Reply to Reviewer Comment C4

This manuscript surely is a valuable contribution to research in paleoclimate and fits well into the scope of CPD. It is generally well written, and the figures meet the common quality standard, though some could be improved.

In this study the authors seek to show the consistence of Portuguese paleoclimate data from four different sources, (1) the Europe-wide annual reconstruction of Luterbacher et al. (2004); (2) local repeated borehole temperature observation from one site in Portugal; (3) paleoclimate simulation and their signature in these boreholes; and (4) precipitation indices from documentary sources since the late Maunder minimum. Noting that the (1) is not consistent with the other results, they propose a new reconstruction based on a two-stage calibration procedure using information from these sources, and compare the obtained results with (4). This is an interesting approach which should be discussed in the community, though I think this manuscript should only published after major revisions.

Reply: We are thankful to the reviewer for his careful reading of the manuscript and for his very constructive and pertinent comments. We believe they have significantly improved our manuscript. Please see our point-by-point replies below. The corresponding changes in the manuscript are also highlighted in blue.

Particular Comments:
1) Borehole data. Though the authors refer the reader to earlier papers, I think that a bit more information would be useful. This concerns mainly the estimation of the geothermal gradient (or heat flow density). The authors correctly mention that their results can only be preliminary, given the small depth of the borehole. However, it would be interesting if the authors could include a discussion of this problem. For example, a plot of HFD(z) could be helpful, in order to see whether we are reaching more or less constant values in the estimation interval. In Fig 1 there is obviously a change in thermal conductivity at about 180 m depth (why not use the interval 140-180 m?). A HFD plot could possibly better justify the assumption that the background heat flow can be approximated linearly. Surely the estimated linear profile is not an steady state geotherm (as stated in the caption of Fig. 1), but may contain the signature of older events back to the last glacial cycle (e.g., Rath et al.2013). In addition, it would give a more direct view of the errors. It would be nice to see a graphic showing how this estimation error translates to the whole profile. Maybe a Monte Carlo study? Clearly, getting a bit more quantitative here would strengthen the study considerably.
Reply: We entirely agree with the reviewer that the heat flux density (HFD) measurements are a very useful tool to test the linearity assumption for the geothermal gradient estimates. Therefore, we have added further information concerning the borehole location and the HFD:

Lines 126-135: *"The borehole is located in a region where the typical vegetation is old cork trees. The vegetation type has not changed in the last hundred years and the topography is subdued, with small elevation variations of tens of meters in the nearest few kilometres. The rock type in the area is hercynian age granite. Its thermophysical properties were measured in four samples, collected in a quarry located in the same granitic body and 1.5 km eastwards of the borehole. Thermal conductivity values of 2.8 ± 0.2 W mK^{-1} and thermal diffusivity values of 1.3 ± 0.1 m^2 s^{-1} were measured on polished surfaces of rock samples. Heat production was calculated as 2 ± 1 W m^{-3} (Correia and Šafanda, 2001). The estimated heat flux density for the borehole is 60 mW m^{-2}, which was confirmed as an a posteriori value of 58 ± 13 mW m^{-2} using the Functional Space Inversion method of Shen and Beck (1992)."*

The following reference was added to the reference list:

Shen, P. Y. and Beck, A. E.: Paleoclimate Change and Heat-Flow Density Inferred from Temperature Data in the Superior Province of the Canadian Shield, Global and Planetary Change, 98, 143-165, 1992.

Despite the aforementioned considerations about the HFD, we would like to mention that the geothermal gradients are estimated with very low errors in the present study. In fact, this is now clearly stated in the manuscript:

Lines 225-228: "*The corresponding root-mean squared error (RMSE) of each estimated linear model is always <0.01ºC (R-square adjusted >99.9%), which means that the errors in the estimation of the geothermal gradients have only minor impacts on the subsequent temperature-depth anomalies.*"

The RMSEs of each linear fit are now also plotted in Fig.1b.

As we will consider each temperature-depth profile separately in Fig. 2a, we believe that these estimation errors are enough for quantifying the uncertainties in the profiles from a statistical viewpoint.

With respect to the change in thermal conductivity at depths around 180 m, we would like to mention that very fine material gradually deposits at the bottom of the borehole and may locally change thermal conductivity. Therefore, the change in temperature the reviewer refers to is due to a local effect and not to a change in the actual thermal conductivity of the borehole geological formation. In fact, the borehole was drilled in a very homogeneous granite batholith. Nevertheless, we agree with the reviewer that for a more adequate estimation of the HFD the depth interval of 140-180 m should be used instead of 140-180 m. We have recalculated these gradients and changed Fig. 1b and Fig. 2a accordingly. We would like to mention, however, that no significant changes were found in the results. We have also added the following explanation to the text:

Lines 221-224: "*Owing to the deposition of fine material at the bottom of the borehole, there is locally a change in thermal conductivity at about 180 m. As the borehole was drilled in a very homogeneous granite batholith, these changes are not due to changes in the geological formation. In the present study, depths >180m are not used for gradient estimations.*"

Moreover, the reviewer states that the estimated linear profile for the temperature log is not linear and may contain the signature of older events back to the last glacial cycle. We do not argue about that. However, we cannot detect this signal because of the short borehole depth. In fact, from a theoretical point of view, the wavelength of the glacial cycle is longer than what can be seen in a 190 m depth borehole.

Lastly, we agree that the term 'steady-state geotherm' is incorrect. We have removed 'steady-state' in Fig. 1 caption, as suggested by the reviewer.

(2) Inversions and forward models. The authors state that "...uncertainties inherent to these inversion methods (Hartmann and Rath, 2005) are avoided in the present study".
I do not agree. Most of the problems mentioned in the article cites are of a physical character. If you lack reliable estimates for the subsurface properties, they also will render the forward model (forced by some simulation or reconstruction) unreliable. I have already mentioned the problem of background heat flow density above. Any model or parametrization error will be present in both approaches and thus can not be simply avoided. The only additional problems here are the procedures related to the solution of the ill-posed inverse problem, e.g. the truncation or damping of the SVD-derived generalized inverse. However, comparable procedures are also implicitly or explicitly ingredient of many (non)linear regression codes, or, for example, in SSA

procedures of different flavors. This is of course not central to this study, but if mentioned at all, it should be discussed in a fair manner.
Reply: We agree with the reviewer that our statement is misleading. In order to avoid a detailed discussion of the inversion vs. forward models and of the ill-posed inverse problem, which we also think is out of the scope of the present study, we have rephrased the entire sentence. It now reads as:

Lines 140-142: "*However, the uncertainties inherent to these inversion models (Hartmann and Rath, 2005), mostly due to errors in the estimation of subsurface parameters, are also present in these forward models.*"

Regarding the HFD estimation problem referred by the reviewer here and in the "Particular Comments: 1) Borehole data", please see our previous reply.

(3) Regression approach. I found the part on the two phase calibration rather difficult to understand, and thus think that it should be expanded and improved. In particular,
It should be made clearer, which assumptions have to be made, and that - as I understand it - the result is a recombination of long period information from the simulations and shorter period information from the instrumental period and the Luterbacher reconstruction. This may be meaningful, but deserves more discussion. Personally I would not call this process a calibration ("adapting uncertain parameters in order to increase agreement of models with available observations"), but a method of reconstruction. From the description of the method at the end of section 2.3 it is not fully clear to me why it "can be used to correct discrepancies between long-term trends of reconstructed and simulated temperature series". One could conclude from this study that the Luterbacher reconstruction is not "valid" in this area in the light of the borehole temperatures (and rogation ceremonies) and simulations based on best current knowledge on climate physics. What observational data are relevant for the Luterbacher reconstruction in this area? Or, more general, why does the reconstruction not capture the trends? These would be the obvious questions following this study. A discussion of these problems could improve the manuscript considerably.
Reply: We agree that some parts of the methodology might not be clear enough and some changes were carried out in the text so as to improve clarity.

First of all, we would like to emphasise that we have not developed a new reconstruction. From our point of view, using the term "reconstruction" is indeed not adequate in our study. We have only performed an adjustment of the Lut2004 reconstruction. Technically, we have carried out a correction of the original Lut2004 through an amplification of its low-frequency variability. Thus, we now explicitly refer to a "post-reconstruction adjustment" throughout the manuscript.

More specifically, the SSA was used to isolate the low-frequency variability in the time series of the simulations and Lut2004 reconstruction. Following this methodology, we have identified a lack of agreement between the low frequency-variability in the Lut2004 reconstruction with the paleoclimatic simulations and their corresponding external forcings. Furthermore, the low-frequency variability in the simulations is in agreement with local borehole data. Therefore, the existing Lut2004 reconstruction was corrected to become more meaningful, i.e. its low-frequency variability was adjusted to the low-frequency variability in the simulations.

The different methodological steps are now more clearly explained in the abstract:

Lines 25-32: "*It is found that the reconstructed annual mean temperature series in Portugal is not consistent with the external forcings from the regional paleoclimate simulations. Furthermore, the low-frequency variability in the simulations is in agreement with local borehole temperature-depth profiles. Therefore, the existing reconstructed series is calibrated by adjusting its low-frequency variability to the simulations (first-stage adjustment). The annual*

reconstructed series is then calibrated in its location and scale parameters, using the instrumental series and a linear regression between them (second-stage adjustment)."

It is also now explicitly stated that we only have performed a post-reconstruction adjustment and not a reconstruction itself:

Lines 36-38: *"Thus, the series resulting from this post-reconstruction adjustment can be of foremost relevance to improve the current understanding of the driving mechanisms of climate variability in Portugal."*

Lines 86-88: *"The identification of possible inconsistencies with the above-referred data sources enables a post-reconstruction adjustment of this time series."*

With respect to the sentence in section 2.3 referred by the reviewer, it can be better contextualized in its full paragraph that now states the following, after some rephrasing:

Lines 191-200: "*Under the assumption that the aforementioned external forcings used in the paleoclimate simulations are mainly manifested by long-term temperature trends in western Iberia, as suggested by Gómez-Navarro et al. (2012), similar trends of reconstructed and simulated temperatures should be expected. As SSA enables isolating data-adaptive non-linear trends in the time series (Ghil and Vautard, 1991), it can be used to correct discrepancies between long-term trends of reconstructed and simulated temperature series. In the present study, this approach was used to adjust the low-frequency variability in the reconstructed series to the paleoclimate external forcings obtained from the simulations (adjustment of the Lut2004 reconstruction). Therefore, instead of developing a new reconstruction, an adjustment of the already existing reconstruction was carried out herein (post-reconstruction adjustment).*"

In section 3.1, some rephrasing was also undertaken to improve clarity:

Lines 256-268: "*The discussion above hints at a remarkable agreement between the low-frequency variability of near-surface temperature from two independent sources (borehole measurements and paleoclimate simulations). However, whereas the paleoclimate simulations agree well with the borehole temperature-depth profiles, the reconstructed temperature for Portugal (Lut2004) fails to capture the long-term trends. In fact, its linear trend is nearly zero over the whole period and there is no signature of cool/warm periods. This striking disagreement between simulations and Lut2004 was already reported by Gómez-Navarro et al. (2011). As such, the low-frequency variability of the Lut2004 reconstruction needs to be adequately adjusted for climate research purposes. Towards this aim, the ensemble mean temperature from the two simulations was low-pass filtered by a 2-order SSA. The filtered series (SSA-trend in Fig. 2b) highlights the signature of the external forcings on near-surface temperature and was then added to the Lut2004 reconstruction. The resulting calibrated series (CalT = Lut2004 + SSA-trend) is also shown in Fig. 2b.*"

Lastly, due to the complexity of methodologies and datasets used in the Lut2004 reconstruction, it is not possible for us to identify the exact causes for its limited reproduction of the low-frequency variability in Portugal. We may argue that the lack of data over Portugal may partially explain this shortcoming, but a more thorough research would be required for testing this hypothesis. Nonetheless, we would like to mention that a previous study has already found a severe loss of low-frequency variance in the temperature reconstructions by applying the principal component regression methodology followed by Lut2004 (von Storch et al., 2009).

Along these lines, we have added the following text to the final discussion:

Lines 350-357: "*The frequent temporal gaps in the pre-instrumental records and the substantial lack of natural proxies with clear climatic signals in Portugal (Alcoforado et al., 2012; Camuffo*

et al., 2010; Luterbacher et al., 2006) may partially explain the inadequate reproduction of the low-frequency variability in the Lut2004 reconstruction. However, a severe loss of low-frequency variance caused by the method used in Lut2004 was also found by von Storch et al. (2009). Nevertheless, a more detailed assessment of the causes for this shortcoming is out of the scope of the present study, as it does not develop a new reconstruction for comparison, but rather an adjustment of an existing reconstruction."

And a new reference was added to the reference list:

von Storch, H., Zorita, E., and Gonzalez-Rouco, F.: Assessment of three temperature reconstruction methods in the virtual reality of a climate simulation, Int. J. Earth Sci., 98, 67-82, doi:10.1007/s00531-008-0349-5, 2009.

Minor items:
General: too often "not shown" - better refer to other publications, or reformulate.
Reply: Two 'not shown' statements were removed from the manuscript. The text regarding the seasonal analysis (former version of the paper, pg 13, Lines 24-28) was removed, as this analysis is not relevant for the current study. The second 'not shown' statement (former version, pg 11, Line 22) was removed because Fig4c and its interpretation was also removed, as suggested in a comment below.

P4, L5: Possibly the reference is wrong - no boreholes mentioned. Christian, H. J., Blakeslee, R. J., Boccippio, D. J., Boeck, W. L., Buechler, D. E., Driscoll, K. T., Goodman, S. J., Hall, J. M., Koshak, W. J., Mach, D. M., and Stewart, M. F.: Global frequency and distribution of lightning as observed from space by the Optical Transient Detector, J. Geophys. Res.-Atmos., 108, 4005, doi:10.1029/2002JD002347, 2003.
Reply: We agree. We regret this mistake. The reference was removed.

P9, L7ff: Please explain shortly why a difference of 2 m may explain the results. Difference in what? Smallest observation depth? What happened to the annual temperature wave, which is dominant in boreholes down to 15 -20 m?
Reply: This sentence was misleading and was therefore removed from the manuscript. As a matter of fact, the sentence just meant to state that the temperature measurements in the M5 temperature log started deeper than in the other temperature logs. As a consequence, the first temperature measured in M5 is lower than in the other temperature logs. This also answers the other reviewer's comment: the annual temperature wave cannot be detected because the temperature measurement in all five temperature logs started at depths of 10 to 12 m.

P9 L15ff: Which ensemble? Never mentioned before. I think It should be mentioned in the section describing the simulations.
Reply: As suggested, the ensemble definition is now clearly stated in section 2.3.

Lines 167-169: "*An ensemble of two paleoclimate simulations (Sim1 and Sim2), only differing in their initial conditions, were used as a broad estimation of the effect of internal variability…*"

P10, line 1: Observations can support simulations, but how can simulations support observations? This is also relevant to the formulation on P12, L15.
Reply: We agree that this statement is unclear and it was removed from the text. The same applies to the terms 'cross-validation' and 'two-way validation', also following the comment below.

P10 L9: what is a "2-order SSA filtering"? Reformulate or explain.
Reply: We agree that the SSA filtering was not clear. Therefore, we have improved its description in section 2.3. It now reads:

Lines 175-190: "*In order to identify low-frequency variability and trends in the paleoclimate simulations, a data-adaptive filtering, based on a singular spectral analysis (SSA), is applied (Ghil and Vautard, 1991). SSA is based on the well-known principal component analysis, in which the multiple dimensionality is achieved by including time-lagged replicas of the original time series. The resulting principal components are thus linear combinations of different lags of this series, which is equivalent to a time filtering with filter-coefficients that are related to the eigenvectors of the lagged-covariance matrix. More formally, SSA corresponds to an eigenvalue decomposition of a lagged-covariance matrix, with a Toeplitz structure, obtained from the original time series of the paleoclimatic simulations. The rank, M, of this matrix is the average of (N/4–N/3), where N is the time series length (Plaut and Vautard, 1994). For the paleoclimatic simulations M=113 (N=390). In this methodology, the original time series can also be decomposed into a sum of M additive components and can be partially rebuilt using only the leading 'signal modes', thus filtering out background noisy components (Elsner and Tsonis, 1996; Vautard et al., 1992). In n-order SSA filtering, the leading n modes are used to rebuilt the original time series. The lower the number of retained modes, the stronger is the time series smoothing. If all M modes are used, the original time series is fully recovered.*"

P4, L5 P13, L5 P10 L26: "cross-validation" and "2-way validation" are misleading here.
Reply: These terms were removed from the manuscript.

P12, L26: The absence of the trend is not unlikely, but a fact. The absence "in reality" is unlikely.
Reply: This sentence was rephrased and it now reads as follows:

Lines 347-350: "*In effect, the absence of clear long-term trends in Lut2004 is not coherent with the significant changes in the radiative forcing throughout the last 400 years and the important role played by these external forcings on temperature variability over western Iberia (Gómez-Navarro et al., 2012).*"

Figures:

General: It would help to have at least one sentence in the captions which tells us what to look for. This is a matter of taste, of course, because this may lead to redundancy with respect to the text.
Reply: As the reviewer also acknowledges, we think that adding further text to the figure captions might be redundant, particularly because some of them would become excessively long.

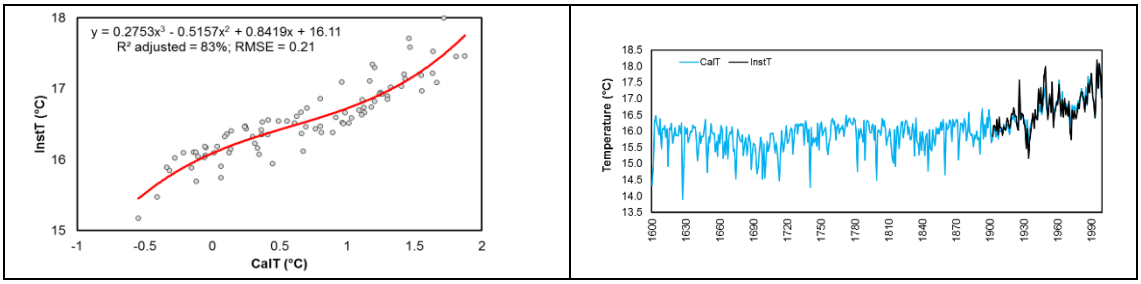
Fig 1: I suggest showing also the estimated background gradient in the top panel, perhaps the reduced temperatures the regression equations in the figure should be in physical units - T and z. As already mentioned I suggest to complement this figure with a HFD(z) plot, but this of course depends on how the authors choose to revise their text.
Reply: We think that adding the regression equations to both panels in Fig. 1 might be confusing for the reader. Hence, we chose to show the equations only in the bottom panel. As

suggested, the regression equations in the panels are now in the appropriate physical symbols (T and z). For the HFD(z) profile see our reply above to point 1.

Fig 2: OK if large enough in the final text.
Reply: Thank you for this suggestion. The size of the figures will be large enough for their correct display.

Fig 3: I guess this is a "robust" regression? Otherwise I would have expected a larger influence of the high leverage points (at the very low & high CalT values), which could be classified as outliers. Also, the distribution of the residuals is clearly not Gaussian.
You can see different behavior at CalT < 0.7 C and above. What does this mean with respect to the statistics? Is this significant?
Reply: The regression line is estimated by a standard least-squares approach (no weighting). In fact, the robust-bisquare approach leads to worse results (R-square adjusted of 78% instead of 82% and RMSE of 0.29 instead of 0.22).

We also agree that e.g. a higher order polynomial fitting would provide better adjustments taking into account the asymmetry of the residuals referred by the reviewer. In fact, a 3-order polynomial fitting, with a robust regression using the bisquare weights method, provides the best adjustment: R-square adjusted of 83% instead of 82% and RMSE of 0.21 instead of 0.22 (left panel in the figure below; equivalent to Fig. 3). Nonetheless, this cubic interpolation has an obvious disadvantage when extrapolating low values (downward spikes on the right panel; equivalent to Fig. 4a). Although using a higher order polynomial leads to a slightly better fit, extrapolation to temperatures outside the range of values used for model fitting, as well as some heteroskadicity (no stationarity in variance) in the residuals, clearly limit its application. This explains the choice of the linear interpolation, which is a compromise solution between the goodness of fit and the reliability of the transformed variable.



This information is now clearly stated in the manuscript so as to enhance this discussion of the model choice:

Lines 282-293: "*The consistency between InstT and CalT has been assessed by a linear regression in their common period (1901–1989). The corresponding scatterplot shows that linear regression provides a good fitting, with a correlation coefficient above 0.90 (Fig. 3), explaining about 82% of the total variance (R-square adjusted), and a RMSE of 0.22. According to the Fisher's test, this least-squares linear regression model is statistically significant at a 99% confidence level (p<0.01). A bootstrap procedure with 10,000 resamples shows that the 95% confidence interval for the correlation coefficient between InstT and CalT is [0.87, 0.93], supporting the Fisher's test. Therefore, CalT clearly reproduces the observed temperature in Portugal in the instrumental period (InstT). A 3-order polynomial fitting, with a robust regression using the bisquare weighting method, provides a slightly better adjustment (R-square adjusted of 83% and RMSE of 0.21), but its extrapolation for the lowest temperatures (outside the range of values used in the model fitting, not shown) is not reliable and was discarded.*"

To improve the statistical significance analysis of this regression estimation we have also added additional adjustment parameters to the legend in Fig. 3.

Fig 4: I suggest to improve or leave out the (c) panels: they should be at the same horizontal scale as the others. I do not find the white "cone of influence". Maybe this refers to another version of the plot? One might also argue whether this panel is necessary, as it does not contain much information. "panels" should be "panel".
Reply: We agree that Fig4c does not provide significant information. Therefore, we chose to remove it from Fig 4, including its related text in the manuscript.

Fig 5: I do not see "0 indices" - also "black edges" instead of "outer lines"?
Reply: The ‘0 indices’ are omitted for the sake of readability of the plot. This is now stated in the caption. As suggested, ‘black outer lines’ were also replaced by ‘black edges’.