Answer Comments Richard Telford - Clim. Past Discuss., 10, 1747-1782, 2014

Dear Richard Telford,

On behalf of all authors I would like to thank you for taking your time to review our manuscript and provide us with challenging comments. In the following we will give a detailed response to your comments and indicate how we will implement them in the revised version of the manuscript.

Sincerely,
Ines Hessler

Comment Richard Telford:
Hessler et al. compile a variety of SST proxies, which for unknown reasons they decide to call sensors, a word I would associate with instrumental data, but exclude diatoms on the grounds that there is not a harmonised data set. This is unfortunate as diatoms can reasonably be expected to be sensitive to SST rather than temperature at a greater depth, unlike several other proxies included.
Reply:
We agree that it is unfortunate that we couldn't include a harmonised diatom data set in our study. However, this is not because no such data sets exists but rather because we were unable to obtain the data from the related working groups despite our efforts.

Comment Richard Telford:
Foraminifera and dinocyst assemblages are used to reconstruct summer, winter and mean annual SST. This may be possible in parts of the ocean, but it is doubtful at high-latitudes where the vast majority of biological production occurs during the warm season. This problem is acknowledged deep in the discussion "The derived seasonal SST reconstructions are not independent but necessarily reflect the covariance among the seasonal SSTs in the modern ocean (Kucera et al., 2005a). This patently unlikely in the case of the MH."
Such concerns are likely to be overlooked by users of the compilation. It would be better to evaluate weather it is possible to make meaningful reconstructions of seasonal SST, and if not, omit them from the analysis.
Reply:
We think it is crucial to include seasonal analysis since the MH is characterised by an enhanced (reduced) seasonal cycle in the Northern Hemisphere (Southern Hemisphere). To this end we adopted the methods used in the MARGO project for the LGM and assigned the sensors to seasons accordingly. We realise that this involves a simplification but in the absence of a robust method to determine the seasonal bias in each proxy and how it may have changed through time we opted to analyse the seasonal values as they are and treat the uncertainty in the discussion. This may appear unsatisfactory, but we are not aware of a robust alternative and we believe that the uncertainty arising from this issue is not larger than that contained in the MARGO LGM compilation. Indeed, it is the message of our analysis that unlike the LGM with a large climatic signal, this

uncertainty is a significant factor hindering a robust early Holocene SST reconstruction.

Comment Richard Telford:
SST reconstructions from planktonic foraminifera are calculated using both the modern analogue technique (MAT) and artificial neural networks (ANN). ANN is described as permitting extrapolation beyond the range of parameters in the calibration set. I have not seen this issue explored.
Reply:
Unlike strict interpolation techniques such as MAT, Artificial Neural Networks with a back-propagation architecture have the ability to extrapolate beyond the range of the training values. This innate property of neural nets can be used in inverse optimisation problems (e.g. which combination of species would yield the highest SST or in real life, which stock portfolio would have given the highest yield). Fortunately, the extrapolation is usually quite limited, unless specifically desired and imposed on the network architecture. An example of a paper dealing with this issue is here:
http://www.sciencedirect.com/science/article/pii/S0022169400002286
The reason this sentence occurred in the methods section was to highlight the difference between MAT and ANN as two entirely independent approaches. To make this clear, we intend to re-word the corresponding sentence.

Comment Richard Telford:
A few cores in the compilation have results from multiple proxies, permitting a direct comparison between the proxies. Unfortunately the results are not consistent, even allowing for the uncertainty in the reconstructions. There are several possible reasons for this that are not fully explored in the paper.

1) Chronological control. The minimum number of radiocarbon dates or other stratigraphic markers is two. Presumably stricter criteria resulted in the exclusion of too many records. With such weak chronological constraints, errors of 1000 years or more can be expected in some cores. This would be a particular problem for the shorter MH window used. I would like the authors to give some consideration to this problem, but suspect the impact is relatively minor.
Reply:
We will address this issue in the revised version of the manuscript and include a short paragraph in the discussion chapter.
Although we applied several quality criteria for the selection of suitable records including a minimum requirement on the chronological control, differences in the SST pattern may be also related to chronological offsets between some cores. However, it is questionable how different the SST signal would actually have been even when considering a chronological error of 1000 years, considering the results of the simple exercise where we used two different definitions of the early Holocene time window. If, as we believe, the early Holocene SST signal was weak, then chronology alone is unlikely to explain the observed difference lest we have made such large error as to compare Holocene and Glacial sediments.

2) Palaeoclimatologists have adopted methods that allow them to report the lowest possible uncertainty on their reconstructions. These low uncertainties make the reconstruction appear good until different proxies are compared, as in Hessler et al., and the reconstructions are not consistent with each other. It is trivial to show that the root mean squared error of prediction of the transfer functions for the dinocysts and planktonic foraminifera is biased low, perhaps by a factor of two, because of spatial autocorrelation violating the assumption of independent between the test and calibration set during cross-validation (Telford and Birks 2005, QSR) and other problems. If the uncertainties were correctly shown, the results might appear less inconsistent.

Reply:

The RMSE of prediction derived by cross validation of the calibration dataset will always by a minimum estimate of an SST reconstruction error. The problem is in the case of a predictive (which is mathematically what proxies are) regression, it is very difficult to produce an upper bound on the uncertainty. We fully agree with the referee that the values we use for transfer functions are quite necessarily minimum estimates. The referee is right that if higher values were used instead, the inability to reconstruct the sign of SST change during the early Holocene would remain but the inconsistency among the different proxies would disappear.

3) Hessler et al. treat all the proxies as being sensitive to (seasonal) SST. However, some of the proxies are probably more sensitive to sub-surface temperatures. There is, for example, good evidence that the dominant foraminifera in the Nordic Seas live sub-surface and that the Holocene temperature trends at the surface and in the sub-surface are different (Andersson et al. 2010 COP). Even if the proxies were perfect, they would appear inconsistent if they are sensitive to different aspects of the water column. Note, if the biotic assemblages have been calibrated against a suboptimal depth/season, the uncertainty will have been inflated.

Reply:

We will address the point of different depth habits of the sensors in a separate chapter in the Matrial and Methods section (2.3) and discuss it in the Discussion.

2.3. Defining the 'sea surface'

The 'sea surface' and its related 'sea surface temperature' have been set to 10-m depth following the decision by MARGO (Kucera et al. 2005a). This decision reflects a compromise allowing a harmonisation of SST estimates among different sensors. This choice does not mean that the authors assumed that all sensors record temperature at that depth. Rather, the decision reflects an assumption that all sensors and proxies record an SST signal which is highly correlated to SST at 10-m depth and that it is therefore possible to calibrate the individual proxies against SST at that depth. In the context of this study where the focus lies on SST anomalies, the principle assumptions of this depth-homogenisation are thus that the SST recorded by each proxy and sensor is highly correlated to SST at 10-m depth and that this relationship remained the same between the present-day and the 6k Holocene time slice. Whereas the SST depth recorded by phytoplankton sensors is limited to the photic zone, the depth range of species of planktonic foraminifera can be broader. The foraminifera-based Mg/Ca SST estimates are based chiefly on symbiont-bearing species with shallow habitat, whose calcification depth has been

constrained to lie within the top 100 m of the water column (e.g., Anand et al., 2003; Regenberg et al., 2009). In contrast, the foraminifera-based transfer function SST are based on analysis of the entire assemblage and as shown by Telford et al. (2013), it is possible that assemblage composition is sensitive to subsurface temperature, particularly in low-latitude regions. This depth mismatch may be significant when reconstructing temperature of the last glacial maximum, but it remains unclear whether it also has an effect on early Holocene SST estimates. Thus, in the absence of a universally applicable set of criteria for assigning depth to SST estimates by different proxies and sensors, we retained the 10-m depth definition used by MARGO, but we acknowledge that depth-misattribution of the reconstructed SST may be an additional source of uncertainty and may account for mismatch among SST proxies, particularly those based on planktonic foraminifera as a sensor.

As indicated in chapter 2.3 (Defining the 'sea surface') the SST pattern reconstructed in this study is also likely biased by sensitivity of planktonic foraminifera assemblages to temperatures at different depths in the water column, as well as by changes in the SST sensitivity or recording depth of the other sensors and proxies between the present-day and the early Holocene. The former is likely to be more significant, because the recording depth of all other sensors and proxies used in this compilation is bound to have remained within the photic zone.

4) The great mismatch between alkenones and other proxies suggests there may be undiagnosed biases in one of more of the proxies.

Reply:

This is indeed the main conclusion of the discussion. If all proxies recorded SST of the season and depth to which they are ascribed, there would not be such large mismatch. Clearly, the diverging results cannot all be right so wither some of the proxies must be recording SST at a different season and depth or not record SST at all. An interesting approach to exploring the origin of the mismatch is presented by Lohmann et al. (2013), who conclude that in many but not all of the records, proxy mismatch is within the range of plausible range of SST variation within the habitat of the sensors.

Comment Richard Telford:

Hessler et al. use a significance test to determine how many records have SST anomalies significantly different from zero. My understanding is that records where the absolute mean anomaly (from cores where there are three or more analyses in the MH window) is greater than twice the standard error of the analyses plus the uncertainty on the reconstruction are deemed significant. This is a non-standard test and its behaviour is not explored. It is equivalent to a t-test for large samples (if the proxy uncertainty is ignored) as the t-distribution resembles a Gaussian distribution in this case, but not when the sample is small as the t-distribution has heavy tails when the sample is small. One critical aspect is whether the reconstruction errors are independent. If the errors are not independent, such that if one analysis is too warm all will be too warm, this test will be too liberal. Conversely, if the errors are independent the test will be too strict. My feeling is that the degree to which the error is independent will be

method specific. Methods where analytical error or aliasing of the annual signal are high will have more independent errors. If the errors are independent, a one-sample t-test can be used – the reconstruction uncertainty is already accounted for in the estimate of the variance. If the errors are not independent, I think a modified t-test can be used, adding the reconstruction uncertainty in quadrature to the standard error. The reconstruction uncertainty could be scaled according to the degree it is thought to be independent. I do not know what effect this would have on the results.

Reply:

This is a very interesting issue. We agree that the way we "diagnose" which anomalies are significant is rather simple and ignores the statistical properties of the variable. These are, alas, bound to be rather complex and heterogeneous. However, we believe the method we use is conservative in the sense of the conclusion we draw – allowing for longer tails in the distribution would likely diagnose more anomalies as not significant, but we were bound to conclude that too few anomalies were significant already with the simple test based on the Gaussian assumption. We will highlight this point at the end of chapter 3.4.1 (Assessment of significance of reconstructed changes in sea surface temperature) in the revised manuscript.

Although we assume all uncertainties are independent a certain level of dependency may exist nonetheless. However, considering the various uncertainties to be dependent would lead to the t-test identifying even fewer records as being significant.

Comment Richard Telford:

Figures S1-S5 are missing from the supplementary material.

Reply:

The figures presented in the Supplementary Material went missing in the course of the first re-submission of the manuscript, which is a sloppy mistake. We will correct it in the revised version of the manuscript and make sure that all figures will be available as indicated in the descriptive paragraph at the beginning of the Supplementary Material.