Reviewer 1

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 changes to the manuscript in blue. We have tracked all changes and a revised manuscript is appended below. Line number refer to this version of the manuscript.

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 (lines 113-114).

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).

We have rewritten this section following the discovery of a mistake in our results. The number of species required is now similar for both methods.

Line 201: Add comma after ‘In WA’

Done.

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 have omitted 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 have followed 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 figure 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 have changed 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 added 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 changed 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).*'

Reviewer 2

We would like to thank the reviewer for their critical look at our manuscript and are grateful for spotting an unexplained mistake in our analysis (species ranking in the North Atlantic). We have repeated all analysis and will make the code publicly available. Whilst the species ranking for MAT in the North Atlantic is indeed different from our original submission, the conclusions of our study remain unaffected: the sensitivity to species pruning provides a useful means to assess the influence of other environmental variables in transfer function-based reconstructions. However, the now correct (and replicable) species ranking leads to changes in the discussion. Below we have copied the reviewer's comments and provided our response in red and proposed changes are indicated in blue. We have tracked all changes and a revised manuscript is appended below. Line number refer to this version of the manuscript.

With kind regards,
Lukas Jonkers and Michal Kucera

================================================================================

This manuscript uses the methods developed by Juggins et al (2015) to determine the importance of each planktic foraminifera for the performance of transfer functions calibrated against sea-surface temperature. These results are used to make models with a greatly reduced set of species which perform almost as well as models based on the full set, however, the reconstructions were different.
The manuscript is generally well written.

I found some of the results surprising. In particular, the low importance of N. pachyderma with the modern analogue technique (MAT). This is an abundant (up to 100%) and common taxon with a clearly defined thermal niche. I found the explanation in the manuscript reasonably persuasive for transfer function performance, but not for importance.
Intrigued, I re-implemented the method and, with a smaller training set, and found N. pachyderma to be among the most important taxon for MAT. Subsequently, I was given access to the authors' code which gave the same result. The authors blame a glitch and now get the same result. I encourage the authors to adopt the techniques described in the British Ecological Society guide to reproducible code.

At least some of the subsequent results will be affected by this glitch.

Repeat analysis showed indeed that the species importance ranking for MAT in the North Atlantic was erroneous. Our new analysis shows that N. pachyderma is the most important species (see new figure 1). This ranking has some consequences for the remaining part of the manuscript:

i) it affects the number of species required to achieve a prediction error similar to the solution with all species;

ii) the species importance ranking is now more similar for both methods;

iii) it affects to some degree the reconstructed temperatures (Figs 3 and 4).

However, the most important finding of our analysis remains unaffected: we still find many different reconstructions with the same prediction error (see figure 6)

Based on the prediction error alone, these reconstructions appear equally valid, thus highlighting a previously unrecognised source of uncertainty in transfer function based environmental reconstructions. We have updated the manuscript to reflect these changes.

Despite extensive efforts, we were not able to reconstruct where the error occurred. It probably resulted from using erroneously an earlier version of the analysis, or an issue with file naming. However, the error only affected the one analysis and after sharing the code between us and the reviewer, we could establish that the new ranking is replicable. We apologise for this mistake and applaud the reviewer for their stringency.

The manuscript gives some consideration as to why some taxa are not important, but it would be interesting to see this expanded. I can think of several reasons. Taxa with low importance might be nuisance taxa, these should have negative importance. Taxa with poorly defined niches, or broad tolerances will have low importance and may be good to exclude from transfer functions. In contrast, taxa with low abundances or occurring in only a small number of sites may have low importance as the analysis will give these low weight even though they might be valuable indicators. It would be useful to try to quantify how much these factors contribute to low importance.

The reviewer points out an interesting aspect of our study and given the comments by reviewer 1 we expand the discussion on this topic. We agree with the reviewers' categorisation, although we would like to point out that we did not find any strict nuisance species with a negative importance. We assume that a species' importance for transfer function is dependent on i) its abundance, ii) its niche width and iii) its sensitivity to the environmental variable that best predicts the entire assemblage. To assess how these factors influence species importance we calculated average and maximum (99th percentile) abundance (in %), the thermal niche width (in deg C) and the temperature sensitivity (expressed as the goodness of fit (r2) of the species abundance to a Gaussian curve) for each species and assessed how well these parameters explained species importance using linear regression with increasing number of variables. We acknowledge that some of these variables are autocorrelated (e.g. temperature sensitivity and abundance are positively correlated, suggesting that the distribution of rare species is not primarily governed by temperature (in agreement with their low importance), or that our metric of temperature sensitivity is inadequate), but we believe this is of minor importance for the exploratory analysis we carry out.

For all ocean regions and for both MAT and WA we find that abundance appears to explain most of the variance in species importance. For the North Atlantic for instance the r2 of a linear model that only includes maximum abundance is 0.79 and 0.91 for MAT and WA respectively. Addition of niche width raises the r2 slightly to 0.83 and 0.93 and further addition of temperature sensitivity has a negligible effect (r2 0.84 and 0.94). This indicates that species abundance is the most important factor controlling species importance. This is probably due to the simple fact that abundant species are consistently present and allow best definition of thermal niche. We agree with the reviewer that rare taxa with a well-defined narrow niche might be good indicator species, but (if they exist) their usefulness for transfer functions is limited by their low abundance, leading to inconsistence incidence. We will add the following to the discussion:

We have added this discussion to the section where we consider species importance (lines 153-168):

*To understand why some species are more important than others, we consider their overall maximum abundance, the width of their thermal niche in the training set and their temperature sensitivity as potential predictors of importance. We define temperature sensitivity based on how well the species abundance in temperature space can be described by a simple Gaussian curve. This analysis reveals that abundance (Fig. 2) is the best predictor of species importance (Table 1). Indeed, multiple regression models that include all three variables perform only marginally better than a model using abundance alone. However, we note that all three variables are correlated to some degree. Interestingly, abundance and temperature sensitivity are positively correlated (r = 0.84), implying that the thermal niche of abundant species is better defined compared to rare species (Fig. 3A). We also observe that temperature sensitivity and thermal niche width are correlated. Counterintuitively, this correlation is positive: species with a narrow thermal niche appear less temperature-sensitive (r = 0.60; Fig. 3B). We attribute this pattern to a combination of low abundance of species with narrow thermal niches (sensitivity being correlated with abundance) and the possibility that their distribution is not primarily governed by temperature, assuming that the narrow thermal niche may be an artefact of adaptation to specific oceanic regions or regimes, only secondarily correlated with temperature.*

Whereas importance can be evaluated objectively, the effect of species inclusion rules on reconstructions cannot be as the truth is not known. Adding more taxa to the re- construction will obviously tend to make the reconstruction more similar to the all-taxa reconstruction. But any individual taxa could make the difference larger. Contrary to line 192, I do not regard the increase in difference with the inclusion of G glutinata as evidence that it is a nuisance taxon. It might be more powerful to run this analysis on the calibration set.

We agree with the reviewer as was reflected by our statement "*The analysis also reveals that there are many uninformative species, but very few – if any – real nuisance species…*". The reason why we assigned G. glutinata a possible nuisance status is for two reasons i) inclusion of this species in the calibration dataset leads to an increase in the prediction error (Fig. 1) and ii) inclusion of this species in the reconstructions leads to a jump in the difference in reconstructed temperature (Fig. 4). Our original submission already included this analysis, as reflected in the sentence "*G. glutinata in the case of WA, which leads to an increase both in the prediction error and the difference of the reconstruction, supporting its potential rating as a nuisance species*".

I am not sure we need to see the reconstructions with in 3A with fewer than the "minimum number" of taxa, as no one should be using these.

We agree. These have been removed from the figure (now 4).

The bias observed in for MAT in figure 3B looks much like what I would expect given N. pachyderma had been omitted.

This figure (now figure 4) will be changed to reflect the revised species importance ranking. The patterns for MAT and WA are now more similar:

It would be interesting (line 210) to compare the sensitivity of WA and MAT on the same range of models to identify the inherent variability of each.

The revised species importance ranking yields a top ten of most important species that overlap by 90 % (see figure above). The sensitivity of the results to the species importance ranking (the inherent variability) is therefore now (using the new ranking for NA) to a large extent assessed. We also note that the uncertainty in importance of the individual species are rather large, rendering the ranking of the unimportant species more uncertain. We see therefore little merit in repeating the analysis with 'swapped' species rankings.

Line 268. "that reconstructions that are highly sensitive to species pruning may indicate that the observed assemblage changes cannot be attributed solely to the environmental variable that is to be reconstructed"

This is an interesting idea, but I don't think this conclusion is justified given the results in the current version of the manuscript.

Our updated results still reveal the same patterns in the reconstructions and we find it hard to identify an alternative explanation, so we prefer to uphold the statement. We hope the revised analysis in the discussion will help to convince the reader that our conclusions are justified.

Whereas the manuscript demonstrates that only a few taxa are important for temperature reconstructions, I would hope that micropalaeontologists continue to count the full assemblage so that a range of questions can be addressed with the data. Only for routine analyses (for example water quality monitoring) is identifying only the important taxa justified.

We agree with the reviewer. Our intention was not to encourage counting fewer species, but to quantitatively investigate the effect of species pruning. This was motivated in part because some legacy datasets include only a limited number of species and we wanted to assess if these could be used. We will include an explicit statement in the revised version to encourage scientists to continue working with complete taxonomic resolution (lines 318-319).

Minor points

Line 176. Celsius not centigrade

Addressed.

Capitalisation of axis labels in all figures needs to be checked.

Done.

Figure 2. Might be better to use scale_fill_continuous(trans = "log", breaks = ...) than to log transform the percent.

OK.

The authors report (line 312) that the code is available on request. It would be much better to archive the code on, for example, github, or better still a permanent archive such as zenodo.org, ideally before review.

We agree with the reviewer. Our code is now available on: https://github.com/lukasjonkers/species_selection and will be made available on zenodo upon acceptance of our manuscript.

# Sensitivity to species selection indicates the effect of nuisance variables on marine microfossil transfer functions

Lukas Jonkers and Michal Kučera
MARUM – centre for marine environmental research, Bremen University, Leobener Straße 8, 28359 Bremen, Germany.

Keywords: planktonic foraminifera, transfer functions, MARGO, WA, MAT

## Abstract

The species composition of many groups of marine plankton appears well predicted by sea surface temperature (SST). Consequently, fossil plankton assemblages have been widely used to reconstruct past SST. Most applications of this approach make use of the highest possible taxonomic resolution. However, not all species are sensitive to temperature and their distribution may be governed by other parameters. There are thus reasons to question the merit of including information about all species, both for transfer function performance and for its effect on reconstructions.

Here we investigate the effect of species selection on planktonic foraminifera transfer functions. We assess species importance for transfer function models using a random forest technique and evaluate the performance of models with increasing number of species. Irrespective of using models that use the entire training set (weighted averaging) or models that use only a subset of the training set (modern analogue technique), we find that the majority of foraminifera species does not carry useful information for temperature reconstruction. Less than one third of the species in the training set is required to provide a temperature estimate with a prediction error comparable to a transfer function that uses all species in the training set. However, species selection matters for paleotemperature estimates. We find that transfer function models with different number of species but with the same error may yield different reconstructions of sea surface temperature when applied on the same fossil assemblages. This ambiguity in the reconstructions implies that fossil assemblage change reflects a combination of temperature and other environmental factors. The contribution of the additional factors is site and time specific, indicating ecological and geological complexity in the formation of the sedimentary assemblages. The possibility of obtaining multiple different reconstructions from a single sediment record presents a previously unrecognised source of uncertainty for sea surface temperature estimates based on planktonic foraminifera assemblages. This uncertainty can be evaluated by determining the sensitivity of the reconstructions to species pruning.

## 1 Introduction

Any method to infer quantitative paleoenvironmental information from (fossil) species assemblages relies on the use of a modern (or near modern) training set to identify a statistical relationship between the species assemblage and the environmental variable to be reconstructed. A range of methods exists, varying broadly in the way the training set is used to define the relation between the species community and the environmental parameter of interest. In this context, one aspect of the training set design that has received little formal attention, is species selection. In applications based on marine microfossils, the assumption has been that (nearly) all species potentially carry useful information and training sets have been designed to include as many species as practically possible (Kucera et al., 2005). However, there are fundamental, theoretical reasons to question if full taxonomic resolution is necessary for accurate reconstructions. Firstly, species may respond to other environmental variables than the one of interest and therefore confound the reconstruction. Such species would negatively impact the predictive power of the transfer function model. Secondly, species could be opportunistic or have a wide ecological niche and hence provide little information on the environmental variable to be reconstructed. These species would add no, or very little, information to constrain the reconstruction. The premise of full taxonomic resolution has been recently tested by Juggins et al. (2015), who demonstrated that transfer function techniques are variably sensitive to the influence of non-informative taxa. These authors used highly diverse assemblages from coastal and lacustrine environments, where the proportion of uninformative or confounding taxa can be expected to be high and where species selection for transfer function models is intuitively appropriate.

However, marine microfossil assemblages (notably planktonic foraminifera) are less diverse and the species display a low degree of endemicity. In these circumstances, it would appear important to retain as many species in the training set as possible. Indeed, up to now, species pruning in marine microplankton applications has been done, if at all, for practical, rather than ecological reasons. Species were removed or lumped because of low abundance, preservation or taxonomic issues (Hayes et al., 2005; Vernal et al., 2001; Zielinski et al., 1998). However, the minimum number of taxa required for a robust transfer function model, or the influence of non-informative or nuisance taxa on such models, has never been formally evaluated. This is remarkable, considering that many species of marine microplankton appear to have similar response to the modelled variables and that there is also evidence for responses to multiple variables (Telford et al., 2013; Siccha et al., 2009; Steinke et al., 2008). Furthermore, from a practical point of view, reducing the number of species in the model may improve counting statistics and reduce counting time. Moreover, reducing taxonomic resolution may also be beneficial for application of automated identification and counting systems (e.g. Beaufort and Dollfus, 2004; Hsiang et al., 2016) and enable the use of legacy datasets with incomplete taxonomy. The lack of a formal evaluation of species pruning in marine microfossil studies is also relevant because of the prolific use of different transfer function techniques. The evaluation of species pruning by Juggins et al. (2015) was carried out for ‘global models’ where the entire training set is used to define the species response curve to an environmental variable. Such methods are used on marine microfossil applications as well (e.g. Imbrie and Kipp method (Imbrie and Kipp, 1971), weighted averaging (WA) or artificial neural networks  (Malmgren and Nordlund, 1997)). However, in the marine studies the most popular approach is the modern analogue technique (MAT). This approach uses similarity measures to identify a small ‘local’ subset of samples from the training set to derive the environmental variable to be reconstructed. Such a ‘local’ approach is fundamentally different and could be sensitive to species pruning in the training set in a different way.

Here we evaluate species importance for transfer function performance for two widely differing methods (WA and MAT), representing both ends of the spectrum between local and global methods. We start by assessing species importance to determine the ranking of species for transfer function performance. We then investigate why some species are more important than others. Finally, because the behaviour of transfer functions models cannot be assessed within the modern training set only, we assess the influence of species selection on paleotemperature reconstructions.

We use planktonic foraminifera as a model group, but the results from this study are likely of wider relevance to paleoecological reconstructions based on marine microfossils. Planktonic foraminifera are ideal for this purpose because of the existence of a large global training set (Siccha and Kucera, 2017) and because of strong evidence that their distribution in surface sediments can be predicted by sea surface temperature alone (Bé and Hutson, 1977; Morey et al., 2005).

## 2 Data and Methods

To derive the transfer functions, we use planktonic foraminifera assemblage data from core top sediments recently compiled in the ForCenS dataset (Siccha and Kucera, 2017). Multi-species categories and morphotypes were removed and in case of replicate samples (at the same location) one sample was randomly selected. Annual mean temperature was assigned to each sample in the training set based on data from climatology (Stephens et al., 2002) averaged over the upper 50 m and within a 100 km radius of each core top sample. To test the effect of species selection on the performance of the transfer function outside of the calibration dataset, we analysed three fossil datasets. To evaluate the effect on assemblages with different diversity, we use two downcore records from the Atlantic: M35003-4 from the Caribbean (Hüls and Zahn, 2000), which spans 0-55 ka BP, and a longer,  0-180 ka BP, record from the Iberian Margin: MD95-2040 (De Abreu et al., 2003). And to evaluate the effect on past spatial SST patterns, we reanalyse the MARGO Last Glacial

Maximum (LGM) dataset from the North Atlantic Ocean (Kucera et al., 2005). The taxonomy of all fossil samples was harmonised with the ForCenS dataset following the same criteria as in Siccha and Kucera (Siccha and Kucera, 2017). Species not reported to be present in the downcore assemblages were assumed to be absent.

To determine how many and which species are needed for the transfer function models we start with assessing species importance following the approach described by Juggins et al. (2015). Briefly, this involves randomly selecting 1/3 of the species in the training set and dividing this reduced training set in a part that is used to build the transfer function model including approximately 63 % of the samples and an out-of-bag (OOB) part that is used to evaluate the model. The proportion of each species in the OOB selection is then randomly permuted and the performance of the model is reassessed. Species for which the permutation leads to an increase in the prediction error are considered important. The species importance is derived from the average difference in prediction between the OOB and the permuted OOB samples across 1,000 bootstrap samples of the training set. The entire approach is repeated 10 times in order to get an error estimate on the species importance. For WA we use inverse deshrinking to obtain the environmental variables and for MAT we use the 5 closest analogues determined using squared chord distance. We note that the way the species importance is assessed and the transfer function models with different numbers are designed is implicitly using a rest group of non-included species. This is because the calculations are based on the relative abundances of the species in the training set with the full taxonomic resolution. This procedure simulates a situation where the total number of fossils of the studied group has been determined, but only a subset of the species has been counted.

Next, we use the species importance to build transfer function models with successively increasing number of species. We start with the two most important species and increase the number according to the species importance ranking. The performance of each model (i.e the prediction error) was assessed using h-block cross validation to account for the effect of spatial autocorrelation in the training set (Telford and Birks, 2009). We used a cut off distance of 850 kilometres, which was shown to be appropriate for the North Atlantic (Trachsel and Telford, 2016). Because of the presence of cryptic species with specific ecological preferences, we have carried out the analyses separately for individual ocean basins following Kucera et al. (2005), assigning the new Red Sea samples of ForCenS to the Indian Ocean. The discussion below focusses on the North Atlantic Ocean because of the abundance of both core top and down core data, but the species ranking for the other oceans is provided in the supplementary information. All calculations were carried out in R (R core team, 2016) using the packages rioja (Juggins, 2017), raster (Hijmans, 2017), vegan (Oksanen et al., 2018) and ggplot (Wickham, 2016).

## 3 Results and discussion

### 3.1 Species importance ranking

Irrespective of which regional training set is used, for both WA and MAT only a small number of species appears important (Fig. 1, supplement). As shown by the example of the North Atlantic, cross validation of the transfer function models with increasing number of species shows that there is a large tail of low-importance species that do not add information and hence do not lead to improved transfer function performance (Fig. 1). In general, it appears that reconstructions with similar errors to the full species reconstruction can be achieved with less than a third or the total number of species. To reach prediction errors at, or very close to, the possible minimum, fewer species are needed for MAT than for WA. The species importance ranking varies somewhat by method and region, but is generally similar (supplement). In the North Atlantic there is a ninety percent overlap of the species in the top ten.

**Deleted:** seventy

**Deleted:** , with the most notable exception being the cold-water species *N. pachyderma*, which appears unimportant for MAT, whereas it is ranked as number one for WA

To understand why some species are more important than others, we consider their overall maximum abundance, the width of their thermal niche in the training set and their temperature sensitivity as potential predictors of importance. We define temperature sensitivity based on how well the species abundance in temperature space can be described by a simple Gaussian curve. This analysis reveals that abundance (Fig. 2) is the best predictor of species importance (Table 1). Indeed, multiple regression models that include all three variables perform only marginally better than a model using abundance alone. However, we note that all three variables are correlated to some degree. Interestingly, abundance and temperature sensitivity are positively correlated (r = 0.84), implying that the thermal niche of abundant species is better defined compared to rare species (Fig. 3A). We also observe that temperature sensitivity and thermal niche width are correlated. Counterintuitively, this correlation is positive: species with a narrow thermal niche appear less temperature-sensitive (r = 0.60; Fig. 3B). We attribute this pattern to a combination of low abundance of species with narrow thermal niches (sensitivity being correlated with abundance) and the possibility that their distribution is not primarily governed by temperature, assuming that the narrow thermal niche may be an artefact of adaptation to specific oceanic regions or regimes, only secondarily correlated with temperature.

**Deleted:** and

**Deleted:** (Fig. 2)

**Deleted:** Species that appear important for transfer function performance are in the first place those that are abundant (Fig. 2).

There are nevertheless differences between the two methods: for MAT inclusion of warm-water species only appears sufficient to obtain a minimum prediction error, whereas for WA cold-water species are also required. This is in line with 'global' nature of the technique, which requires characterisation of thermal niches of the species across the entire temperature range of the training set. The former likely reflects the implicit inclusion of information on the abundance of the remaining (cold-water) species and analogues are found because the similarity is based on relative abundances with respect to all species.

The analysis also reveals that there are many uninformative species, but very few – if any – real nuisance species, i.e. species that lead to an increase in the prediction error when included in the transfer function model. The single possible species in this category are *G. glutinata* in the case of WA (Fig. 1). *Globoconella inflata* is a species with a very wide thermal tolerance and a high degree of cryptic diversity (Darling et al., 2017; André et al., 2014) suggesting that its response to temperature may be complex.

### 3.2 Effect on reconstructions

Despite the apparent lack of importance of many species in the transfer function model development, their inclusion matters for the reconstruction based on fossil assemblages. Transfer function models with pruned species numbers yield reconstructions that are systematically different from reconstructions with all species included (Fig. 4). The differences can be up to several degrees celsius and variable in time and space, with important implications for paleoceanographic interpretations. As expected, the average difference between a reconstruction with all species included and pruned species reconstructions, decreases with increasing number of species (Fig. 5). Importantly, species pruning does not only result in more variability in the reconstructions, but leads to a real bias towards either lower or higher temperatures than a reconstruction with all species (Fig. 5).

Inclusion of the most important species leads to rapid decrease in the difference between species-pruned and all-species reconstructions, but unlike the species importance assessment (Fig. 1), adding more species leads to further stepwise decreases in the difference (Fig. 5). Importantly, steps in reconstruction difference occur with the addition of different species in different cores, indicating that they result from specific downcore changes in the assemblage composition. The exception appears to be *G. glutinata* in the case of WA, which leads to an increase both in the prediction error and the difference of the reconstruction, supporting its potential rating as a nuisance species. When adding species to the transfer function model, the reduction in the difference from the reconstruction with all species is initially accompanied by a reduction in the prediction error (Fig. 6). However, once the most important species are included, the prediction error stabilises whilst the difference continues to decrease. This means that for the same fossil assemblage, multiple different reconstructions with the same prediction error are possible. This pattern is visible in data sets from different faunistic and climatic regions and different time spans. Moreover, it holds for both methods, suggesting that it is a general pattern of the effect of species selection on SST

**Deleted:** P

**Deleted:** and *G. bulloides* for MAT

**Deleted:** *Both*

**Formatted:** Font: Italic

**Deleted:** are

**Deleted:** their

**Deleted:** 3

**Deleted:** centigrade

**Deleted:** 4

**Deleted:** 4

**Deleted:** Especially in the case of the MARGO reconstruction using MAT, the reconstruction with the minimum number of species is fundamentally different from the one that uses all species (mean absolute difference > 4 °C), even though the prediction errors are virtually identical. For WA the average LGM cooling is relatively insensitive to species pruning (Fig. 4), yet, as for MAT, the spatial pattern of reconstructed temperatures is, showing consistently lower temperature off northwest Africa (Fig. 3B).¶

**Deleted:** 4

**Deleted:** 5

reconstructions using planktonic foraminifera assemblages. In WA species pruning has a greater effect on the reconstruction error than in MAT, but for both methods species pruning leads to different reconstruction with the same error. Taken at face value, and in the absence of independent evidence, such inherent ambiguity renders it impossible to decide which of the reconstructions is more realistic. This adds a previously unrecognised source of uncertainty to quantitative assemblage-based reconstructions.

**Deleted:** on the differences between the reconstructions. The opposite is visible for

### 3.3 Sensitivity to species pruning

To evaluate the cause of the sensitivity of reconstructions to species selection, we calculate for each fossil sample a sensitivity measure based on the standard deviation of the reconstructions using different number of species and evaluate its predictability by properties of the assemblages (Fig. 7). The pruning sensitivity is all fossil datasets higher for MAT than for WA (Fig. 7) despite the fact that the minimum number of essential species is higher for MAT. This suggests that MAT is inherently more sensitive to species pruning, perhaps because the method does not rely on extrapolation to define a species' niche.

**Deleted:** 6

**Deleted:** 6

**Deleted:** . This partly reflects the fact that WA requires more species to reach minimum prediction error (Fig. 1) and the range of reconstructions used to calculate the pruning sensitivity includes fewer transfer function models with more species included than for MAT.

Intuitively, it would be expected that the effect of species pruning will be larger in low-diversity assemblages. However, for both techniques and all datasets, pruning sensitivity is unrelated to the number of species present in the fossil sample (Fig. 7). The fact that pruning sensitivity is not related to richness confirms that diverse assemblages contain many unimportant (redundant or uninformative) species. Especially for MAT, it could be expected that fossil samples with a composition that differs most from the training set will be most sensitive to pruning. This could be so if some species that are judged as uninformative in the training set carry environmentally relevant information downcore. Indeed, we observe such effect of analogue quality, with poorer analogue assemblages being more sensitive to pruning (Fig. 7). As expected, the effect is stronger for MAT, but it is most clear in the downcore records. However, the fact that the effect is generally weak and the absence of a clear common pattern in species pruning sensitivity (Fig. 7) highlights the difficulty in attributing assemblage changes to a single environmental factor and suggests that each site, or time slice, has unique species turnover dynamics. The uniqueness of species dynamics in each dataset is also reflected in the different position of the distinct steps in the difference between the reconstructions with all species and the pruned reconstructions (Fig. 4). Because these steps are not accompanied with a change in the transfer function performance, they indicate changes in the fossil assemblages that either reflect a temperature sensitivity of the species that is not captured in the training set (inadequate training set) or are not related to temperature (non-analogue condition (Hutson, 1977)). Non-analogue situations could arise either secondarily from post-depositional changes in the assemblages, or primarily and reflect a biological response to another environmental

**Deleted:** 6

**Deleted:** 6

**Deleted:** only

**Deleted:** visible

**Deleted:** 6

**Deleted:** 3

variable than temperature, which is not apparent in the training set because of temporally variable covariation between this predictor variable and temperature.

Post-depositional changes in species assemblages could arise from processes like dissolution (Berger, 1968) or sediment (and thus fossil) mixing due to bioturbation (Hutson, 1977; Kucera et al., 2005). The effect of mixing should be most pronounced around intervals of assemblage change and may create fossil species assemblages with poor analogues in the training set. We have assessed to what degree analogue quality (dissimilarity from the training set) varies as a function of change in the assemblages for the downcore records investigated here (Fig. 8). The observed relationship has the a sign consistent with our hypothesis, but is weak and not apparent in each time series, suggesting that analogue quality varies as a function of multiple parameters and does not simply reflect sediment mixing. In the analysed fossil samples, dissolution should be negligible and since expatriation is unlikely to have changed considerably between the training set and fossil samples. We thus infer that the major reason for the observed ambiguity of the reconstructions is the effect of nuisance (non-temperature) variables on fossil assemblage composition.

**Deleted:** , expatriation

**Deleted:** 7

**Deleted:** right

**Deleted:** direction

### 3.4 Implications for paleoecological reconstructions

Despite its intuitive simplicity and in spite of apparent statistical support (Morey et al., 2005), relating species assemblage change to a single environmental variable is not trivial (Juggins, 2013; Telford and Birks, 2009; Telford and Birks, 2005; Telford et al., 2013). This means that quantitative reconstructions based on transfer functions must be interpreted with caution and evaluated in view of additional, independent evidence. Our analysis provides further proof for the temporal emergence of apparently non-temperature related changes in fossil planktonic foraminifera assemblages that are not captured by calibration (Telford et al., 2013; Steinke et al., 2008; Siccha et al., 2009). The ambiguity in reconstruction due to species selection thus highlights the potential of environmental variables that appear unimportant in shaping species communities today, in explaining changes in past assemblages. An important question is therefore: how can this ambiguity be considered when interpreting individual reconstructions? We recommend as a first step to use the approach outlined here (ranking of species importance for each basin is provided in SI) to quantify the sensitivity of reconstructions to pruning and use this as an indication of the influence of secondary/nuisance variables on the reconstruction. This requires that the data are available at full taxonomic resolution and we explicitly do not encourage researchers to only count the apparently important species. Although a method to translate the effect of pruning into a mechanistic and quantifiable uncertainty estimate is not available, one may conclude that reconstructions that are highly sensitive to species pruning indicate that the observed assemblage changes cannot be attributed solely to the environmental variable that is to be reconstructed.

**Deleted:** evidence

**Deleted:** So, the

**Deleted:** and evaluate its potential geological and biological drivers

Ultimately, we need a more mechanistic understanding of the factors that determine species assemblage composition in the sediment. Importantly, planktonic foraminifera species inhabit vertically and seasonally distinct habitats (Jonkers and Kučera, 2015; Rebotim et al., 2017). Thus, in many cases the species in a sediment assemblage have never actually lived together and their abundance is thus unlikely to reflect the same forcing. Moreover, the controls on vertical and seasonal abundance variability are species specific, adding even more complexity to deriving a single environmental variable from an assemblage of different species. Ecological models that explicitly simulate assemblages from multiple environmental variables (Kretschmer et al., 2018; Lombard et al., 2011) may aid to improve our capabilities to quantitatively reconstruct past environmental change from species assemblages. In a climate modelling context, such forward modelling of fossil assemblages is likely more fruitful than directly comparing the inferred temperatures.

**Deleted:** almost all

## Conclusions

There are both theoretical and practical reasons to investigate whether full taxonomic resolution is required to infer paleoenvironmental data from microfossil assemblages. We have addressed this issue using planktonic foraminifera, but we believe that our results are relevant to other groups of (marine) microfossils too. We have ranked species according to their importance for transfer function model development and shown that less than a third of the species is needed to derive a model that performs as good (or better than) a model with full taxonomic resolution. Nevertheless, even though addition of more species has little effect on the transfer function model performance, sea surface temperature reconstructions with increasing number of species, are different. Thus, multiple different reconstructions are possible and their reliability cannot be assessed by transfer function performance in the training set. Our analysis suggests that fossil assemblages do not uniquely reflect a single environmental variable (sea surface temperature in this case), but rather provide an integrated response to biotic (but not temperature related) and abiotic (sediment mixing) factors. We have identified a new way to detect uncertainty (ambiguity) in transfer-function reconstructions due to nuisance variables. The sensitivity of reconstructions to species selection can be quantified using the approach outlined here, facilitating an assessment of the robustness of the reconstruction.

## Acknowledgements

We are grateful for the constructive comments by two anonymous reviewers. We thank Steve Juggins for advice at the initiation of this project. LJ acknowledges support from the climate modelling initiative PalMod, funded by the German Federal Ministry of Education and Research (BMBF).

## Data and code availability

All datasets used in this study are in the public domain: ForCens: https://doi.pangaea.de/10.1594/PANGAEA.873570; WOA: https://www.nodc.noaa.gov/OC5/WOA01/pr_woa01.html; Core MD95-2040: https://doi.pangaea.de/10.1594/PANGAEA.66811; Core M35003-4: https://doi.pangaea.de/10.1594/PANGAEA.55756; MARGO LGM: https://doi.pangaea.de/10.1594/PANGAEA.227329.

Code is available at https://github.com/lukasjonkers/species_selection

**Deleted:**

**Deleted:** upon request from LJ.

**Formatted:** Not Highlight

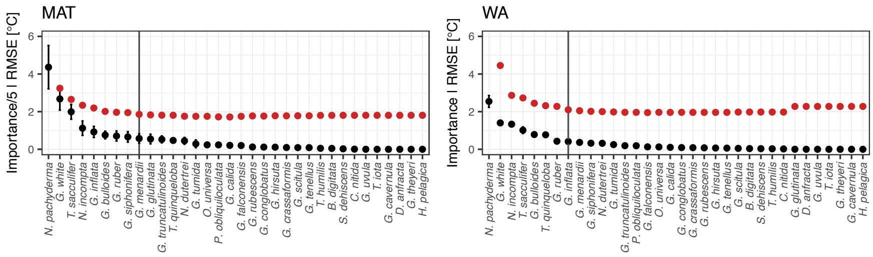

Figure 1: species importance (black) and transfer function performance (red) for models including all species to the left of the point (e.g. the first red point for the MAT transfer function for the North Atlantic denotes the prediction error (RMSE) of a model with *G. ruber* (white) and *T. sacculifer*). Error bars on the species importance show the standard deviation of 10 replicates; see methods for details. The dark vertical lines indicate the species needed for a transfer function model where including more species only leads to an marginal reduction in the reconstruction error (arbitrarily set at when the prediction error is within 10 % of the minimum). Note that for display purposes the species importance for MAT has been divided by a factor 5.

**Deleted:** with minimum error (MAT), or

**Deleted:** incremental

**Deleted:** WA

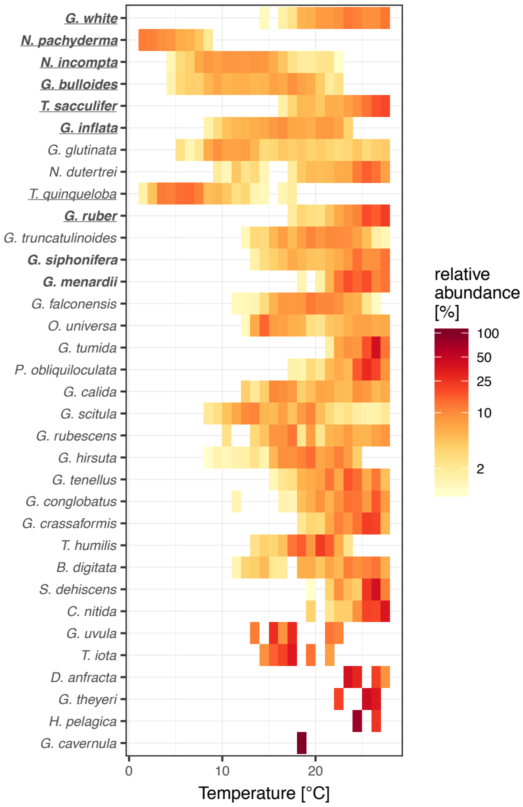

Figure 2: species thermal niche, ranked by average sedimentary abundance (top to bottom). Colours indicate average abundance (on a logarithmic scale) within one-degree centigrade bins. The essential 400 species (see Fig. 1) for MAT are marked in bold and those for WA are underlined. Despite their extremely well-defined thermal niche, rare species are not essential for transfer functions and those that are important are all abundant.

**Deleted:** i.e. needed for a transfer function model performance with minimum prediction error

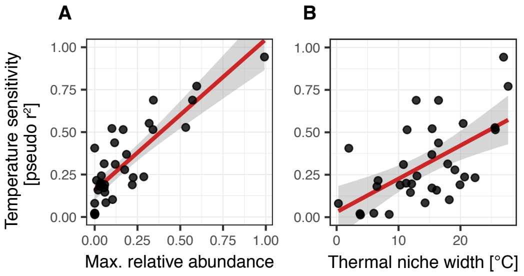

Figure 3: Relationships between temperature sensitivity and (A) maximum relative abundance and (B) thermal niche width of planktonic foraminifera species in the ForCenS dataset. The positive correlation between temperature sensitivity and abundance suggests that the thermal niche of abundant species is better defined than that of rare species, rendering abundance a good predictor of species importance. Panel B suggests that species with a narrow thermal niche are relatively unsensitive to temperature, suggesting the distribution of these, often rare, species is not primarily governed by temperature.

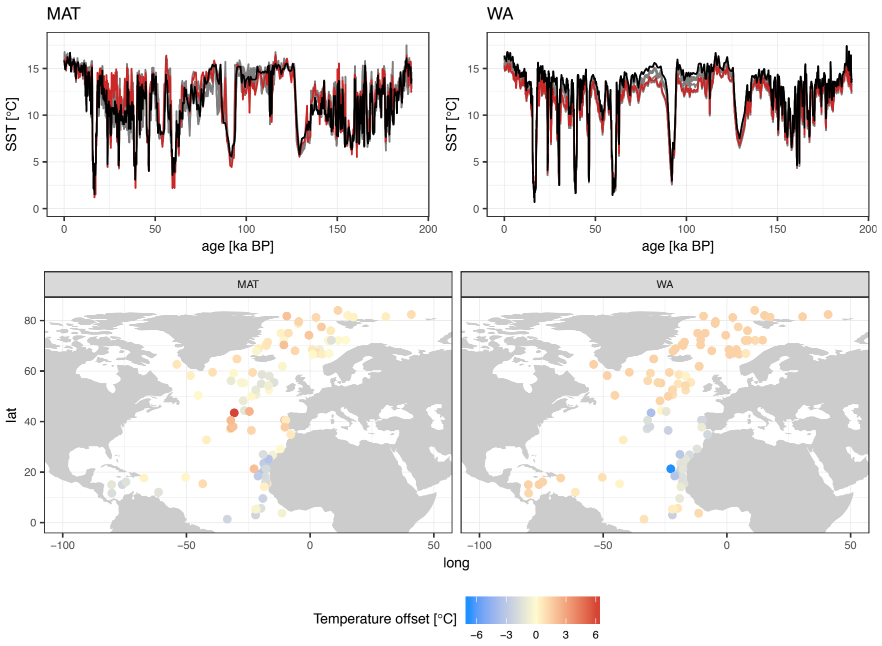

Figure 4: species selection affects SST reconstructions, examples from two independent data sets, showing the effect on temporal and spatial patterns in reconstructed SST. A: SST reconstructions for core MD95-2040 from the Iberian Margin using 33 possible transfer function models with the number of species increasing according to their ranking. Red lines show the reconstructions with essential species only, dark grey lines the reconstructions with more species and the black line is the 'final' reconstruction with all species. B: Spatial pattern of the difference between the SST reconstruction for the LGM with the minimum and with all species included.

**Deleted:** 3

**Deleted:** changes

**Deleted:** Thin pink lines are the reconstructions with fewer than the minimum number of species, r

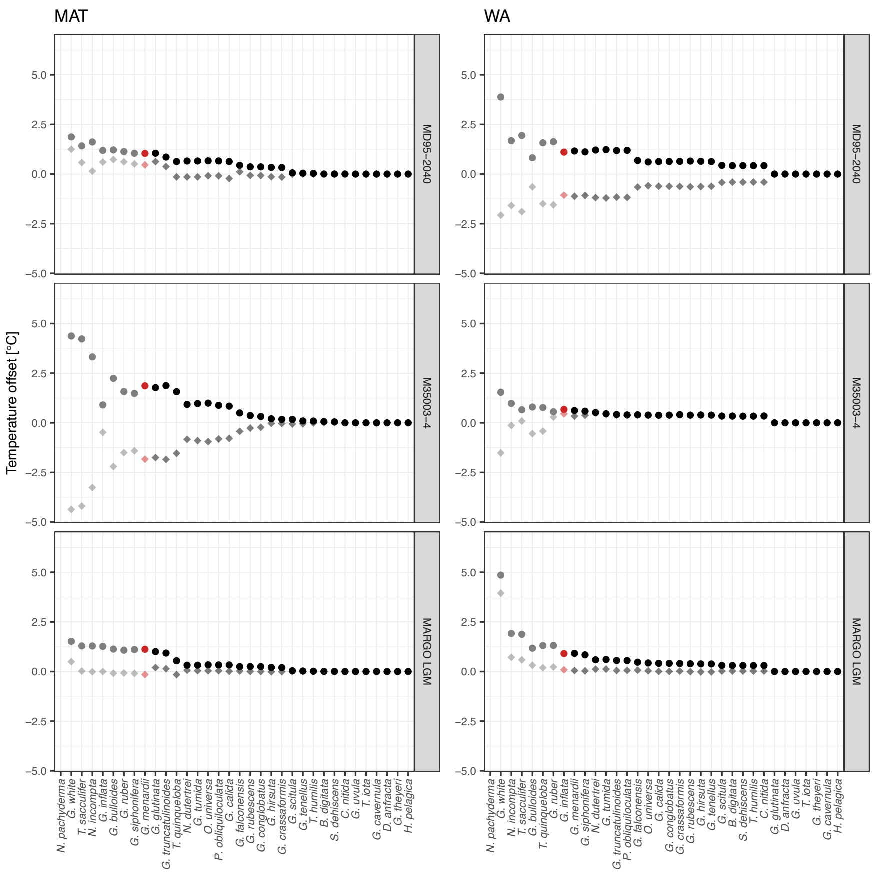

Figure 5: effect of species selection on SST reconstruction. Graphs show mean difference between reconstructions using species-pruned transfer function models and the model that includes all species. The results are ordered according to species importance, with each dot representing the result from a 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.

**Deleted:** 4

**Deleted:** stepwise

**Deleted:** transfer functions with the species up to the marked point

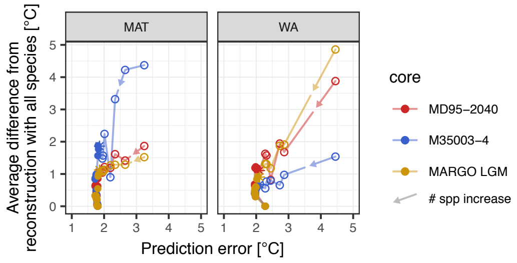

Figure 6: general effect of species pruning on reconstructions and on the prediction error. Each dot represents a reconstruction with n species and its associated prediction error; arrows indicate direction of increasing species numbers and decreasing species importance. The first point details a transfer function model with two species and one species is added at each subsequent point. Open symbols highlight reconstructions with the essential species of which the inclusion leads to a reduction in the prediction error. The reconstruction with minimum error and minimum number of species is marked by a star. All other reconstructions are plotted as dots. Note that for MAT and WA the inclusion of the essential species leads to a reduction of the prediction error, but that once these species are included there are still multiple reconstructions with the same error possible.

**Deleted:** 5

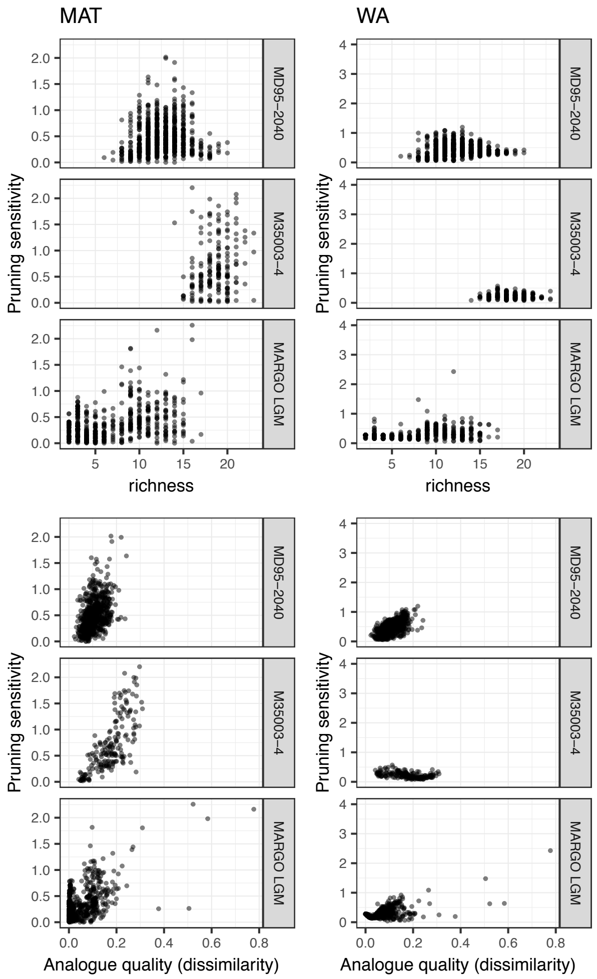

Figure 7: exploring the effect of species richness and analogue quality on the sensitivity to species pruning. This sensitivity is defined as the standard deviation of the reconstructions with more than the minimum required number of species (i.e. based on transfer function models with uninformative species, those to the right of the red dot in Fig. 4). Analogue quality is defined as the average dissimilarity from the five most similar samples in the training set; zero values thus mean perfect analogues. A clear effect of richness cannot be observed, yet for some of the downcore reconstructions (MD95-2040 and M35003-4) poor analogues appear associated with higher pruning sensitivity, particularly for MAT.

**Deleted:** 6

**Deleted:** different numbers of species, but similar prediction errors (i.e.

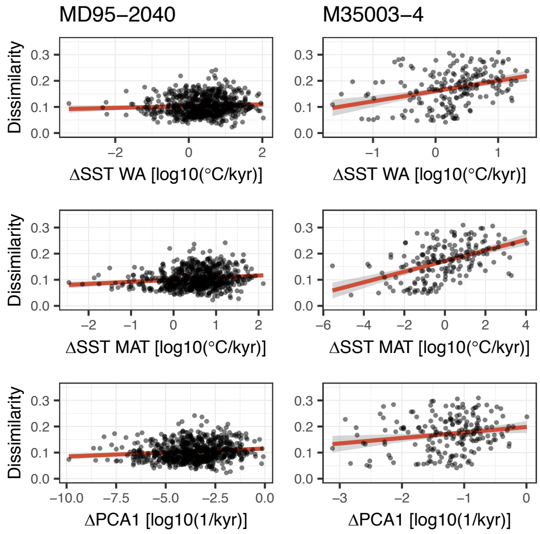

Figure 8: a possible role for sediment mixing in determining analogue quality (dissimilarity, see Fig. 6). Sediment mixing would create poor analogue assemblages when different species communities are mixed. The difference between subsequent species communities in the sediment core is here defined in two ways, i) as the difference in inferred temperature (top and middle row) and ii) as the first derivative of the loadings of the first axis of a PCA on the species abundances. Linear regression including 95 % confidence interval, is shown as a red line with grey shading. Only for core M35004-3 a weak positive relationship is visible, indicating that sediment mixing is a possible contributor to creating poor analogue assemblages, but cannot explain them entirely.

**Deleted:** 7

Table 1: Factors explaining species importance. Correlation coefficients of linear regression between species importance and maximum abundance, thermal niche width and temperature sensitivity.

| Method | Abundance | Temperature sensitivity | Niche width | Abundance + niche width | Abundance + niche width + temperature sensitivity |
|---|---|---|---|---|---|
| MAT | 0.79 | 0.66 | 0.21 | 0.83 | 0.84 |
| WA | 0.91 | 0.74 | 0.31 | 0.93 | 0.94 |

**References**

André, A., Quillévéré, F., Morard, R., Ujiié, Y., Escarguel, G., de Vargas, C., de Garidel-Thoron, T., and Douady, C. J.: SSU rDNA Divergence in Planktonic Foraminifera: Molecular Taxonomy and Biogeographic Implications, PLOS ONE, 9, e104641, 10.1371/journal.pone.0104641, 2014.

Bé, A., and Hutson, W.: Ecology of planktonic foraminifera and biogeographic patterns of life and fossil assemblages in the Indian Ocean, Micropaleontology, 23, 369-414, 1977.

Beaufort, L., and Dollfus, D.: Automatic recognition of coccoliths by dynamical neural networks, Marine Micropaleontology, 51, 57-73, https://doi.org/10.1016/j.marmicro.2003.09.003, 2004.

Berger, W. H.: Planktonic Foraminifera: selective solution and paleoclimatic interpretation, Deep Sea Research and Oceanographic Abstracts, 15, 31-43, https://doi.org/10.1016/0011-7471(68)90027-2, 1968.

Darling, K. F., Wade, C. M., Siccha, M., Trommer, G., Schulz, H., Abdolalipour, S., and Kurasawa, A.: Genetic diversity and ecology of the planktonic foraminifers Globigerina bulloides, Turborotalita quinqueloba and Neogloboquadrina pachyderma off the Oman margin during the late SW Monsoon, Marine Micropaleontology, 137, 64-77, https://doi.org/10.1016/j.marmicro.2017.10.006, 2017.

De Abreu, L., Shackleton, N. J., Schönfeld, J., Hall, M., and Chapman, M.: Millennial-scale oceanic climate variability off the Western Iberian margin during the last two glacial periods, Marine Geology, 196, 1-20, https://doi.org/10.1016/S0025-3227(03)00046-X, 2003.

Hayes, A., Kucera, M., Kallel, N., Sbaffi, L., and Rohling, E. J.: Glacial Mediterranean sea surface temperatures based on planktonic foraminiferal assemblages, Quaternary Science Reviews, 24, 999-1016, http://dx.doi.org/10.1016/j.quascirev.2004.02.018, 2005.

Hijmans, R. J.: raster: Geographic Data Analysis and Modeling. R package version 2.6-7. https://CRAN.R-project.org/package=raster. 2017.

Hsiang, A. Y., Elder, L. E., and Hull, P. M.: Towards a morphological metric of assemblage dynamics in the fossil record: a test case using planktonic foraminifera, Philosophical Transactions of the Royal Society B: Biological Sciences, 371, 10.1098/rstb.2015.0227, 2016.

Hüls, M., and Zahn, R.: Millennial-scale sea surface temperature variability in the western tropical North Atlantic from planktonic foraminiferal census counts, Paleoceanography, 15, 659-678, 10.1029/1999PA000462, 2000.

Hutson, W. H.: Transfer functions under no-analog conditions: Experiments with Indian Ocean planktonic Foraminifera, Quaternary Research, 8, 355-367, https://doi.org/10.1016/0033-5894(77)90077-1, 1977.

Imbrie, J., and Kipp, N. G.: A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Caribbean core, in: The late Cenozoic glacial ages, edited by: Turekian, K. K., Yale University Press, New Haven, 71-181, 1971.

Jonkers, L., and Kučera, M.: Global analysis of seasonality in the shell flux of extant planktonic Foraminifera, Biogeosciences, 12, 2207-2226, 10.5194/bg-12-2207-2015, 2015.

Juggins, S.: Quantitative reconstructions in palaeolimnology: new paradigm or sick science?, Quaternary Science Reviews, 64, 20-32, http://dx.doi.org/10.1016/j.quascirev.2012.12.014, 2013.

Juggins, S., Simpson, G. L., and Telford, R. J.: Taxon selection using statistical learning techniques to improve transfer function prediction, The Holocene, 25, 130-136, doi:10.1177/0959683614556388, 2015.

Kretschmer, K., Jonkers, L., Kucera, M., and Schulz, M.: Modeling seasonal and vertical habitats of planktonic foraminifera on a global scale, Biogeosciences, 15, 4405-4429, 10.5194/bg-15-4405-2018, 2018.

Kucera, M., Weinelt, M., Kiefer, T., Pflaumann, U., Hayes, A., Weinelt, M., Chen, M.-T., Mix, A. C., Barrows, T. T., Cortijo, E., Duprat, J., Juggins, S., and Waelbroeck, C.: Reconstruction of sea-surface temperatures from assemblages of planktonic foraminifera: multi-technique approach based on geographically constrained calibration data sets and its application to glacial Atlantic and Pacific Oceans, Quaternary Science Reviews, 24, 951-998, 2005.

Lombard, F., Labeyrie, L., Michel, E., Bopp, L., Cortijo, E., Retailleau, S., Howa, H., and Jorissen, F.: Modelling planktic foraminifer growth and distribution using an ecophysiological multi-species approach, Biogeosciences, 8, 853-873, 10.5194/bg-8-853-2011, 2011.

Malmgren, B. A., and Nordlund, U.: Application of artificial neural networks to paleoceanographic data, Palaeogeography, Palaeoclimatology, Palaeoecology, 136, 359-373, http://dx.doi.org/10.1016/S0031-0182(97)00031-X, 1997.

Morey, A. E., Mix, A. C., and Pisias, N. G.: Planktonic foraminiferal assemblages preserved in surface sediments correspond to multiple environment variables, Quaternary Science Reviews, 24, 925-950, http://dx.doi.org/10.1016/j.quascirev.2003.09.011, 2005.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Henry, M., Stevens, H., Szoecs, E., and Wagner, H.: vegan: Community Ecology Package. R package version 2.5-2. https://CRAN.R-project.org/package=vegan. 2018.

Rebotim, A., Voelker, A. H. L., Jonkers, L., Waniek, J. J., Meggers, H., Schiebel, R., Fraile, I., Schulz, M., and Kucera, M.: Factors controlling the depth habitat of planktonic foraminifera in the subtropical eastern North Atlantic, Biogeosciences, 14, 827-859, 10.5194/bg-14-827-2017, 2017.

Siccha, M., Trommer, G., Schulz, H., Hemleben, C., and Kucera, M.: Factors controlling the distribution of planktonic foraminifera in the Red Sea and implications for the development of transfer functions, Marine Micropaleontology, 72, 146-156, https://doi.org/10.1016/j.marmicro.2009.04.002, 2009.

Siccha, M., and Kucera, M.: ForCenS, a curated database of planktonic foraminifera census counts in marine surface sediment samples, Scientific Data, 4, 170109, 10.1038/sdata.2017.109, 2017.

Steinke, S., Yu, P.-S., Kucera, M., and Chen, M.-T.: No-analog planktonic foraminiferal faunas in the glacial southern South China Sea: Implications for the magnitude of glacial cooling in the western Pacific warm pool, Marine Micropaleontology, 66, 71-90, https://doi.org/10.1016/j.marmicro.2007.07.008, 2008.

Stephens, C., Antonov, J., Boyer, T., Conkright, M., Locarnini, R., O'Brien, T., and Garcia, H.: World Ocean Atlas 2001. Volume 1, Temperature, in: NOAA Atlas NESDIS 49, edited by: Levitus, S., U.S. Government Printing Office,, Washington DC, USA, 167, 2002.

Telford, R., and Birks, H.: Evaluation of transfer functions in spatially structured environments, Quaternary Science Reviews, 28, 1309-1316, 2009.

Telford, R. J., and Birks, H. J. B.: The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance, Quaternary Science Reviews, 24, 2173-2179, 2005.

Telford, R. J., Li, C., and Kucera, M.: Mismatch between the depth habitat of planktonic foraminifera and the calibration depth of SST transfer functions may bias reconstructions, Clim. Past, 9, 859-870, 10.5194/cp-9-859-2013, 2013.

Trachsel, M., and Telford, R. J.: Technical note: Estimating unbiased transfer-function performances in spatially structured environments, Clim. Past, 12, 1215-1223, 10.5194/cp-12-1215-2016, 2016.

Vernal, A. d., Henry, M., Matthiessen, J., Mudie, P. J., Rochon, A., Boessenkool, K. P., Eynaud, F., GrØsfjeld, K., Guiot, J., Hamel, D., Harland, R., Head, M. J., Kunz-Pirrung, M., Levac, E., Loucheur, V., Peyron, O., Pospelova, V., Radi, T., Turon, J.-L., and Voronina, E.: Dinoflagellate cyst assemblages as tracers of sea-surface conditions in the northern North Atlantic, Arctic and sub-Arctic seas: the new 'n = 677' data base and its application for quantitative palaeoceanographic reconstruction, Journal of Quaternary Science, 16, 681-698, doi:10.1002/jqs.659, 2001.

Wickham, H.: ggplot2: elegant graphics for data analysis, Springer, 2016.

Zielinski, U., Gersonde, R., Sieger, R., and Fütterer, D.: Quaternary surface water temperature estimations: Calibration of a diatom transfer function for the Southern Ocean, Paleoceanography, 13, 365-383, doi:10.1029/98PA01320, 1998.