Reviewer #1

This paper tests several different methodological choices that are typically made (or could be made) in paleoclimate reconstructions using DA. I think this presents a good and valuable presentation and discussion of these choices. The findings and suggestions for future reconstructions are very helpful to the community performing these types of reconstructions.

We thank the Reviewer for the positive comments on our work and for the suggestion how to improve our figures.

Section 3.1: Is there any specific justification for the choice of the ratio of L_z and L_m being 2:1? Could this, or some other ratio, be justified by looking at the correlation length scale in observational data?

> The idea behind a longer length scale in zonal direction than in meridional direction is based on the zonal flow in the atmosphere. On multi-annual to multi-decadal time scales multiple processes act in meridional direction, e.g. a widening/shrinking of the Hadley cell, shifts of the ITCZ or changes in atmospheric modes like AMO or NAO. These can shift the the zonal circulation northward or southward but the zonal coherence will be less effected. That is the reason why we had the hypothesis that may have longer decorrelation distances in zonal direction. We will explain this hypothesis in the revised version of the manuscript.

Section 3.1: Is there any justification for the specific localization values that you chose for each variable that was reconstructed? Are these values data-driven or just educated guesses? Were any experiments done to test on optimal localization value? I would assume that if these values were used based on weather DA experiments, they might not apply on the longer paleo time scales where one would generally expect the correlation length scales to be larger.

> The localization length scale parameters were defined based on the spatial correlation of the variables in the monthly ECHAM model simulation fields. In Section 2.4 we refer to the paper by Franke et al. 2017 how the localization was done in the original setup. We used the same localization length scale parameters for localizing the sample covariance in most of our experiments to evaluate improvements in comparison with this initial setup. For this study, we calculated the latitudinal dependency of correlation of the state variables from a bigger ensemble of the model than in Franke et al (2017). The result suggested that the longer length scale parameters can be applied in the tropics and the predefined length scale parameter of precipitation is probably too strict. Based on the rather strict decorrelation length scale in the previous study and the assumption that the covariances can be better estimated from a bigger ensemble, we used doubled length scale parameters in some of the experiments for localizing the climatological covariances. In this case, the L for temperature is 3000 km, which means that the correlation is decreased close to zero approximately 6000 km away from the observation. We did not carry out further experiment to find the optimal localization value because even double the localization distance hardly changed the reconstruction skill. Hence, our system does not appear to be very sensitive on the localization distance as long as it remains in a reasonable range. On the one hand, we do not further restrict the localization because that would limit updates to a small regions around observations. On the other hand, our experiments without localization showed negative reconstruction skill in locations far away from observations, even with the error covariance matrix is calculated from climatology. We will provide this additional explanation in the revised manuscript.

Section 4.2.1: When you are comparing the distributions, you say that for example,the most skillful reconstruction is obtained from the 100c_PcL experiment. What is the basis for saying it's the best? What aspect of the distribution are you comparing? The median or some other specific value(s)?

Yes, for comparison we used only the median.

Many of the distributions shown in the figures look very similar so it was hard for me to feel confident about the statement that one particular set of reconstruction choices was better than another. Are the distributions statistically distinct?

We agree with the Reviewer that the distributions of the skill of the experiments over the extratropical Northern hemisphere look similar. We have not checked whether the distributions are statistically distinct. In the revised paper we will provide some statistical evaluation of the experiments.

Instead of comparing the distributions, would it be possible to show the differences compared to the "original"reconstruction (i.e., you'd compute the difference in the skill score for each location and then summarize this distribution of differences in the plots)? I'm wondering if this,or something similar, might make the differences more clear. Because currently when I look at the distributions, many of them look very similar and perhaps even statistically indistinguishable.

Thank you for your suggestion. We will make the plots as it was suggested and if the differences become more distinguishable we will replace the original figures (Fig. 4-7 and 9).

Fig 8 & 10: It would be very helpful to give a little more explanatory information/labeling on each panel, such as was done in Fig 3.

We will add more labels to the figures.