We would like to take this opportunity to thank the reviewer for their helpful comments on our manuscript. We will implement all suggested changes in a revised manuscript. Below we have copied the reviewer's comments and added our response in red and proposed changes to the manuscript in blue.

With kind regards,
Lukas Jonkers and Michal Kucera

==========

I have reviewed the manuscript presented by the authors and have found it to be well written, interesting and concise. The topic it addresses has indeed been genrerally overlooked by the wider community and there is no doubt that more work of this sort is necessary in the field of paleoclimatology. They begin by ranking the species by importance. The method considered from Juggins et al. (2015) seems valid, but I wonder whether other methods could be considered and if the results would be significantly different; I imagine not so much. The authors then showed that for both MAT and WA, it is possible to obtain rather different paleo reconstructions when using a different number of species for the calibration of the model, even though those models had very similar error of prediction in the training dataset. This underlines that the prediction error in the training dataset is not by itself a sufficient metric for characterizing the uncertainty in reconstructions. This also applies to other microfossils. It would be interesting to analyze the spatial patterns of the prediction errors in order to identify the sources of variation related to species-pruning.
The manuscript leaves many questions unanswered, but I understand that not all questions can be answered in a single article. I will be looking forward for future work concerning the quantification of the species-pruning uncertainty, and the identification of non-climatic biotic and abiotic factors leading to the uncertaitnties.
I found the treatment of the authors satisfying and did not find any important flaw with the analysis presented. Therefore, I recommend that it can be published without major revisions, and below I include a list of specific points which should be taken into account to improve the quality of the manuscript, particularly many figure captions should be revised.

We thank the reviewer for the encouragement. We agree with the evaluation that some issues are left unanswered. One of these – the ecological reasons for species ranking – was also highlighted by reviewer 2, and we thus decided to expand the discussion on this particular topic. In addition, we would like to highlight at this point that due to an error discovered by reviewer 2, the species importance ranking for MAT in the North Atlantic has changed. This change only applies to the importance ranking of the species, it has no fundamental implications for the conclusions of our manuscript, but it leads to substantial changes to the discussion. For more details, please refer to our response to reviewer 2.

Minor corrections:

Line 15: ‘information [about] all species’

Done.

Line 108: If they were not reported, I imagine they were not counted. I am not sure what the authors did, but for consistency, such species should then also be removed from the calibration dataset.

We agree with the reviewer that this is potentially an important point. However, when counting foraminifera assemblages for SST reconstructions it is common practice to count all species and we assume here that species that were absent in the entire dataset (e.g. tropical species in a polar assemblage) were simply not reported (or, conversely, the authors only reported species which occurred at least once in their dataset). Such treatment of missing species is a common practice with synthesis datasets in paleoecology (e.g. MARGO).

Lines 113: How many samples are kept for this part used to build the transfer function? I found the brief explanation hard to understand and had to read Juggins to understand. I would suggest rewriting this paragraph in a more straightforward way.

Sample selection for the bootstrap and OOB is carried out randomly with replacement. This means that approximately 63 % of the samples are used to build the model and the remainder is used as the OOB. We will add this to the description of the method.

Line 118: I do not understand why the approach is repeated 10 times after the 1000 bootstrap. Couldn’t the error estimate be obtained from the bootstrap estimates directly? Is it that the 1/3 selection of species is the same for any given set of 1000 bootstraps?

We realised that a very large number of bootstrap samples is required the obtain stable results. One could of course also obtain the error directly from the bootstrap estimates, but the results would be similar and in this way we could make use of R packages without too much modification of the code. The species selection is variable and random across all bootstraps.

Line 123: Does this mean you do not renormalize the species abundances after selecting a subset of species? If yes, then couldn’t this lead to problems if for example we have two sites with a similar composition regarding the ‘useful’ species which are sensitive to say temperature, but one has a large amount of irrelevant species temperature-wise while the other has almost none.

The reviewer is right in that there is no re-normalisation; all remaining species are included in (a virtual) rest group. In this way, the potential issue, quite correctly raised by the reviewer, is circumvented. This procedure also mimics a common situation where researchers who only counted the most abundant species also often provide the total number of planktonic foraminifera.

Line 150: I would indicate the number of species in parenthesis for MAT (around 6 species) and for WA (around 9 species).

Will do.

Line 201: Add comma after 'In WA'

We will reword the sentence.

Line 240: Is expatriation a synonym for sediment mixing or a different process? It is mentioned, but neither defined or discussed really.

Expatriation refers to the incidental/occasional advection by ocean currents of species outside their normal habitat. We realise that this is in fact not a post-depositional process and will omit it here.

Figure 1: Not sure that incremental is the right work, an incremental change could be large. Maybe "marginal" or a synonym would be more appropriate.

We agree and will follow the suggested rewording.

Figure 3: Could you define what years you are considering for the LGM. Even better would be to add a shaded background over that period which you averaged in the top row graphs.

We follow the MARGO definition of LGM (23 to 19 kyr BP), but this is strictly speaking irrelevant here. The different figures are showing two entirely independent datasets. We will highlight this in the caption. The two datasets are used to show the influence of species selection in two different situations (time-slice and time-series).

Figure 4: The caption could be rewriten more clearly, it is not clearly stated that the difference is with the reconstruction with the full taxonomic resolution.

We apologise, some text appeared to have gone missing. We will change the caption to: '*Figure 4: effect of species selection on SST reconstruction. Graphs show mean difference between reconstructions using species-pruned transfer function models and the model that includes all species. The results are ordered according to species importance, with each dot representing the result from a transfer function model with the species up to the marked point (similar to Fig. 1). Grey symbols are the reconstructions with fewer than the minimum number of species and red symbols the reconstructions with the minimum number of species. Dots are the average of the mean absolute difference, diamonds the mean difference.*'

Figure 5: Is the prediction error calculated using the full timeseries or only the LGM as previously? Also, it might be useful to indicate the number of species used for the 1st point, does it start at n=2? I imagine that the increments of increasing number of species is simpy 1, i.e. ni+1=ni+1. Cool figure.

The reviewer is correct and the start is at two species and the increments are always a single species. We will add this to the figure caption. However, the prediction error is based on the cross-validation of the transfer function model in the core top data set and thus independent of the reconstruction.

Figure 6: Second sentence could be revised grammatically speaking.

For clarity we will change this sentence to: '*This sensitivity is defined as the standard deviation of the reconstructions with more than the minimum required number of species (i.e. based on transfer function models with uninformative species, those to the right of the red dot in Fig. 4).*'