R. Telford's comments

- In *Italic* : referees' comments
- In normal font : our answers

---

1. *The uncertainty of the reconstructions is estimated at 11%, which is ignored when model skill is estimated. Hence the test of model skill is not fair; even a perfect model might appear to have low skill because it cannot match the reconstructions' errors. Methods to incorporate reconstruction uncertainty into estimates of model skill were developed by Hargreaves et al (2013).*

Indeed, the uncertainty of the reconstructions was not taken into account when calculating the RMSE. This is discussed extensively in the revised version. Nevertheless, we are not able to use the method proposed in Hargreaves et al. (2013). To compute the proposed skill estimate, a reference is needed:

$$S = 1 - \sqrt{\left(\frac{\sum_{i=1}^{N} (m_i - o_i)^2 - \sum_{i=1}^{N} (e_i^2)}{\sum_{i=1}^{N} (n_i - o_i)^2 - \sum_{i=1}^{N} (e_i^2)} \right)}$$

where for each model grid i for which a proxy-based reconstruction is available: m_i is the model results, o_i the data, n_i the reference, and e_i the data error. N is the number of observations.

The natural reference here corresponds to no change between the mid-Holocene and pre-industrial conditions. This is also the first option proposed in Hargreaves et al. (2013). In that case, the term $\sum_{i=1}^{N} (n_i - o_i)^2$ is smaller than $\sum_{i=1}^{N} (e_i)^2$ and the skill becomes undefined as discussed in Hargreaves et al. (2013).

Because of this, the interpretation of the figure 6 has been changed following your comments. Instead of concluding that the models have no skill at local scale, we explain that it is impossible to quantitatively evaluate the skill of the models since the signal of the data and of the models is very low and since the data error is of the same order as their signal. Furthermore, as suggested, we focus on the physical consistency of the pattern rather than on the numerical value of the error. The abstract, the conclusions and a full paragraph in section 3.2 have been updated in this sense.

2. *As regards the uncertainties of the data:*

We agree that the question of uncertainty is a difficult one. However, this is out of the scope of our study to discuss it in detail. Here we take the simplest option to keep the value proposed by de Vernal et al. (2013) since it is the only study that gives quantitative estimates with error bars. We insist in the revised version on the limitations implied by this choice.

Reference
de Vernal, A., Hillaire-Marcel, C., Rochon, A., Fréchette, B., Henry, M., Solignac, S., and Bonnet, S.: Dinocyst-based reconstructions of sea ice cover concentration during the Holocene in the Arctic Ocean, the northern North Atlantic Ocean and its adjacent seas, Quaternary Science Reviews, 79, 111–121, doi:10.1016/j.quascirev.2013.07.006, 2013.

Hargreaves, J. C., Annan, J. D., Ohgaito, R., Paul, A., and Abe-Ouchi, A.: Skill and reliability of climate model ensembles at the Last Glacial Maximum and mid-Holocene, Clim. Past, 9, 811–823, doi:10.5194/cp-9-811-2013, 2013.