Reply to anonymous reviewer comments, 29.2.2014, concerning the manuscript of Rehfeld, Trachsel, Telford & Laepple in discussion for Climate of the Past (doi:10.5194/cp-2016-13)

04/15/16

Summary

We would like to thank the reviewer for her/his insightful remarks, which will help us to improve a revised manuscript.

Reviewer's comments are given in grey. Emphasis in *italics* was added to highlight main points.

Point 1: Combination of climate parameters governing vegetation composition in the past

The paper by Rehfeld et al. deals with the pollen-based climate reconstructions. The authors use climate model data and modelled vegetation to explore the reliability of reconstructions of different climate parameters in pollen-based reconstructions. The advantage in such an experiment "in an ideal model world" is that the past climate and vegetation are known at all times (6 ka to present), allowing to assess the reliability of the reconstructions. *The authors show that reconstructing multiple climate parameters can be misleading, as it is possible that in reality there is only one climate parameter which drives the spatial and temporal vegetation change, and the reconstructions of other climate parameters show temporal variability which is caused by the fact that these less important parameters are spatially correlated with the important parameter in the modern spatial data used for constructing the transfer function.* This is certainly nothing new, most of the palaeoecologists using pollen data have been aware of this problem, but it is useful to have a special study where this problem in explicitly explored using novel approaches.

I find it easy to agree with the authors that "the temporal changes of a dominant climate variable are imprinted on a less important variable, leading to reconstructions biased towards the dominant variable's trend" and that the high r2 in the cross-validation is of limited use to identify which variables can be reconstructed, as r2 can be high not only for the variable which is really important for vegetation or pollen, but also to non-important variable which are spatially correlated with the important variable. The authors suggest assessing the amount of fossil vegetation variance explained the reconstruction output and expert knowledge as possible means to select the climate variables. *The latter one has been used in pollen-based reconstructions, but unfortunately the expert knowledge almost invariably is limited to present ecological setting.* It is possible, or even likely, that *if we go back in time enough, the combination of climate parameters governing the vegetation composition have been fundamentally different from the present.*

Non-analogue combinations of climate and other physiologically relevant variables certainly occurred in the past and will result in less accurate reconstructions. For example, the effect of low-CO2 concentrations on LGM pollen-based reconstructions is difficult to quantify and is hence largely ignored. Climatic conditions in the Holocene, the period analysed in this study are unlikely to have been sufficiently non-analogous to cause serious problems, and non-equilibrium vegetation may be more of a general problem.

Point 2: Number of PFTs

There is one striking problem with the paper. Given that the authors use model data only, they are restricted to use plant functional types (pft), not pollen types or plant species. In the real world, the WA-based climate reconstructions often comprise over 100 pollen types, not pfts. Modern analogue-based reconstructions use pfts, but even in them the number of pfts is generally 20-30. In a striking contrast, the number of pfts in the current study is eight - in other words extremely low. I am surprised that the palaoeclimate reconstructions with such a low number of variables make any sense in the first place, given that they are based on a few, extremely broad pft classes.

We thank the reviewer for raising this point. We agree that the number of Plant Functional Types (PFTs) in our model study is lower than what is generally used in a real-world large-scale reconstruction exercise (eg. in Davis et al., 2003, Mauri et al., 2014, Mauri et al., 2015). We use 9 PFTs (one of which is representing bare soil, or desert fraction), whereas e.g. Mauri et al. (2014) use 22 (two of which are virtual). However, as pointed out in the response to Reviewer 1, what is ultimately relevant for the calibration and reconstruction efforts is the information contributed by the PFTs or taxa; in other words how many of them actually contribute to the pollen (or PFT) diagram in a relevant way. This number is much lower than the number of PFTs, or the number of taxa in a pollen diagram, but as we outline below, this number is comparable between our modeled PFTs and real-world pollen diagrams.

The median palynological richness in the 4990 sites in the European Pollen Database is 31. However, the median effective number of species (given by Hill's N2 (Hill, 1973)), discounting very rare taxa which would not be relevant for reconstructions, is 4.9. Of the 458 grid points we use, the median N2 for the fossil data (shown in Fig. 1d in the manuscript) is 2.9, and for the modern calibration data it is 2.7. Therefore, although the number of PFTs is much

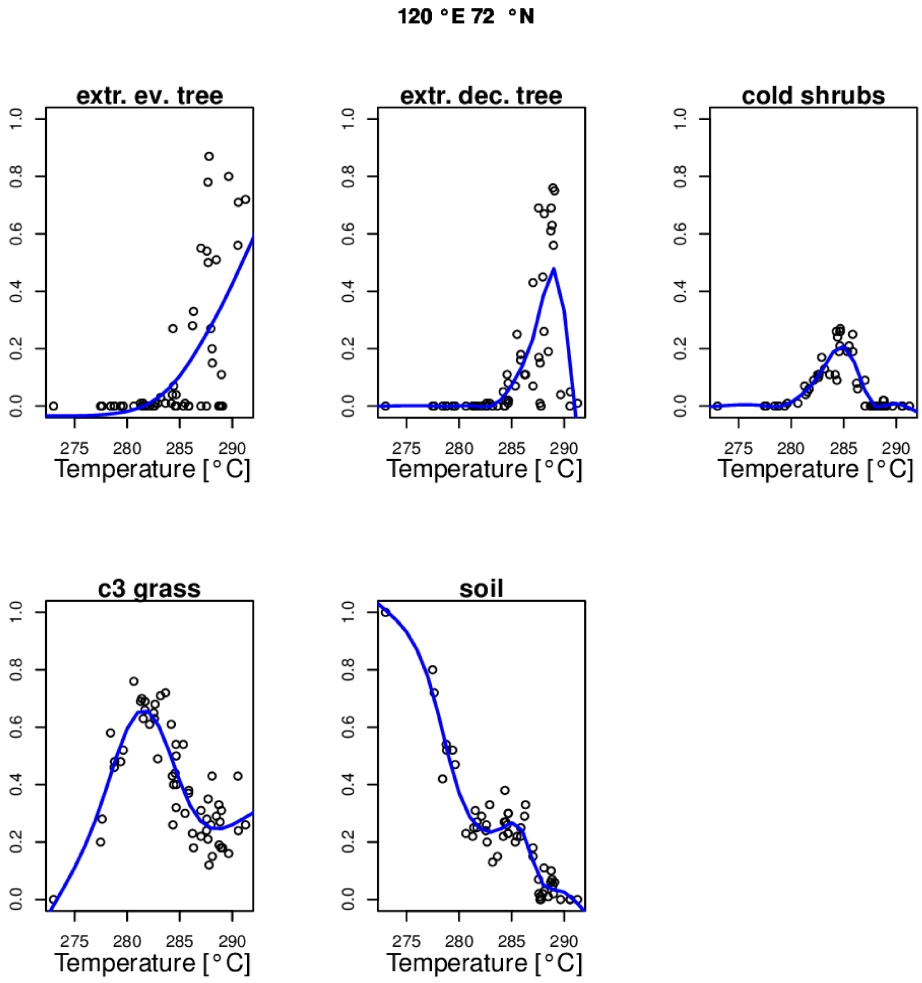
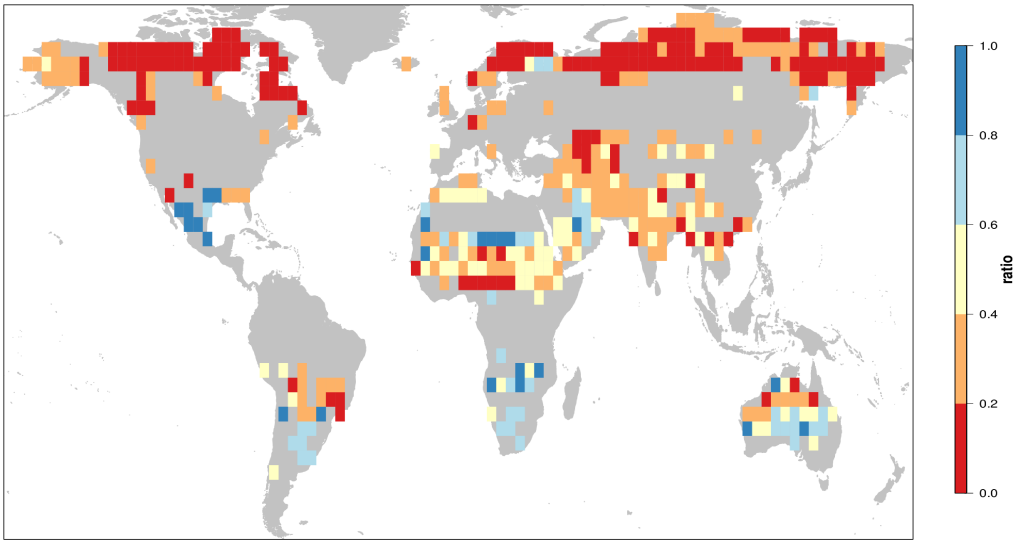


Figure R1: Species response curves for the calibration radius around example site 120E,72N. Only non-zero taxa are shown.

Figure R2: Ratio of the standard deviation of the MTWA climate variable at modern analog sites over the standard deviation within the training set.



lower in the model, the diversity and effective number of species is not much lower than that many actual pollen-climate reconstructions. Moreover, we are setting bounds on N2 and turnover to avoid pathological problems due to too few taxa. This is an important aspect we will discuss in more detail than before in a revised manuscript.

Point 3: Multiple Analogues

I suspect that the reconstructions using modern analogue must have included some serious problems which are not reported in the paper. The *problem of multiple analogues* (where the many modern analogues for the fossil sample are present, often in very different climatic settings) would be unavoidable with eight pdfs only.

We thank the reviewer for raising this important point. The multiple analog problem could arise if the species response curve (e.g. with respect to the climate variable, e.g. MTWA) within a modern calibration radius was multimodal. However, analyzing the species response curves at several sites suggests that this is not the case. As an example, Fig. R3 shows the species response curves for all taxa effectively present in the Siberian site for which we also show the complete reconstruction workflow in the manuscript (Fig. 2). The species response curves are not multimodal.

Furthermore, the overall high transfer function r2 (Fig. 7 in the manuscript) shows, that analogs are not picked at random from the training set, and underlines that multiple analogs are not a problem . To exemplify this we calculate the ratio of the standard deviations of the temperatures at the analog sites, and the standard deviation of the temperatures across the whole training sets (Fig. R4) . The ratios are generally smaller than 0.5, thus illustrating that the analog sites are not randomly drawn from the training set.

In the revised version of the manuscript we will explicitly demonstrate that arbitrary analogs are not a problem here, by including the above discussion.

Point 4: Error estimates and uncertainties

The *error estimates of the calibration sets and the fossil reconstructions are not presented or discussed in the paper, but they most likely are extremely high*. I therefore wonder if the *difference in the reconstructions (in Fig. 8 and 9, for example) are inside or outside the error*

estimates?

For the sake of brevity we have not given explicit plots of all quality metrics within the manuscript. We do, however, give standard error estimates (RMSEP, cross-validation r2) for the example grid point reconstruction workflow (Fig. 2 in the manuscript), and for all grid points >50N (Table 1). The calibration r2 for all climate variables is given in Fig. 7 in the manuscript, and Fig. 7 in the supplement compares the different error estimates for downcore RMSE and RMSEP for warmest month temperature. While we have given the summary numbers for the RMSEP and the r2 in Table 1, we agree that it would be helpful to show these statistics explicitly for each gridpoint, projected onto a map. We show the results for the temperature variables below in Fig. R5. As suspected by the reviewer, the RMSEP, particularly for MTCO, is not small, however, it is not much larger than that in real-world large-scale reconstructions (c.p. Frechette et al., 2008; Mauri et al., 2014). The regions with high RMSEP are largely consistent with the regions where the calibrations' r2 is low (Fig. 7 in the manuscript). We will improve Figs. 4 and 5 in the manuscript by masking grid points with an r2 below 0.5, as the results show the covariance across time and space of winter and summer temperature reconstructions, and could be affected by this. Furthermore we plan to incorporate Fig. R5 in the supplement of the revised manuscript, and we will also better explain and more prominently discuss Table 1 in the revised manuscript.

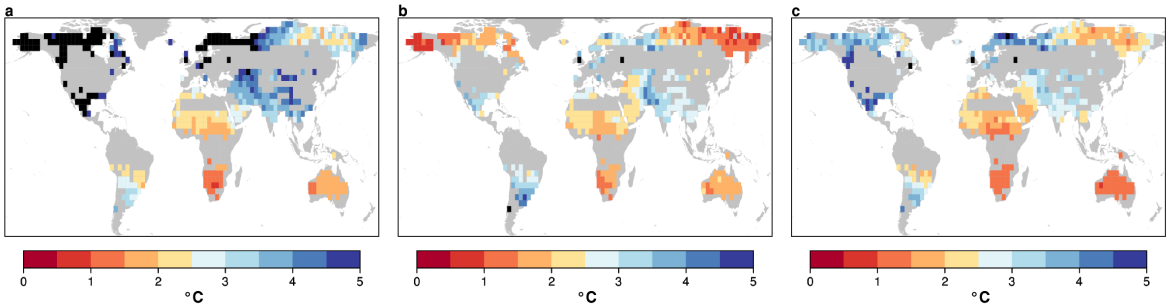


Figure R3: RMSEP for 10-fold-leave-group-out cross validation of MTCO (a), MTWA (b) and MAT (c). Black gridcells indicate a RMSEP larger than 5C.

The question whether or not the temperature changes against time shown in Fig. 8 and 9 in the manuscript are significant or not using the calibration RMSEP is not straightforward. A standard assumption in paleoclimate reconstructions is that errors in time and space are independent (as assumed e.g. in Marcott et al. 2013, Fedorov et al., 2013, Shakun et al., 2012). This assumption would result in a standard error of 0.008C (1se) for the difference of -0.72C between the calibration at 0k vs. 6k in Fig. 8 (taken across all 27 gridpoints >70N and 197 time points). For Fig. 9 the difference at the 0k is 0.32C, at 6k it is .44C and the standard error at these time points is smaller than 0.13C. Both differences in the reconstructions would therefore be highly significant. On the other hand, there are good reasons to expect spatial and temporal correlations in the errors. In the (unrealistic) extreme case of a complete dependency or errors, the differences would be not significant. In reality the true uncertainty likely lies between the two extremes assumed here but a mechanistic understanding of processes causing the proxy uncertainty is required to provide better error estimates. In the revised manuscript, will discuss these aspects of uncertainty in Section 4 and include the uncertainty estimates for both cases.