Response to reviews of Telford, Li and Kucera (2012)

We are grateful for the three reviews our manuscript received. Reviews 2 and 3 are mostly positive, with review 3 requests additional information and analyses. The more negative comments in review 1 are based, we believe, on a misunderstanding of the methods used.

Review 1 first argues that the Modern Analogue technique is not affected by the correlation structure in the calibration set environmental data (point 1)

> *The authors stated that many problems that previous transfer function methods may encountered, for example, the multiple environmental interferences or joint controls of more than one environmental variables that may caused the transfer function estimates invalid or in large errors. This is very true especially to the transfer function methods that adopt regression equations (such as Imbrie-Kipp Method) as the co-variance problem always occurs when it was regressed against one variable dependent on more than one parameters. However, in this paper the authors actually used MAT (Modern Analog Technique) that is a method based on dissimilarity coefficients between fauna assemblage data without any assumption of correlation or response function models. I am not clear that the "errors" generated by the MAT is that closely related to the co-variance or multiple environmental control problem inherent from our core top data base;*

Review 1 implies that because the Modern Analogue Technique does not assume a response model between the environmental data and species abundances the assumptions of transfer functions (Birks et al. 2010) do not apply. This is an error equivalent to arguing that the non-parametric Wilcoxon test makes no assumptions because it does not assume a Gaussian distribution for the data.

How the assumptions of transfer functions affect MAT has been little studied, with the exception of spatial autocorrelation, but Guiot and de Vernal (2007) have shown that MAT is sensitive to the correlation between summer and winter SST. Below we demonstrate, using two arguments that MAT is sensitive to the correlation structure in the data.

First, consider a case with two environmental variables *x* and *y* which are correlated in the calibration set. Even if x is not ecologically important, it will appear possible to reconstruct *x* with MAT because of its correlation with *y*. If, in the past, the correlation between *x* and *y* was different, reconstructions of *x* will be spurious.

Second, foraminifera assemblages are potentially influenced by multiple environmental variables, for example temperature at different depths and in different seasons, food availability, salinity, etc. It is mathematically not possible to reconstruct all such potential controlling variables in this high-dimensional environmental space experienced by the foraminifera as there are fewer dimensions, i.e. species, in the assemblage data, than the number of potential controlling environmental factors. The system is underdetermined. However, some environmental variables are more important determinants of foraminifera assemblages, and other variables are correlated with these, so many environmental variables appear possible to reconstruct. But if this correlation structure changes, the reconstructions may be spurious.

For methods that use the global relationship between species and the environment, correlations between environmental variables across the whole calibration set are most important. For methods like MAT that use only a local subspace for any reconstruction, it is the correlation between variables in this subspace that is most important. It is not obvious that the problem is circumvented by needing to consider many subspaces rather than the whole calibration set.

Following on from this, review 1 (point 2)

> *I don't see any sign of "errors" on the downcore estimations is directly coming from the multiple controls. The subsurface warming, sometimes shown in the downcore records (for example, V22-222 in Fig. 4), may reflect true signals as the thermocline became thick during the glacial near the subtropical gyre in the north Atlantic. It's hard to say any error in the downcore estimates except that we have independent evidence (geochemical SSTs, etc.);*

One interpretation of our reconstructions of marked warming from the LGM to the Holocene in the subsurface coupled with little near-surface temperature change at V22-222, would be that the mixed layer was thinner during the LGM. An alternative interpretation is that at least one of these reconstructions is incorrect. It would be trivial numerically to make reconstructions of a great many environmental variables, of temperature and nutrient status for different seasons and depths, but we would not expect all of these to be equally valid. Likewise, we need to be cautious before accepting both the surface and subsurface reconstructions from V22-222. The low dimensionality of most fossil time series suggests that it is unlikely that multiple independent reconstructions can be generated from a single set of fossil data. The methods presented in our current paper and in Telford and Birks (2011) can help identify which reconstructions are best: these methods show that the subsurface SST reconstruction explains more of the variance in the fossil data than the near-surface reconstruction. This, together with the results from figure 3 that shows that MAT has limited ability to reconstruct tropical near-surface temperature in parts of the modern ocean lead us to prefer the subsurface reconstruction.

Review 3 (point 1) asks to modify the references that are used to support some of our arguments on foraminifera ecology and ocean circulation. We thank the referee for pointing out several valid issues and we will modify the references as appropriate. The reason we were sparse in citing literature on these issues is that the points we make are not contentious (if not trivial) and we feel we only need to provide one reference as example that supports these statements. Any other approach will require a thorough historical literature review, which we believe is not needed in this case.

Review 3 (point 2) asks why we used the MARGO foraminifera calibration dataset rather than dataset of Salgueiro et al. (2010), which is somewhat larger, having added over 100 observations from the Portuguese margin to the MARGO data set. We believe that expanding the calibration dataset will make little change to the results, except perhaps for sites in the area augmented by Salgueiro et al. (2010), and no change to our conclusion that it is not appropriate to use a single calibration/reconstruction depth across the entire ocean and that 10m SST is an inappropriate calibration target at least in parts of the Atlantic. We thank reviewer 3 for their suggestion of additional sites and agree that the inclusion of additional sites would important in a paper discussing the palaeoceanographic implications of our conclusions, which we hope to write shortly, but it will not change the conclusions of this paper, which we see more as a proof of concept.

Review 3 (point 3) asks us to explain what depth range we are referring to with “subsurface” vs. deeper (permanent thermocline?) temperatures (e.g. p. 4082 lines 21-22). In this context, we were simply describing figure 5 with no explicit reference to the water column structure. We will try to clarify this text.

We will add some oceanographic information to p. 4082 where it might help interpretations as requested by review 3 (point 4). Consequently this text will not become shorter as requested by review 2 (point 2), but should become more interesting.

Review 3 (point 5) raises the interesting question of the role of surface and deep dwelling taxa in causing reconstructions from some depths to explain more of the variation in the fossil data than others.

> *p. 4081 and ff. and Fig. 4: Not being familiar with the faunal data I wonder how much of the variability seen at the different depth levels is driven by changes in the faunal composition. So I would like to see two records added for each core site: 1) a sum of the major surface dwelling planktonic foraminifers (e.g., G. ruber + G. sacculifer); 2) a sum of the major deep dwelling planktonic foraminifers (e.g., G. menardii, G. tumida, G. truncatulinoides, G. hirsuta)*

This is a valid point and we will explore this issue before submitting a revised manuscript. We fear that the requested records for each site of the sum of the deep and surface dwelling taxa might not be very informative, as variability rather than abundance might be more important. An alternative strategy we will explore is to make reconstructions with the two sets of species, with the expectation that reconstructions of near surface SST will explain more of the variance in the surface fauna whereas reconstructions from deeper water will explain more of the variance in the deep dwelling fauna.

Reviews 2 (point 1) and 3 (point 7) suggested that we use the same y-axis for the plots in figure 5. While we understand the motivation for this, we do not think it useful in this instance as it is the shape of the profile, not the absolute values, that is most important. At least part of the difference in absolute value between sites is due to differences in species richness: the null expectation is for a reconstruction to explain more of the variance in a species-poor polar foraminifera record than in a species-rich tropical or subtropical record. We will improved the caption for this figure and rearranged the plots so nearby sites are adjacent (also in Table 1); they were previously arranged strictly north-south.

Regarding figure 5, review 1 (point 3) asks

> *As the authors adopted MAT approach to estimate SSTs, I don't understand what are the "proportion of variance explained by —" (in Fig. 5). The proportion of variance is normally generated by using a fauna matrix decomposed from coretop data in interpreting downcores. Are these numbers on the Y-axis of Fig. 5 dissimilarity coefficient?*

The calculation of the "proportion of variance explained by the reconstruction" is explained in the methods section. We use redundancy analysis, a constrained ordination. Our calculations are not related to communality nor are they a dissimilarity coefficient. We hope our expanded caption to figure 5 will remind the reader.

Reviews 2 and 3 both request additional discussion or analysis of the CMIP5 analysis shown in figure 6.

Review 2

> *3) Discussion of Fig. 6 results is a bit too brief: it would be interesting to know where the most similar profiles in the pre-industrial are located and if this location is the same for all models.*

Review 3

> *6) Comparison to climate models: please give more information on the models as not every reader will be familiar with the particularities of each model. For example, is any of the model a transient model or are the years given in Table 2 the year after spin-up for the respective time slice. Since the model data is available why is there no comparison to/ discussion on the subsurface to "deeper" temperature variations seen in the climate model(s) in comparison to data variability shown in Fig. 4?*

There is a wealth of information in the CMIP5 runs and previous CMIP and PMIP model runs have been under-exploited. However, we believe that further CMIP5 analyses are outside the scope of this paper, the main aim of which is to show that 10m is not necessarily the most appropriate depth for reconstructing temperature from foraminifera assemblages. We hope, in the near future, to write a second paper that will use the ideas from the current paper and develop a new understanding of conditions in the Atlantic at the LGM, incorporating foraminifera assemblage-temperature reconstructions, other proxy data, and a detailed analysis of the CMIP5 runs.

With regard to the information about the climate models, the simulations and output used are described in section 2 (p4080 L8-14); we will clarify that these are time-slice simulations, as specified by PMIP3. It is standard practice to document the simulation years from such time-slice (i.e., equilibrium) experiments such that others may reproduce the results. Since these model simulations are not transient, it is not possible to generate time series analogous to the reconstruction time series shown in Fig 4.

Review 1 (point 4) writes that they can't read the "Euclidean distance" on Figure 6. We will increase the font size on this figure.

Review 1 (points 5) suggests that we should show some examples and why and where the forward models could improve the results.

> *The authors concluded that "the forward models" of planktic foraminifer assemblage may improve the results. To accomplish a complete article in the "Climate of the Past", I suggest that*

the authors should show some examples and why and where the forward models could improve the results.

We already cite both available examples of forward models and describe how forward models could be used. We had hoped that the remainder of the text would show that traditional transfer functions for reconstructing SST have unresolvable problems that limit their potential, which would argue for the potential of alternative techniques to be explored. We will revise the text to make this clearer and consider citing successful applications of forward models in other areas of palaeoecology.

We accept most of the minor changes the reviewers suggest.

References

Guiot, J. and de Vernal, A.: Transfer functions: methods for quantitative paleoceanography based on microfossils. In: Developments in marine geology, 1: 523-563, doi:10.1016/S1572-5480(07)01018-4, 2007.

Telford, R. J. and Birks, H. J. B.: A novel method for assessing the statistical significance of quantitative reconstructions inferred from biotic assemblages, Quaternary Sci. Rev., 30, 1272–1278, 2011.