in CMIP/PMIP nomenclature). In order to obtain results over the full millennium, we adopt the approach from Klein and Goosse (2018) and merge the first ensemble members (r1i1p1) of the past1000 experiment with the corresponding ensemble members of the historical experiment. Although not continuous, there is no large discrepancy in 1850 C.E. between the two merged simulations (e.g., Klein and Goosse, 2018)."

6. In Section 2.2, for all models, excluding the CESM ensemble, which realization (r1i1p1) is used? Is it the only realization available? If not, why was this one selected? Please clarify.

   Yes, until now only one simulation for CCSM4, GISSE2-R, IPSL-CM5A-LR, MPI-ESM-P and BCC-CSM1-1 is publicly available. This will be clarified in the revised version of the manuscript:

   "For CESM1, an ensemble of simulations is available, providing an estimate of the internal variability as simulated by this model, but for CCSM4, GISSE2-R, IPSL-CM5A-LR, MPI-ESM-P and BCC-CSM1-1, there is only one simulation available."

7. L. 160: ": : :for Mill Island, the heat flux is set to zero: : :" How realistic is this? Furthermore, this is a different technique than the other sites. Could this influence the results? If the heat flux is set to zero, how is the steady-state temperature calculated?

   In the case of Mill Island, the hole is shallow (120 m), but the ice sheet is very deep at the site. At sites with such a deep ice sheet, and with a high accumulation rate, the conditions at the base are not impacting more than roughly the bottom 1000 m, so it is perfectly reasonable to model the top of the ice sheet only, with a zero heat flux at the bottom. The validity of this assumptions is discussed in detail in the original paper. Here is a quote: "The optimal surface temperature history was found to be essentially independent of the location of this bottom boundary condition for depths in excess of 180 m below the surface" (Roberts et al., 2013). The steady state temperature is calculated in the same way as the other models, and the only difference is the bottom boundary condition. This will be specified in the revised version.

8. L. 161: "For WAIS, a vertical step of 1 m for the upper 500 m and up to 25 m for the deepest part, and for other sites where the depth of borehole is close or less than 500 m, the step is set to 1 m for overall depth." Why are various techniques used again? What is the benefit of this?

   WAIS is the only very deep borehole, and we use a coarser model resolution for the deepest part to save some computer time as in Orsi et al 2012. This is not required for the shallower cores for which the computation time is lower and it is the reason why we keep a fine resolution for all the depth of the core. This will be specified in the revised version.

9. L. 183: "At WAIS-Divide, the spread of the sensitivity tests is lower than the spread if the different scenarios." What is meant by scenarios? Is it the different models being analyzed?

   The different scenarios mean the different simulated borehole profiles driven by different climate model results. This will be clarified in the revised version of the manuscript:

   "At WAIS-Divide, the spread of the sensitivity tests is lower than the spread in the simulated borehole profiles driven by different climate model results (solid lines in color in Figure 2 (a) and (b))."

10. L.196: ": : :but the deviation in the top 100 m show that there is climate information stored in the upper part of the profile, and that this profile cannot be fully determined by boundary conditions." Climate is not the sole reason why the top 100 m would show deviation. How can you be sure it is climatic information?

    We totally agree that the surface temperature change is not the sole reason why the top 100 m would show deviation. For instance, we can find that the initial and basal temperature have some impacts on the shape of simulated borehole temperature in the top 100 m shown in the Figure 2 (e). Meanwhile, we expect that some climate information is stored in the top 100 m from the comparison between the simulated borehole temperature profiles (solid lines in the Figure 2) driven by different GCMs with the stationary temperature profile (thick dash-dot line). This paragraph will be rephrased in the revised version of the manuscript:

    "At Styx, the boundary conditions are adjusted to reproduce the slope of the temperature profile in the deeper part (100-200 m). Compared with stationary temperature profile, the simulated borehole profiles driven by GCMs (solid lines in the Figure 4 (e)) show a deviation in the top 100 m, which suggests that there is climate information stored in the upper part of the profile. Meanwhile, at the depth shallower than 50m, the effect of boundary conditions is weaker than the differences in the temperature histories from the different model, which means the borehole temperature data can be used to discriminate between temperature histories provided by the different models."

11. At the start of the paragraph at L.198, it is stated that internal climate variability and the different characteristics of the climate models are the main sources of differences. The results from the CESM ensemble have not been discussed in this section. To strengthen this point, I recommend adding in a discussion of it.

    As suggested by the reviewer, a discussion of the internal variability in the CESM ensemble will be added at the start of the paraph at L.198 in the revised version:

    "The internal variability also has significant impact on the shape of the simulated borehole profiles. At these four sites, the range of simulations driven by CESM ensemble is much larger than range of the different sensitivity tests in the top of 50 m (shown as the shaded area in Figure 4 b, d, f, h), which conforms that the dominant source of uncertainty in a model–data comparison, at least in the top 50 m, is from the internal variability."

    Furthermore, the statement that internal climate variability and different characteristics of the climate models being the main source of differences does not hold true for Mill Island. In Figure 1, only different depths of the zero heat flux are considered. More tests must be added to conclude the importance of the influence of internal variability and different model characteristics to the differences at this site.

    Yes, in Figure 2(e), only the upper 50 m shows that there is a noticeable spread between the colored lines, illustrating that different climate model scenarios result in different temperature profiles, and that this difference is larger than the spread between the dashed lines (the sensitivity to model parameters). We will clarify it and include more sensitivity tests for Mill Island, as requested in the revised version.

12. In Figure 3, why are different smoothing techniques used? Can it influence the results?

    The reason why we use different smoothing is to facilitate a comparison between the reconstruction and climate model results. The reconstruction provides a smoothed history of the past surface temperature changes, but the smoothing itself depends on the time and the characteristics of the site. We have tried to mimic this as much as possible by using variable smoothing in the plot. As the reconstructions at WAIS and Styx preserve mainly the centennial and multi-centennial variabilities, we applied longer smoothing (50-year) to the climate model result at WAIS and Styx. Similarly, at Larissa and Mill Island, the reconstructions show the multi-decadal and decadal variabilities, so we choose 10- year smoothing at Larissa and 3-year smoothing at Mill Island.
    Using these different smoothing techniques is thus justified and not influencing significantly our conclusions, because our goal here is to perform a visual model-data comparison in the time domain in order to see if the reconstruction is within the range provided by the ensemble. Since the reconstructions have much wider ranges than those ones from the climate model results, the basic compatibility between model and model will not be changed.

13. The reconstructions from the climate models presented in Figure 3 are calculated using what technique? Their errors bounds are also not presented. How does this influence the results? Does the reconstruction from the climate model always lie within the error bounds? Please clarify in the manuscript.

    The temperatures displayed in Figure 3 come directly from the surface temperature calculated by the climate model, based on its own dynamics and the forcing applied as discussed in section 2.2. Single time series are available for each model experiment without error bounds but providing an ensemble of experiments gives a range of current state-of-the-art models. This range provides a kind of uncertainty associated with model results but relating this to a precise estimate of the error is unfortunately a complex issues as models are for instance not independent of each other, sharing similar parametrizations, and may have common biases, due in particular to the relatively coarse resolution of climate models (Abramowitz et al., 2019;Knutti et al., 2017;Sanderson et al., 2015).

14. L.215: "In order to remove the bias on the mean state for each climate model, anomalies are shown using the total period covered by each reconstruction as reference." Which figure is being referred to? The paragraph starts discussing Figure 3 but these are not anomalies.

    Here, we show an example to explain the methodology. In Figure R1, the original climate model result is shown as the red curve. Its mean over the period 850-2000 C.E. is different from the reconstruction. To remove this bias, we applied a very simple bias correction to climate model results, ensuring that after the adjustment the climate models have the same mean over the reference period as the reconstruction (Yellow curve in the Figure R1). This will be clarified in the revised version of the manuscript as follows:

"In order to ensure that the climate model results have the same mean over the reference period as the reconstruction, we applied a very simple, constant correction to remove the mean bias of the climate model results as shown on the Figure 3."

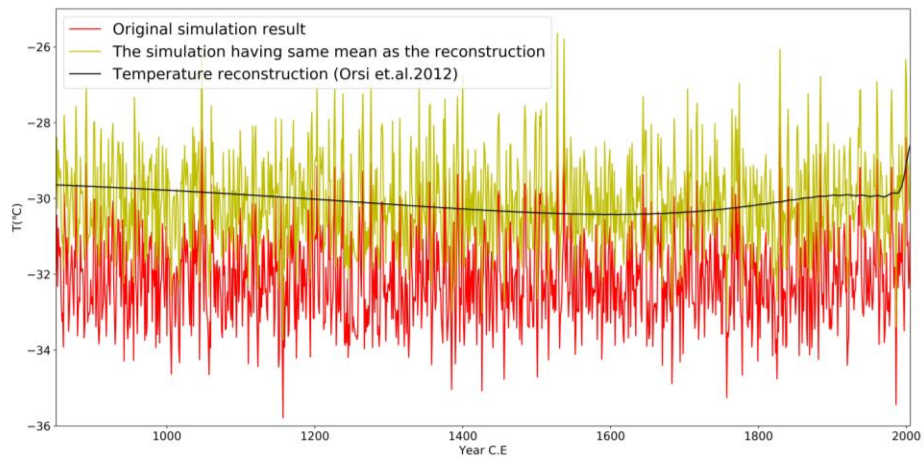

Figure R1. Comparison between reconstructed surface temperature series at WAIS and the climate model outputs at the grid cell-point closest to WAIS.

15. Since the temperature variability increases as you go back in time, there is less confidence with respect to the timing of events. Timing of events varies within climate models. Could this further explain any discrepancies of the timing of events? Please discuss.

The timing of the events differs indeed between the simulations if those events are related to internal variability and not caused by a specific forcing. This influence of internal variability can be estimated from the difference between the CESM members as discussed in Section 4. In addition, from the Figure R2, the range of CESM members does not increase back in time. This suggests that the temperature variability between the different members of CESM, and thus the associated uncertainty, does not change a lot over the time.

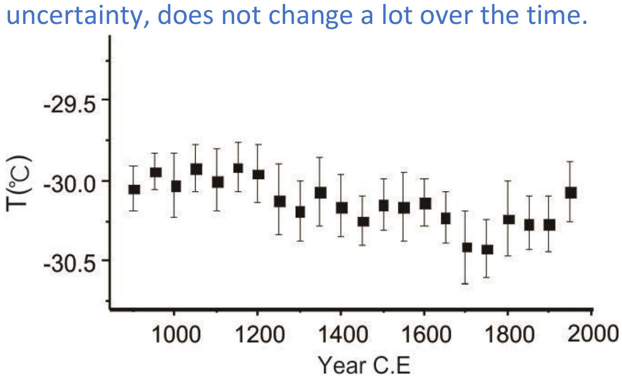

Figure R2. Temperature variability of the CESM ensemble at the grid cell of WAIS. The black square represents the mean of the CESM ensemble in the corresponding time. Their error bound present the 1 standard deviation (1σ) ranges of the CESM ensemble.

16. What causes the decrease in temperature at 1980 and 2000 in Styx and Larissa (Figure 3)? Is it climatic in origin or an artifact of the reconstruction technique?

    The four borehole reconstructions in the manuscript are from the original papers. In the papers related to the Styx (Yang et al., 2018) and Larissa (Zagorodnov et al., 2012), the authors have shown that the reconstructions from borehole are consistent with the weather stations, and ice core isotope-derived records. Consequently, the decrease in temperature in 1980 C.E. and 2000 C.E. at Styx and Larissa likely reflect climatic signals.

17. L.390: "Fig. 8 shows the spatial correlation in the Antarctica Peninsula (AP)." Do you mean the spatial correlation of the gridcells? Please clarify

    As suggested, we rewrite this sentence as follows:

    "Figure 8 shows the spatial correlation between the temperature from 1825 to 1925 C.E. at Larissa and other grids cells for each climate model."

18. L390-393: "Despite the correlation coefficient decreasing as the grid getting far away from the Larissa, the values, at least around Larissa for each model, are higher than 0.6, showing that this metric is representative of the whole peninsula region, and not extremely site-specific." A correlation coefficient of 0.6 means that it only explains 36% of the variance. How can you conclude that it is representative of the entire peninsula?

    Yes, we totally agree with the referee that a correlation coefficient of 0.6 is not that high and our view was probably a bit too optimistic. We have changed this sentence following the suggestion:

    "Despite the correlation coefficient decreasing with the distance from the Larissa, the values, at least around Larissa for each model, are higher than 0.6, showing that this metric is representative of part of the AP region, and not extremely site-specific."

19. Why are the CESM ensemble members not presented in Figure 8? How is this metric influenced by internal variability?

    Figure 8 shows the correlation between the temperature from 1825 to 1925 C.E at Larissa and other grids cells in the Whole Antarctica for each climate model. Since there are no significant differences between each member in CESM ensemble, we just show one member (CESM1) as an example in the manuscript.
    This will be mentioned explicitly in the revised version of the manuscript and we will add the figure shown below (Figure S1) in the supplement:

    "As there are no significant differences between each member in CESM emsemble (see in the Figure S1), only one member CESM1 and other GCMs are present in the Figure 8."

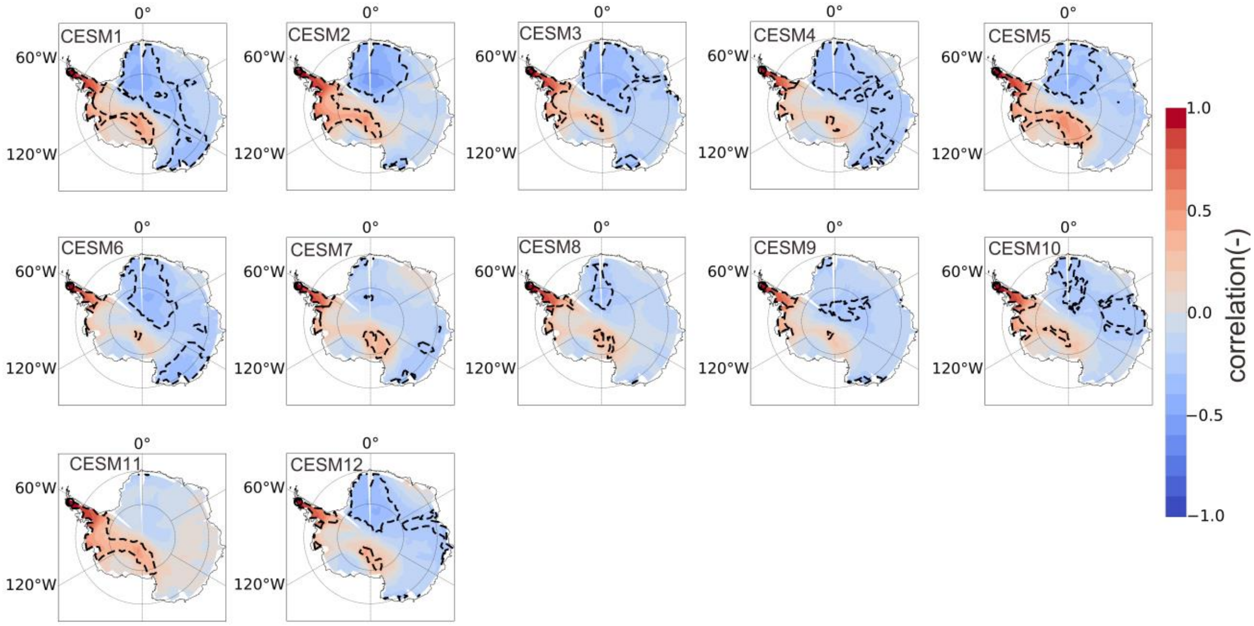

Figure S1. The correlation map (blue-red shading area) showing the relationship between the temperature from 1825 C.E. to 1925 C.E at Larissa and other grid cells in Antarctica for each CESM member. The black dotted contour lines show a significant correlation at the 99 % significant level.

20. L.426-428: "A model that responds clearly to the Ozone forcing, and has a strong SAM signature should exhibit this dipole pattern, and it is interesting that some models do not show it, indicating that the Ozone forcing is not dominating over internal variability." The CESM ensemble members are not seen in Figure 10 and 11 nor discussed. How can this be concluded?

    We agree with the reviewer that we jump a bit too quickly on the conclusions and a dedicated analysis should be performed to prove this. In particular, the correlation pattern can be also strongly influenced by the spatial response of each model to the ozone forcing itself. To avoid a long discussion on a point not central to our analysis, we have preferred to remove the part of the sentence relating to the respective role of ozone forcing and internal variability.

21. Borehole temperature profiles and their ground surface temperature histories are compared with those from climate models. They ability of the climate models to reconstruct the ground surface temperature was evaluated and three distinct metrics were created. From all of this, how do you think climate models could improve? From your analyses, what are their areas of weaknesses? It would be beneficial to add a section outlining this to the conclusions.

    We agree with the reviewer that this is a very interesting issue. Our goal was to provide a test to estimate the performance of climate models. This is of course the basis for model improvements but the link between a bias in one diagnostic and a model improvement is not straightforward and can be very different for different models. Making specific suggestions would require many additional diagnostics, comparisons and tests. Without that, we fear that any additional material we could add on this subject would be too general or speculative to be really informative for the reader. This is the reason we do not plan to develop this in the revised version.

Technical Points: There are many grammatical errors throughout the article impeding the reader's comprehension. They are not all outlined below but should be addressed in the revised manuscript.

Thank you for noticing all the following errors that will be corrected. We will check all the details in the grammar and improve it in the revised manuscript.

1. L36: Please define acronym AP

   Thanks, we will add the definition of AP (Antarctic Peninsula).

2. It would facilitate comprehension if the depths of each borehole were added to Table 1. This would help explain the various time periods for the reconstructions found in Figure 1.

   We will add a column for depth in the Table 1.

3. Please add the units of elevation to the map in Figure 1.

   We will add them.

4. In Figure 2, in the boxes below Figures 2 a,b and 2c,d, there is a typo in the word accumulation. The thermal diffusion used along with its units should be included. In the caption, " 2) sensitivity tests using the temperature history of once CESM member. . .", do you mean one CESM member? Also "The shade area represents the simulated subsurface temperature ensemble driven by CESM" should read "The shaded area. . ."

   As suggested, we will add the unit for the thermal diffusivity.
   We will modify "once CESM member" to "one CESM member" and also correct the phraseology of "shade" to shaded.

5. L.192: "At Mill island,. . ." Should read Mill Island to be consistent throughout the text.

   As suggested, we will correct it.

6. L.192: "..the ice thickness is much deeper. . ." Ice thickness cannot be deeper. Should read thicker.

   This will be modified.

7. In Figure 4, the y-axis of 4a,e and f are crowded. Either decrease the amount to y-ticks or increase the figure size. Some of the symbols, in particular the yellow triangle of CCSM4, are difficult to see. I would recommend increase the size of the markers for the climate models and the reconstruction. Also, the labels on 4c and d are cut off by the below figures. Please fix.

   We will fix these problems in the Figure 4, and it will be updated with a clearer figure in the revised version.

8. L.254: "Larissa shows a temperature minimum in 1940's. . ." should read . . .1940s.

This will be modified as suggested.

9. L.270 a period is missing at the end of the sentence.

   This will be modified as suggested.

10. In Figure 5, please use a different colour for the observations.

    We will modify the Figure 5 as suggested.

11. For consistency, use CCSM or CCSM4

    As suggested, we will check and replace CCSM by CCSM4.

12. For the techniques/metrics elaborated in Section 4, please be consistent with the use of grid, grid-point, and gridcell. Since you are comparing with data from the gridcell, I'd recommend the use of that word to facilitate the reader's comprehension.

    Thanks for your recommendation. We have replaced grid and grid-point by grid cell.

13. L.372: "For most of the models,. . ." It would be best to include a number or percentage of models to really illustrate your point.

    As suggested, we will add a percentage in the corresponding sentence : "75% models show WAIS displays a larger cooling from 1000 to 1600 C.E. than other locations in Antarctica (shown in blue) but with magnitude similar to other grid cells in West Antarctica.".

14. In Figures 6,9,12, please add the units to the colour bar as these are surface temperatures tendencies

    We will add the units to the colour bar in Figure 6, 9, 12.

15. In Figures 6,7,9, the circle illustrating the location of the observations is not clear. Maybe make it bolder or another colour.

    Done.

16. L.394: "Figure 9 shows the same temperature trend (1825-1925) for all models." Do you mean surface temperature since Figure 9 shows varying trends.

    Thanks for your remark. The previous sentence has been modified: no model is able to capture the observed temperature trend from 1825 C.E. to 1925 C.E..

17. L.395-396: “A majority of the CESM members(CESM1, 7, 8, and 9). . .” Do you mean minority? 4/12 is not a majority.

    Yes, we made a mistake here. Thank you for that. We propose to replace “A majority of the CESM members(CESM1, 7, 8, and 9)” by “Only four member of CESM (CESM1, 7, 8, and 9) show a cooling trend over AP, but the magnitudes of them are still less than the observed one”.

18. L.403: There is a typo in the word overestimation.

    Corrected.

19. In Figure 8, the dotted contour line is not clear to the reader. Also, indicate that the colour bar represents the correlation coefficient.

    Done.

20. L.427 please define acronym SAM.

    We will add the definition of SAM (Southern Annular Mode).

21. Figure 10, some of the numbers in the colour bar appear to be cut-off. The red dashed line is not visible to the reader. Please correct.

    We will fix these problems in the Figure 10, and it will be updated with a clearer figure.

22. Figure 11, the y-label of d is overlapping with c. Please clarify that it is the linear trends of surface temperature in the caption.

    We will fix these problems in the Figure 11.

23. Figure 12 is not referenced in the text

    Yes, we made a mistake here. We will remove it.

Reference:

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam., https://10, 91-105, 10.5194/esd-10-91-2019, 2019.

Klein, F. and Goosse, H.: Reconstructing East African rainfall and Indian Ocean sea surface temperatures over the last centuries using data assimilation, Climate Dynamics, 50, 3909–3929, https://doi.org/10.1007/s00382-017-3853-0, 2018.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophys. Res. Lett., https://10.1002/2016gl072012, 2017.

Orsi, A. J., Cornuelle, B. D. and Severinghaus, J. P.: Little Ice Age cold interval in West Antarctica: Evidence from borehole temperature at the West Antarctic Ice Sheet (WAIS) Divide, Geophys Res Lett, 39(9), 1–7, doi:10.1029/2012GL051260, 2012.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble, J. Clim., 28, 5171-5194, https://10.1175/jcli-d-14-00362.1, 2015.

Yang, J. W., Han, Y., Orsi, A. J., Kim, S. J., Han, H., Ryu, Y., Jang, Y., Moon, J., Choi, T., Hur, S. Do and Ahn, J.: Surface Temperature in Twentieth Century at the Styx Glacier, Northern Victoria Land, Antarctica, From Borehole Thermometry, Geophys Res Lett, 45(18), 9834–9842, doi:10.1029/2018GL078770, 2018.

Zagorodnov, V., Nagornov, O., Scambos, T. A., Muto, A., Mosley-Thompson, E., Pettit, E. C. and Tyuflin, S.: Borehole temperatures reveal details of 20 th century warming at Bruce Plateau, Antarctic Peninsula, Cryosphere, 6(3), 675–686, doi:10.5194/tc-6-675-2012, 2012.