Review of 'A new global climate reconstruction for the Last Glacial Maximum '

The manuscript by Annan et al. is an interesting contribution to the literature on LGM temperature changes. The authors use a wide range of proxy-based reconstructions and available simulations. The resulting temperature fields thus can be seen as an aggregation of knowledge on LGM climate over the last few decades. Several important methodological advancements for the application of data assimilation in the paleoclimate context are presented in the manuscript, in particular the de-biasing of the prior ensemble and the a priori model selection to obtain more independent ensemble members. However, it is at times difficult to follow the presented results and explanations. Therefore, I recommend major revisions of the manuscript after which it should be a valuable resource for the community.
My major issues with the manuscript are as follows:

1. There are too few plots and the discrete color scales with often wide temperature steps make the current plots not very informative. The majority of SAT anomalies are between -2 and -8 K and should be represented by more than two different colors. The colors of the dots (proxy data) in Fig. 1 and 3 are difficult to identify. I was missing visualizations of
    - the difference between de-biased and not de-biased simulation ensemble (while the small difference in GMST is discussed in the text, a visualization of the spatially distributed difference is missing)
    - pairwise model similarity (the current discussion makes it impossible for the reader to understand how similar simulations needed to be such that only one of them was retained)
    - the validation results
    - the sensitivity tests
    - difference maps between the new reconstruction and discussed previous reconstructions (ideally including where differences are statistically significant)

2. A statement on data and code availability is missing. As the resulting fields will be a valuable community resource they need to be made available. To follow the choices made by the authors and compare them with previous approaches, code to reproduce the results should be made available.

3. The used error metrics in the validation and sensitivity sections are strongly focused on the posterior mean and the global mean temperature. As this is a Bayesian reconstruction of climate fields, a stronger focus on the spatial structures of the reconstruction (and how different choices discussed in the sensitivity tests influence it) and of the full posterior probability distribution would be more informative (e.g. maps of cross-validation results, coverage frequencies to study the meaningfulness of the posterior uncertainties,

probabilistic score functions). The included proxy data tends to be spatially clustered. Therefore, I wonder how much leave-one-out cross-validation is influenced by spatial autocorrelation and whether leaving our more data at once would be a better choice (e.g. h-block or leave-N-out cross-validation).

4. The abstract is too short and not very informative outside of the stated global mean temperature anomaly. I recommend to specify the used proxy data, simulations, and methodology. Actually describing the results from investigating the differences compared to Tierney et al. (2020) would be more informative for the interested readers than just writing "We discuss the reasons for this discrepancy".

5. The explanations on how statistical and data processing choices influence different features of the reconstruction are very valuable and should be a useful guide for future applications of data assimilation in paleoclimatology. However, they are often very general (e.g. p. 1 l. 2-3, 7; p. 2 l. 8-9; p. 3 l. 10, 26; p. 5 l. 16; p. 6 l. 12-13; p. 8 l. 9-10; p. 11 l. 12-14; p. 14 l. 4-5; p. 15 l. 5, 7-8, 19-20; see also the specific comments below) such that readers are forced to believe the authors and cannot trace the explanations in the results. Therefore, I recommend to specify these explanations.

Specific comments:

- Replacing "climate" by "temperature" in the title would be more precise.
- p. 1, l. 18-19 (and several subsequent instances): In which sense are the resulting fields 'physically consistent'? While the individual ensemble members are physically consistent, it is not clear to me how physically consistent the fields are after de-biasing and applying the ensemble Kalman filter.
- p. 2, l. 8-10: What are examples of absent localized features and suspected small-scale artifacts? Specifying the magnitude and spatial-scale of these features would be valuable to better interpret the results.
- p. 3, l. 10: What does "unusual ocean boundaries" mean? Has this been reported elsewhere? Given the numerous papers employing the PMIP3 ensemble, this finding of potential bugs in these simulations sounds relevant for others.
- p. 3, l. 26: What was used as cutoff values for RMS / pattern correlation? Giving more details seems necessary to interpret / reproduce the present details.
- p. 5, l. 16: What is the GMST anomaly of this outlier? What is the difference to the closest ensemble member? Given how much the GMSTs are used throughout the manuscript, plotting the GMST anomalies of the ensemble members (selected and removed ones) would be useful for the readers.
- p. 6, l. 12-13: How much of the variance is explained by the first four EOFs? How close is $\mathbf{q}$ to the final posterior mean?
- p. 7, l. 3: Cleator et al. (2020) augmented the Bartlein et al. (2011) data by Australian records from Prentice et al. (2017). These ones could be added to the dataset employed here. Given the new temperature reconstructions from ice cores that have been produced over the last decade, are the values from Schmittner et al. (2011) still up-to-date?

- p. 8, l. 13-15: Is the selection of TEA over MARGO based on the grid box averages or on the level of individual records? Given that TEA uses different archives than MARGO (inclusion of δ^{18}O in TEA, use of foraminiferal assemblages in MARGO) there might be some inconsistencies when the selection of MARGO over TWA is performed on the grid box level, in case there are systematic differences between different proxy types.

- p. 9-10: How is the specific implementation of the ensemble Kalman filter selected? How does it compare with the ensemble square-root Kalman filter employed in Tierney et al. (2020)? Are there previous examples of applying ensemble Kalman filters to multi-model ensembles? Are the assumption of it still satisfied for multi-model ensembles?

- p. 11, l. 12-14: A quantification of what "noticeably smoother" and "visually less structured" means would be very helpful.

- p. 14, l. 4-5: How is "no sign of over-fitting" measured? What would be considered over-fitting?

- p. 14, l. 18: It would help me if the rank histograms were shown and not just described.

- p. 14, l. 27-28: Which numbers of EOFs were tested? Assuming that the general cooling pattern is mostly contained in the first EOF (as they are not shown, I cannot determine that for sure), it would likely be more interesting to compare not just the GMST but the spatial patterns.

- p. 15, l. 5: How is the worsened fit to the data quantified?

- p. 17: Given the extensive sensitivity tests, are there general recommendations for the future usage of PMIP ensembles in climate field reconstructions and potentially for the design of future PMIP cycles? How could/should multi-model ensembles be designed for effective usage in climate field reconstructions?

References

Bartlein, P. J., Harrison, S. P., Brewer, S., Connor, S., Davis, B. A. S., Gajewski, K., Guiot, J., Harrison-Prentice, T. I., Henderson, A., Peyron, O., Prentice, I. C., Scholze, M., Seppä, H., Shuman, B., Sugita, S., Thompson, R. S., Viau, A. E., Williams, J., and Wu, H.: Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis, Climate Dynamics, 37, 775–802, https://doi.org/10.1007/s00382-010-0904-1, 2011.

Cleator, S. F., Harrison, S. P., Nichols, N. K., Prentice, I. C., and Roulstone, I.: A new multivariable benchmark for Last Glacial Maximum climate simulations, Clim. Past, 16, 699–712, https://doi.org/10.5194/cp-16-699-2020, 2020.

Prentice, I., Cleator, S., Huang, Y., Harrison, S., and Roulstone, I.: Reconstructing ice-age palaeoclimates: Quantifying low-CO2 effects on plants, Global and Planetary Change, 149, 166–176, https://doi.org/10.1016/j.gloplacha.2016.12.012, 2017.

Schmittner, A., Urban, N. M., Shakun, J. D., Mahowald, N. M., Clark, P. U., Bartlein, P. J., Mix, A. C., and Rosell-Melé, A.: Climate Sensitivity Estimated from Temperature Reconstructions of the Last Glacial Maximum, Science, 334, 1385–1388, https://doi.org/10.1126/science.1203513, 2011.

Tierney, J. E., Zhu, J., King, J., Malevich, S. B., Hakim, G. J., and Poulsen, C. J.: Glacial cooling and climate sensitivity revisited, Nature, 584, 569–573, https://doi.org/10.1038/s41586-020-2617-x, 2020.