**Climate
of the Past**

Discussions

Interactive comment on "Testing the analog method in reconstructing the global mean annual temperature during the Common Era" *by* Juan José Gómez-Navarro et al.

Juan José Gómez-Navarro et al.

jjgomeznavarro@um.es

This is an interesting paper. I don't intend to give a comprehensive review but would like to add a few comments.

The authors thank the time and interest shown in reading our manuscript and making constructive suggestions.

The method of analogue selection seems very similar to what Goosse and co-workershave presented and described in a large number of papers (dating back at least to2006) as a simplified particle filter, so I'm surprised that this does not come up for discussion anywhere in the current manuscript. Admittedly the method of Goosse

et al does differ slightly in that they usually re-initialise the model from the selected best state in order to generate the prior for the subsequent year, but since there is very little predictive skill on the annual time scale, this difference in approach is relatively unimportant (as the authors here also note). Our own work (Annan and Hargreaves Climate of the Past 2012) also supports this view. While Goosse et al usually use a sample size of around 100 ensemble members, these do all share an appropriate external forcing whereas a large majority of the CMIP/PMIP ensemble of opportunity used here will consist of simulated years when the forcing and its recent history is (relatively) far from the truth. This will not matter if the forcing is small enough to be unimportant, of course, which may be the case for much of the last millennium. Our own work used an ensemble of approximately 10,000 model years in some idealised testing and found it inadequate for reconstructing local temperatures but useful on the hemispheric scale. It is not clear from the presentation of the results to what extent this is also the case in this manuscript and I suspect that some more examination of the EOF analogue might be instructive in this respect, since the adequacy of ensemble size for such particle based methods depends strongly on the dimensionality of the problem.

Certainly we overlooked these references. We will re-formulate part of the introduction and the discussions of the results to improve these aspects of the manuscript. As pointed out by J. Annan, there are some differences between Goose's method and ours. One important difference is that Goose et al use in their work a climate model of intermediate complexity, with a fairly coarse spatial resolution - this is the price paid to be able to run a relatively large ensemble of simulations. The price we pay is that, as also pointed out by J.Annan, our analogues are not tied to the truth by a common external forcing. The analogue reconstruction is just driven by the available proxy data. Assessing to what extent this is permissible and yields reasonable results is exactly the aim of this study. The rationale behind it is that the patterns of variations are sufficiently constrained by the available proxies, and are not so much dependent on the particular forcing that dominates in a given year, either solar variations, greenhouse gas, etc. This might not always hold, but the study is essentially a test of this assumption.

Although correlation is widely used as an assessment of reconstruction performance in the paleoclimate community, it is not standard in data assimilation and reanalyses and has the potential to convey a somewhat rosy view (in my opinion) of the actual quality of the results. I suggest that the authors also calculate the skill in more conventional terms, e.g. reduction in RMS error compared to the prior (in this case being the simple mean of the ensemble of model runs), both locally and hemispherically if that is desired. Finally, the authors only seem to use one model (GISS) as the source of truth. I wonder if this might have biased the results, e.g. if GISS is atypical in some respect. If not too expensive, I'd hope the authors could repeat the experiments (or at least a subset of them) using a much wider range of models as ground truth.

We used additional performance metrics, including RMSE, but decided focus on just one for the sake of brevity and clarity in the manuscript. We chose correlation for being a rather intuitive metric but we will include other measures of skill in the revised version.

Regarding the choice of GISS as target, we will evaluate the impact of using alternative targets and include the discussion of this result in the manuscript. We believe however that the results will be consistent, as there exists also consistency we show between the results in PPE and in the calibration period with actual data.