Dear Anonymous Referee #2

For clarity, we repeat the reviewer's comments in blue italic font and the replies in black. Since general comments 1, 2, 3 are similar to the comments 4, 3, 2 by the anonymous referee #1, some of the replies are used in common.

The authors present and analyze a novel approach to directly assimilate proxy information into GCM simulations to reconstruct past climate. They find that while assimilation of isotopic proxies is possible and is clearly beneficial in idealized simulations, the actual benefit of assimilating proxy data is limited due to model errors and the small number of assimilated proxies. Data assimilation in paleoclimatology has attracted a lot of attention recently and the science and methods are developing rapidly. This manuscript represents an important contribution to the field in that for one of the first times, proxy data (rather than reconstructed climatic variables) are assimilated directly for climate reconstructions. Therefore, I recommend this article to be published after the outstanding issues detailed below have been addressed.

Thank you very much for the positive and valuable comments.

General comments:

1. *The sensitivity experiments conducted in this study only 'explain' a small fraction of the difference in correlation between the idealized setup (CTRL) and the application to real proxy data (REAL). The reasons for such a reduction in quality are manifold and include GCM model errors and errors in the proxy forward model that are not quantified in the current analysis. Proxy model errors are shortly discussed at the end of section 4, but it is not clear to me how one could attribute errors to the proxy model or the GCM in the absence of controlled experiments (as also stated by the authors in L504). While performing such controlled experiments with alternative proxy model / GCM combinations is clearly beyond the scope of this paper, I suggest the authors carefully reword the respective paragraphs.*

   Thank you for the comments. We understand that there are multiple factors other than model errors for the low skill in REAL experiment and that we do not know their relative contribution. Thus, we carefully modified the abstract and Sect. 6 in which we clearly mentioned that there remains a lot unexplained and avoided arguing that the model errors are the only reason for the degradation in REAL experiment.

2. *In addition to trying to quantify the limitations of the current proxy DA setup by performing sensitivity experiments, the authors also try to answer a second question: namely whether direct assimilation of proxy data is superior to assimilating climatic variables (here temperature) reconstructed from the proxy data. In contrast to the approach pursued here, it would seem easier to address this question using the REAL experimental setup. Based on this setup, one could derive reconstructed (gridded) temperature data from the exact same proxies that have been used in the REAL experiment and assimilate these reconstructed temperatures instead. Such an experimental framework would be instructive as to whether empirical proxy models (i.e. reconstructed temperatures) outperform the physics-based on-line proxy models. Alternatively, one could devise idealized experiments similar to the ones performed in the study in which one compares assimilations based on the assumption of a perfect proxy model. In contrast to the comparison presented here, one would need to compare the CTRL (or any other of the synthetic proxy experiments) to the corresponding experiment in which the proxy data (+ noise) from the truth run has been used to reconstruct temperatures which are then assimilated. Such analysis, however, may be beyond the scope of this paper and I would be perfectly happy if the authors decide to focus on the main message of the manuscript – the proxy data assimilation and partial attribution of its limited skill to quantifiable sources – only.*

Thank you for the comments. We modified the experimental setting for T2-Assim following your suggestion. In the modified experiment, temperature is reconstructed from the isotopic records which is used in CTRL by simple regression-based method. Proxies whose correlation with local temperature during calibration period (1871-1950) is not statistically significant ($p < 0.10$) are removed following Mann et al. (2008). This screening process reduced the available data from 94 to 81 grid points. Based on the correlation between isotope ratio and local temperature, SNR can be estimated through the equation (Mann et al., 2007):

$$\mathrm{SNR} = \sqrt{\frac{r^2}{1-r^2}}$$

where r is the correlation. SNR is shown in Fig. 8. Subsequently, this reconstructed temperature (T_r) is assimilated. The assimilated result is shown in Fig. 7. The result is slightly degraded in T2-Assim compared with CTRL due to relatively large error in T_r (Fig. 8). As shown in Dee et al. (2016), the reconstruction skill is somewhat compensated by the structure of Kalman gain. Figure S1 shows the correlation scale length to show the difference in the structure between CTRL and T2-Assim. The correlation scale length was found by computing point correlation between the prior (temperature) and the prior-estimated observation (temperature and δ^{18}O for T2-Assim and CTRL, respectively) for the observation grids, binning these correlations by distance, and computing the mean of each bin. The correlation is consistently high in T2-Assim, which means that the observation information is more effectively used to update the analysis. To sum up, the accuracies are not substantially different among proxy DA and reconstructed DA. However, we should note that this is only the case as long as the relation between temperature and isotope remain the same.

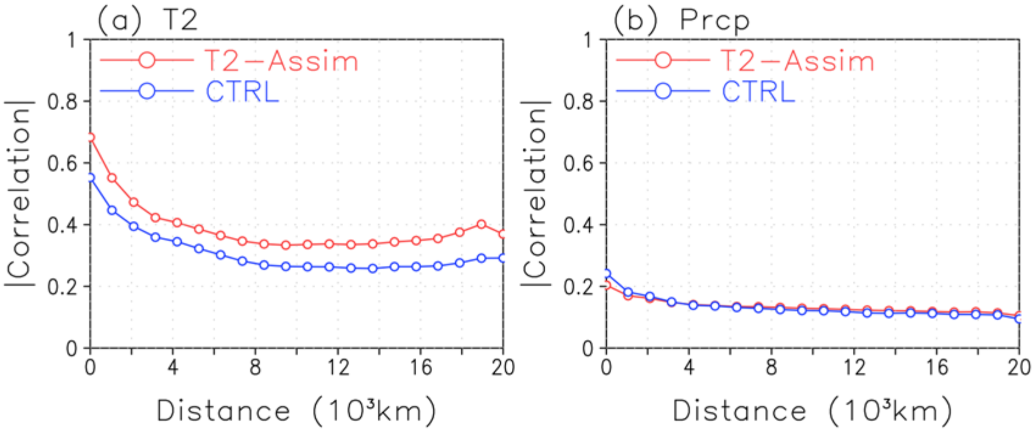

Figure S 1 Mean correlation scale length for T2-Assim (red) and CTRL (blue). The prior is (a) temperature and (b) precipitation. The prior-estimated observation is temperature and δ^{18}O in coral for T2-Assim and CTRL, respectively for the both panels.

3. *The data assimilation method is not described at all. Please add a short section on the data assimilation method with the relevant references. I suggest to focus on the choices and setup specific to this study and to provide the appropriate references; an in-depth introduction to the data assimilation method would only be needed if you chose a non-standard assimilation method that is not documented elsewhere. If, as suggested by the final paragraph of the manuscript, an EnKF has been used, then I suggest to also analyse the spread to error ratio or compute rank histograms to get an impression whether the analysis spread matches the analysis error and the analysis is well calibrated. Lack of calibration (usually overconfidence) is likely due*

to a misrepresentation of the observation error matrix (either underestimation of observation error or correlated errors).

The description of data assimilation method is included in the revised manuscript (L133-137). We used EnSRF (Houtekamer and Mitchell, 2001) with slight modification following the previous studies (Bhend et al., 2012; Steiger et al., 2014). As described in the manuscript, we did an offline approach, in which the analysis is not cycled to the simulation and the same background is used for every analysis step.

Figure S2 shows the spread and RMSE for surface temperature in REAL experiment. The posterior spread matches with RMSE for the first half of the period but it gradually diverges from the RMSE. This reflects the fact that the system has a difficulty in reconstructing temperature in mid- to high-latitude (Fig. 3), where temperature has been increasing in the period. However, the discrepancy does not necessarily mean that the system is not well calibrated. The relatively scarce observation and short correlation length scale must hamper the reproducibility there. On top of that, we speculate that the metrics such as spread-to-error ratio or rank histogram may not suit for the evaluation of the offline DA. In general, it takes several cycles for the spread to match with RMSE through improving the error covariance matrix in the online DA. Contrarily, because the offline DA uses the same background for every analysis step, the quality of the analysis error covariance remains the same (c.f. $\mathbf{P^a} = [\mathbf{I} - \mathbf{P^f H^t}(\mathbf{H P^f H} + \mathbf{R})^{-1}\mathbf{H}]\mathbf{P^f}$). Therefore, the spread will not tell how well the system is calibrated.

Instead, we show the sensitivity of the system to parameters (observation error and localization scale) in REAL experiment to show that the system is how optimal. The results show that the skill is moderately dependent on both the observation error and the localization scale. For the observation error, the results become better with larger error in the investigated range for the both variables. On the other hand, the sensitivity to the localization scale varies from variable to variable. For temperature, the correlation become better along with the scale in the investigated range. For precipitation, the localization scale of 12000km resulted in the best correlation.

The sensitivity is also different by which metric to be used. For instance, RMSE for precipitation becomes larger along with the observation error (not shown).

Given the results above and because the choice of the parameters does not change the main conclusion of the study, we keep using the original value of the parameters.

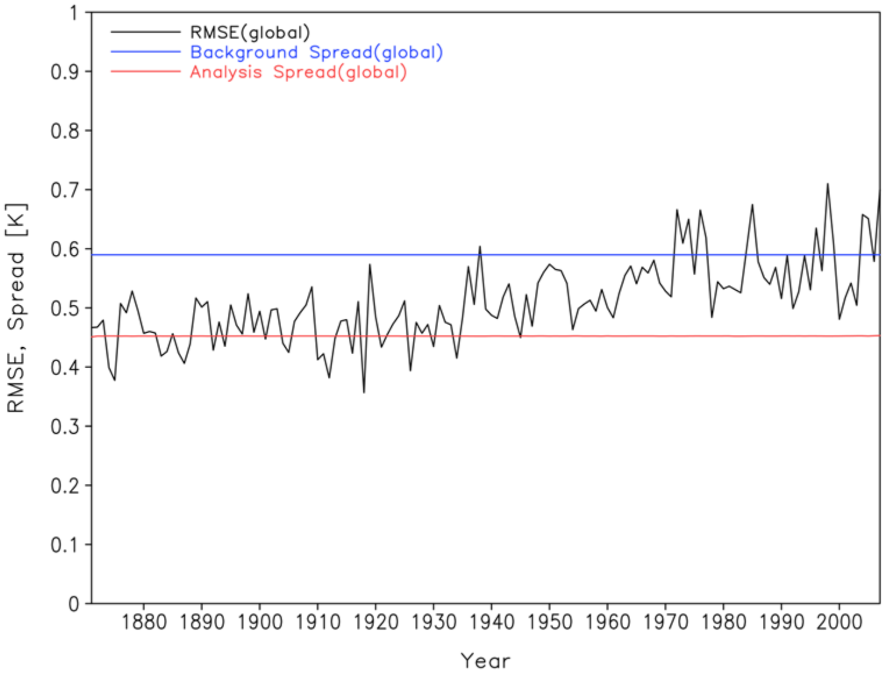

Figure S 2 Global mean of the spread and RMSE for surface temperature in REAL. The spread of background and analysis and RMSE are shown in blue, red, and black, respectively.

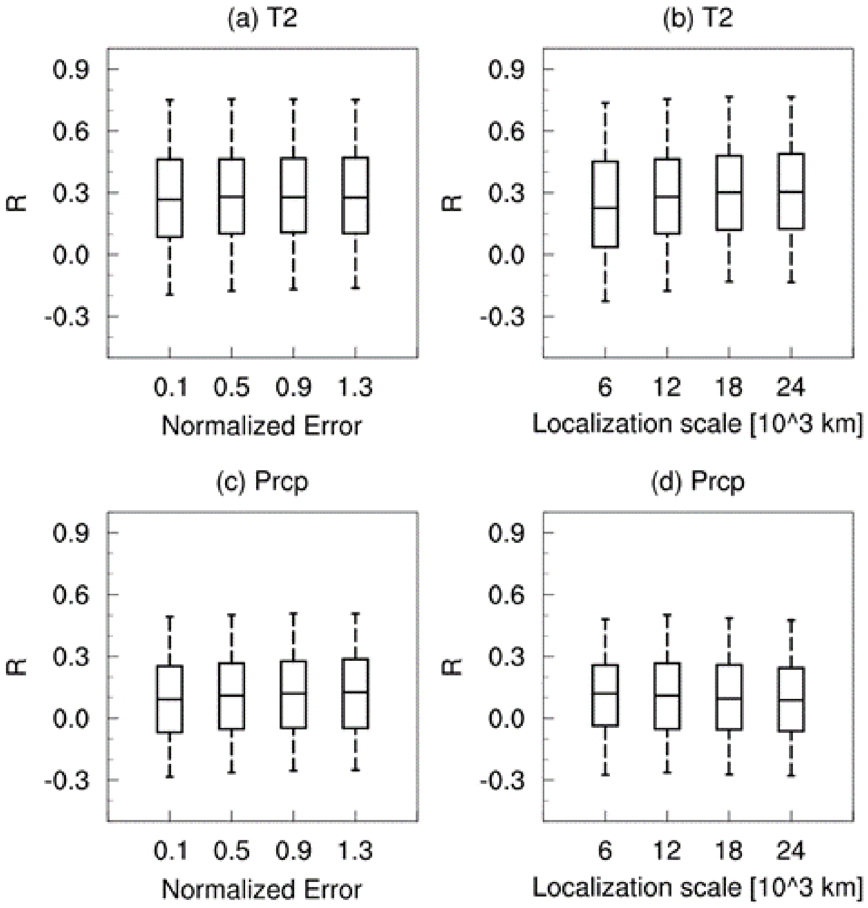

Figure S 3 Box-whisker plot of the distribution of all spatial values for the correlation for (a and b) temperature, and (c and d) precipitation in REAL. (a) and (c) shows the sensitivity to the observation error, and (b) and (d) shows the sensitivity to the localization scale.

4. *Use of the term ‘accuracy’: The authors repeatedly use the term ‘accuracy’ to describe the quality of the analysis. This use of language is somewhat misleading, as accuracy in forecast verification has a specific meaning and the appropriate verification score to measure accuracy would be the mean squared or mean absolute error, whereas the correlation is a measure of forecast / analysis association (e.g. Murphy, 1993). I suggest to either rephrase and write of “improved assimilation”, “enhanced correlation” etc. or to clearly state that accuracy refers to correlation throughout the manuscript.*

   Thank you for the comments. We used the term ‘reconstruction skill’ or ‘skill’ instead of ‘accuracy’ in the revised manuscript. In addition, we clearly stated that we use the correlation coefficient as a measurement of skill (L119-120).

Specific comments:

1. *L112: This issue seems important and I think it would be worth revisiting in the conclusions.*

   Thank you for the comment. We revisited the issue in the conclusion in the revised manuscript (L530-532).

2. *L267: stemming from*

   Corrected.

3. *L363-365/7: Is this a direct quote from the Xu et al. paper? If so I suggest labelling this as such by using quotation marks.*

   No, it is not. We modified the sentence for better readability (L380-385).

4. *L385: for precipitation*

   Corrected.

5. *L440: slightly more accurately?*

   Section 5.1 were substantially modified following the general comment #2. Accordingly, the sentence was not used any more in the revised manuscript. Thank you.

6. *L487ff: if the only difference in simulations is observed vs. simulated SSTs, I suggest the authors refrain from using the term forcing in the following lines for better readability.*

   In the revised manuscript, we substituted the term ‘SST’ for the term ‘forcing’. Thank you.

7. *L499ff: The discussion of the differences of the various sensitivity experiments is hard to read. I suggest to streamline and reword this section along the lines of “Imperfect SST used to drive the CGCM simulation resulted in a slight reduction of correlation compared to the CTRL experiment with perfect SST.”*

   The corresponding sentences were rephrased in the revised manuscript following the suggestion (L507-509; 512-514). Thank you.

8. *L513: non-climatic factors.*

   Corrected.

9. *L514: add reference, e.g. Appendix B of Compo et al. 2011*

   We added the reference (Appendix B of Compo et al., 2011) in the revised manuscript.

10. *L525: I suggest to mention that not in all cases direct proxy DA will be beneficial compared to assimilating empirically reconstructed variables. Also, while assimilating more data is expected to increase the quality of the analysis, care has to be taken in assimilating dependent information (e.g. direct assimilation of proxy data and reconstructed variables derived from the same proxy data).*

    The both sentences were included in the revised manuscript (L542-543; 544-546). Thank you.

11. *Figure 4: The figure labels denote EOF2 whereas only EOF1 is mentioned in the text. Please fix.*

    Thank you for your pointing out. The figure was replaced with the correct figure.