# Author response to reviewer and editor comments

### Manuscript: “Assessing performance and seasonal bias of pollen-based climate reconstructions in a perfect model world” by K. Rehfeld, M. Trachsel, R.J. Telford and T. Laepple (doi:10.5194/cp-2016-13)

05/17/16

## Response to editor comments

The main criticism, expressed by both reviewers, concerning our work was that the number of Plant Functional Types in our model world is much lower (9) than in comparable real-world reconstructions based on PFTs (22, e.g. Mauri et al., 2014), or on pollen directly (where sometimes more than 4 times as many taxa are distinguished). In our response we argue that, for the purpose of climate reconstructions, the number of taxa present in relevant proportions is more important than the overall number of taxa distinguished, and that our efficient number of taxa was not unusually low. We illustrated this by showing the difference between the median effective number of species (4.9), and the median palynological richness (31) in the  European modern Pollen database. We added a new paragraph to the main text and included a new supplementary figure (SFig. 1) to reflect these findings.

The editor suggested that this point could be further discussed and extended as it might imply that field work and reconstructions could be done more effectively. We performed own preliminary analysis of European Holocene pollen stratigraphies, which confirm that reconstructions obtained from greatly reduced datasets are highly similar to full reconstructions. The picture may, however, be different in different climatological and ecological settings, e.g. in the Tropics. Furthermore, the presence of rare taxa may play a very important role in paleoecological studies, which, to many researchers performing the field studies, are more relevant than paleoclimate reconstructions. Therefore, we would like to refrain from extending this disucssion in the manuscript, as it would deviate from our key point, the different relationships across space and time in vegetation and climate and their effects on the reconstructions. Having said this, we agree with the editor that this is an interesting topic for a further study.

# Summary of changes

In the revised version of the manuscript we have addressed all points made by the reviewers. We have added several new paragraphs and four new figures in the supplementary material. We improved the linkage between the main text and the figures and tables in the main text and the supplementary. We furthermore revised two figures, masking out grid points which showed a low transfer function performance, and rephrased statements throughout the manuscript. Finally, we corrected initials in several citations.

We give our previous response to the two anonymous reviews below. At the bottom of each key point made by the reviewers, and below our response, we outline in **dark green** what changes we have made in the manuscript. Line and page numbers refer to the attached document with highlighted changes.

# Response to anonymous reviewer #1

### Summary

We would like to thank the reviewer for his/her detailed comments which will help to improve the clarity and quality of the manuscript.

The reviewer mainly comments on limitations of our study related to the coarse representation of the simulated vegetation and the simulated climate. These are valid points which we are aware of. We will ensure that they are more properly represented in the revised manuscript. However, as we demonstrate below, the reviewer's points are overstated and the limitations brought up by the reviewer do not affect our main conclusions.

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

**Point 1: Number of taxa used in the study**

> **Reviewer's comments:**
> The authors use a vegetation model and climate model to simulate the process of reconstructing climate from pollen data, and in turn to assess the ability of pollen-based methods to accurately reconstruct seasonal Holocene climate change.
> This is a interesting and novel approach, and although similar virtual experiments have been conducted with other proxies, this is the first time that I know of where it has been applied to pollen. Pollen-based climate reconstructions have been widely used in data-

> model comparisons, and large discrepancies have been found between these reconstructions and climate model simulations during the Holocene, particularly in terms of seasonality. Investigation of potential seasonal bias in pollen-based reconstructions is therefore of particular interest and importance.
>
> *The study is generally well written and presented, but has a number of critical issues that I do not think can be easily resolved. The most obvious of these is the unrealistically low number of virtual taxa, or in this case PFTs, used in the transfer-function.* To some extent the authors themselves acknowledge this (lines 550-554) “Furthermore, the methods we have tested are limited by the low number of plant functional types, as large-scale PFT-based pollen reconstructions use roughly 2-3 times the number of PFTs (as e.g. in Davis et al., 2003; Mauri et al., 2014).” *This is actually an underestimate, since of the 8 PFT’s used by Rehfeld et al, 3 are tropical (tropical evergreen trees, tropical deciduous trees and C4 grasses), leaving just 5 PFT’s for the extratropics such as Europe.* Davis et al. 2003 and Mauri et al. 2014 use 22 PFT’s for Europe, which is more than 4 times the number used by the authors in their study. *Taxa based pollen-climate transfer functions commonly use upwards of 50-60 taxa.* These numbers are important because the individual behavior of the PFTs/Taxa and their climatic tolerances constitute the degrees of freedom necessary to reconstruct multiple climatic variables, and particularly those that may show close co-variance as cited as a potential problem by the authors.

As we state in the manuscript, 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). In a given region, the number of contributing PFTs is lower, as some PFTs only appear in some region, thus leading the reviewer to conclude that the difference between our modeled and the real world PFTs is up to factor of four.

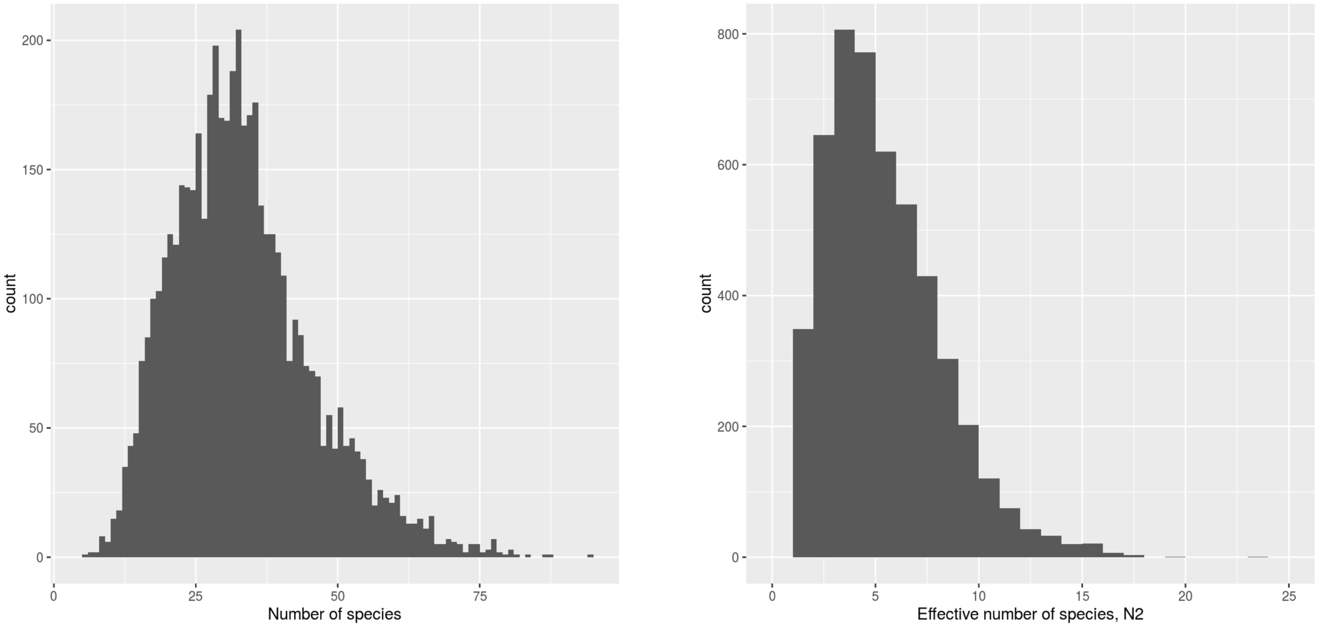
However, what is ultimately relevant for the calibration and reconstruction efforts is the information contributed by the PFTs or taxa; in other words how many of them actually contribute to the pollen (or PFT) diagram in a relevant way. This strongly differs from the number of PFTs/ the number of taxa.

The effective number of species can be quantified by the Hill's number N2, which is an entropy-based measure for the vegetation diversity (Hill, 1973). An analysis of 4990 sites in the European modern pollen database, which formed the basis of Mauri et al. (2014, 2015) and many other large-scale pollen-based reconstructions shows that the median effective number of species (N2) is 4.9, much lower than the median palynological richness of 31 taxa (Fig R1 below) . If the pollen data of the European Pollen Database were assigned to PFTs we would expect the N2 for the PFTs to be even lower.

In our study we only use sites (grid points) with a Hill's number larger than 2 (p.4, l. 19). 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 lower in the model, the diversity and effective number of species is not much lower than that in actual pollen-climate reconstructions.

In the revised version of the manuscript, we will discuss the differing number of taxa as well as the difference in the effective vegetation diversity.

*Figure R1: Distribution of the number of species (left) and effective number of species (N2) in modern samples in the European Pollen Database.*



**Implemented changes in the revised manuscript:**

**- inserted new discussion and definition of Hill's N2 on p. 5 l. 19.**

**- main text paragraph p.5 l. 30 – p. 6. l. 10 on N2 vs. richness**

**- new Supplementary Figure SFig. 2**

**Point 2: Winter vs. summer temperature reconstructability**

> Furthermore, the PFT’s used in the study by Rehfeld et al. have extremely broad climatic tolerances (deciduous trees, evergreen trees, grass..) that can be expected to have little diagnostic power. No pollen-climate transfer function should or would be based on such a low number of taxa/PFT’s with such broad climatic sensitivity, and it is therefore disingenuous of the authors to compare their own over-simplified approach with the approach used in actual pollen-climate reconstructions. For instance the authors infer that because they were unable to reliably reconstruct winter temperatures, this should also be a problem for actual pollen-climate reconstructions. In reality, the problem with winter temperatures is just as likely to be a result of the authors over-simplified experimental design and the use of a limited number of PFT’s with limited winter temperature sensitivity.

In our study we have followed the standard workflow for pollen-based reconstructions. The reconstructability, or non-reconstructability, of climate variables is often inferred from the transfer function r2 and the RMSEP in cross-validation (e.g. in Mauri et al., 2014, Frechette et al., 2008). These test diagnostics are based on the modern calibration data alone. As we show in Fig. 7 in the manuscript, the transfer function estimated r2 for winter temperatures is similar to that for other temperature variables. Therefore, the transfer function diagnostics suggest, that winter temperature is reconstructible. We agree with the reviewer that the true reason for low *actual* reconstructability of winter temperatures may well be that winter temperatures have little influence on the modeled vegetation (which might be realistic in at least in some regions of the world, such as Siberia). However, even if the winter sensitivity of the model vegetation were unrealistically low, this would only strengthen our conclusion that transfer function diagnostics based on modern calibration data alone are not sufficient to characterize reconstructability.

**Implemented changes in the revised manuscript:**

- **new statement on resolution (p. 4 l. 6 – 7)**
- **masking of grid points with low transfer function performance in Figs. 4 & 5**
- **discussion of high RMSEP for winter temperature (p. 10 l. 7-10)**
- **new  Supplementary Figure SFig. 1**

**Point 3: Resolution of the calibration climate dataset**

> This problem is likely to be compounded by the use of climate data for calibration from a climate model with low spatial resolution, and where the spatial variability of climate is highly smoothed compared to the real world. On the one hand this reduces the variance of climate and vegetation in the training set and on the other, it greatly increases the propensity for spatial auto-correlation that the authors also highlight as a problem in their study.

We of course agree that the spatial climate fields from our climate model simulations have a much lower resolution than for example the 0.5 minute resolution interpolated instrumental dataset (Hijmans et al. 2005) used in Mauri et al. (2015).
However, the resolution in itself is neither the determining factor for the variance of the climate explanatory variable, nor for the spatial autocorrelation. Most importantly, the covariance structure between the different climate variables (e.g. summer and winter temperatures), is not directly a function of the resolution. Given the large-scale structure of spatial climate and especially temperature variations (e.g. Hansen, & Lebedeff, 1987) we do not expect a strong influence of the resolution, except in some areas where elevational gradients, not well represented in the coarse model topography.

This is demonstrated in the Figure R2 below, which compares the distribution and relation of the gridpoint winter and summer temperatures in the 0.5 minute resolution temperature field (Hijmans et al. 2005) with the gridpoint temperatures from our low resolution ECHAM5-MPIOM simulation in Europe (30-60N, 0-30E) including several mountain ranges. The 61 land grid-points from the climate model cover most of the phase-space spanned by the 13 million grid points of the 0.5min resolution field, except high-altitude regions represented by the lower-left tail. The model field further shows a similar correlation between the seasons. Thus we see no reason to expect that the low resolution would bias our results towards less skill in reconstructing multiple variables. We will include a detailed discussion in the revised manuscript.

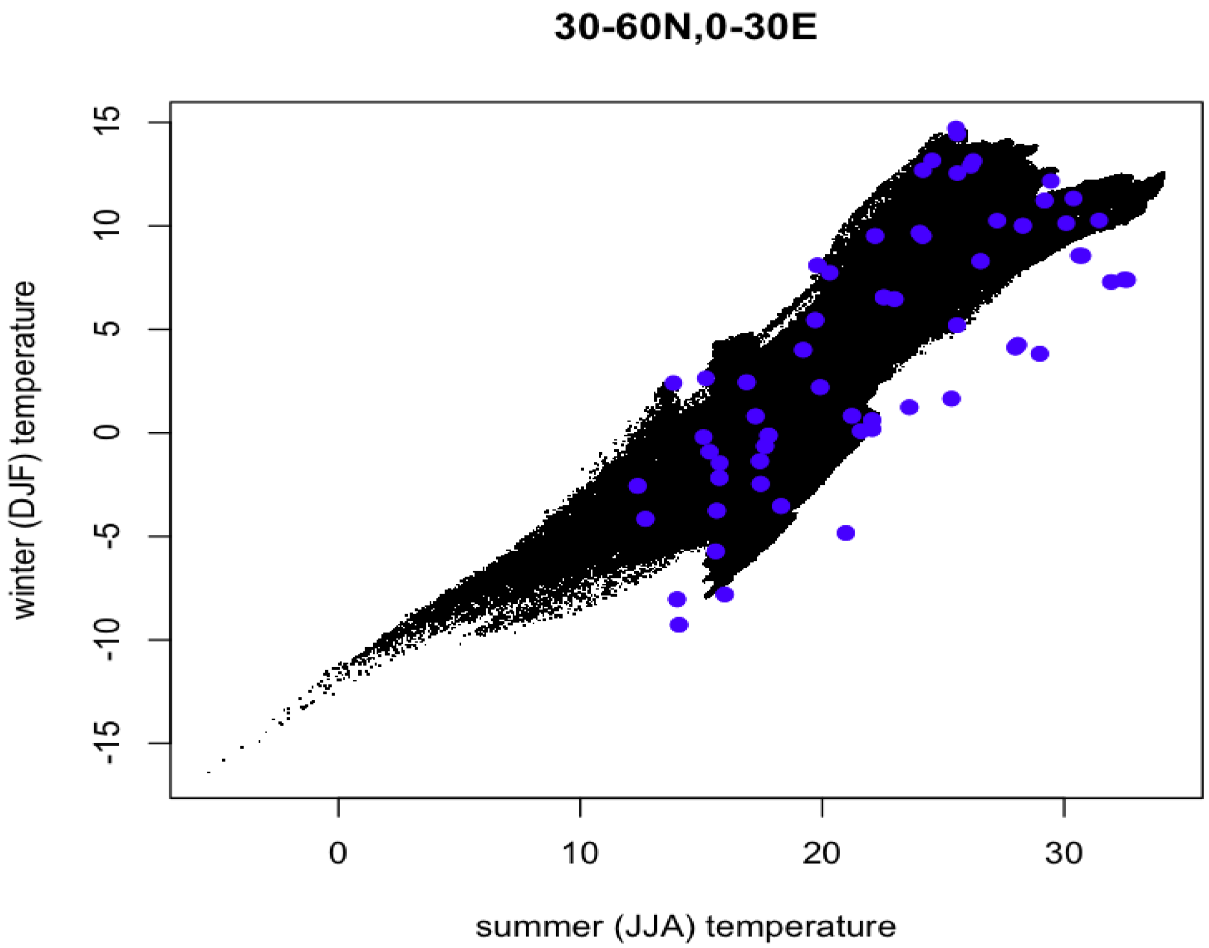


*Figure R2: Distribution and relation of gridpoint winter and summer temperatures in the 0.5'-resolution temperature field of (Hijmans et al., 2005) and the gridpoint temperatures of the ECHAM5-MPIOM simulation (Fischer & Jungclaus, 2011) used in this study.*

Spatial autocorrelation is often not considered in papers reconstructing climate from pollen (e.g. in Bartlein et al., 2011), and will tend to be a larger problem in the densely sampled pollen databases than in our low resolution data, as each pollen site has many geographically close neighbors which can be used as an analogue in the modern analogue technique.

**Implemented changes in the revised manuscript:**

- **new statement on resolution (p. 4 l. 6 – 7)**
- **new  Supplementary Figure SFig. 1**

**Point 4: Simplification**

> *Whilst some simplification should be expected in a 'virtual' study like this, it is important not to over-simplify to the point where the study itself is so far removed from any actual application that the results are not comparable.* The problem here is that the authors consistently conflate their results with those from actual pollen-climate reconstructions (as in the title), and therefore are at risk of presenting a fallacious argument that the average reader who is not so familiar with the topic will likely interpret at face value.

We fully agree that the complexity of the vegetation representation in the model as well as the simulated climate evolution are a strong simplification of the reality. Therefore, results on the Holocene evolution of specific PFT's, the actual spatial pattern of PFT's, or the reconstructability of a certain climate variable in a certain region should not be directly translated  to the real-world.

On the other hand, conclusions about reconstruction methods and the relation of spatial calibration and downcore reconstruction only require a consistent dataset of climate and vegetation parameters in space and time and do not depend on details of the climate evolution or vegetation response, as long as the dataset is realistic enough that we can apply the real world reconstruction workflow. The major factor shaping these results is that the modern spatial relationships between climate variables is different from the changes in the relationships over time, which is a robust feature related to the transient insolation forcing.

In the revision, we will check in detail again if all our statements are either independent from the model-world specifics , or are clearly marked that they just apply to the model world. Furthermore, we will emphasize the limitations of our study further by extending paragraph 4.1, and by highlighting their impact in the conclusion paragraphs. We do, however, find the title of our manuscript appropriate, as it clearly expresses that we work with pollen-related methods in an ideal model world.

**Implemented changes in the revised manuscript:**

- **Throughout the manuscript, in particular in paragraph 4.1 on limitations and the Conclusions  we have evaluated and highlighted in a better way, which statements pertain to the model world experiment here, and which can be generalized to real-world pollen-based reconstructions.**

**Point 5: Repeating the study with a more elaborate vegetation model (LPJGuess)**

> The subject of the paper is nevertheless interesting, and one that would otherwise be worthy of publication. I would therefore encourage the authors to collaborate with someone who has more experience in pollen-climate modeling, and to use a vegetation model such as LPJGUESS which can simulate a greater number of PFT's/Taxa so that the analysis can be more comparable with how pollen-climate transfer functions are actually applied.

As we have shown above, our analysis is comparable to actual pollen-climate transfer functions, as the effective number of species present in the model data is within the range of the numbers observed for the European Pollen Database, which formed the basis e.g. for (Mauri et al., 2014; Mauri et al., 2015).

We agree that using a more complex vegetation model, such as LPJGUESS, would be worthwhile in a future study as this would not only increase the number of PFT's but also include a more realistic interaction between climate and vegetation. This would allow to also test other properies of the reconstruction e.g. effects on the time-scale dependent variability, or potential time-lags. We will thus include this proposal in the outlook of our study.

However, we expect that the main result of this manuscript, the limitations of spatial modern calibrations, would only be strengthened. This is because a more realistic climate-vegetation response will  even differ more in time versus space, if for example the modern vegetation is not yet in equilibrium with the modern climate state.

**Implemented changes in the revised manuscript:**

- **We state that it would be worthwile to repeat the study with a vegetation model with more realistic vegetation and more PFTs (p. 18 l. 16-25)**
- **We emphasize the limitations of our approach (p. 15 l. 2- p. 16 l. 17)**

## Response to anonymous reviewer #2

### 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.

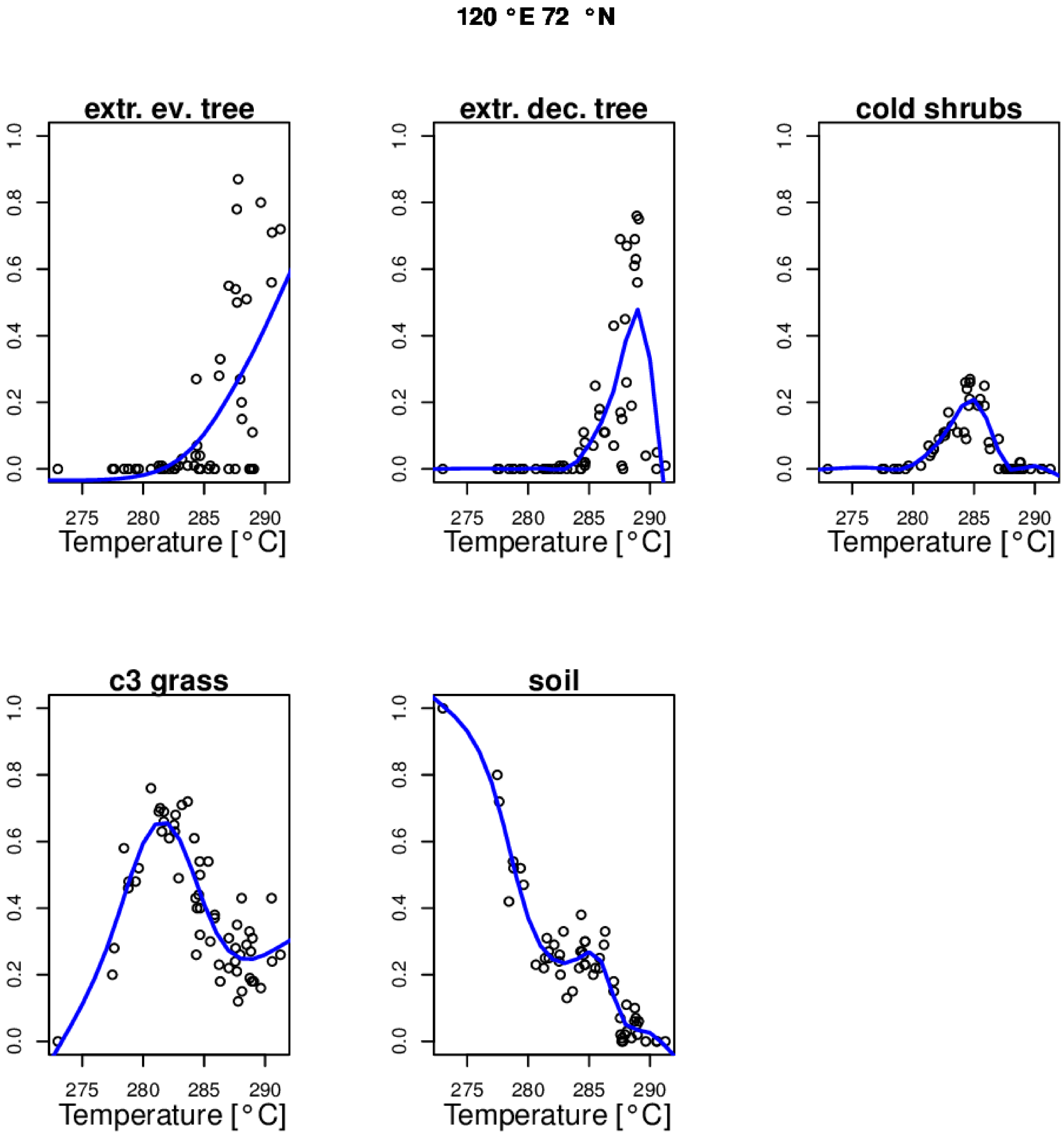


*Figure R3: Species response curves for the calibration radius around example site 120E,72N. Only non-zero taxa are shown.*

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 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.

**Implemented changes in the revised manuscript:**

- **inserted new discussion and definition of Hill's N2 on p. 5 l. 19.**

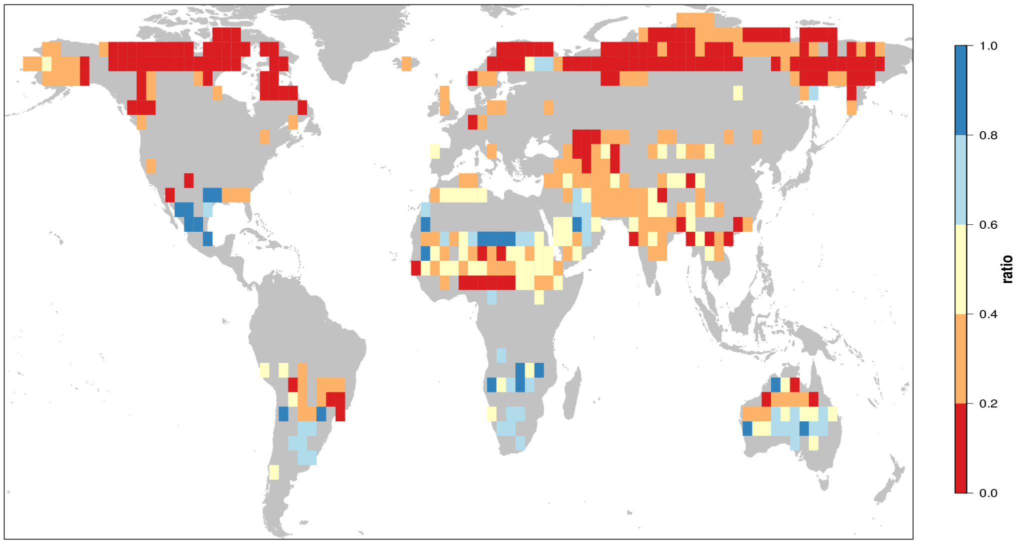


*Figure R4: Ratio of the standard deviation of the MTWA climate variable at modern analog sites over the standard deviation within the training set.*

- **main text paragraph p.5 l. 30 – p. 6. l. 10 on N2 vs. richness**
- **new Supplementary Figure SFig. 2**
- **We emphasize the limitations of our approach due to the low number of taxa (p. 15 l. 2- p. 16 l. 17)**

**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.

**Implemented changes in the revised manuscript:**

- **inserted new discussion paragraph (p. 5 l. 30- p. 6 l. 10)**
- **new Supplementary Figure SFig. 3**
- **We emphasize the limitations of our approach due to the low number of taxa (p. 15 l. 2- p. 16 l. 17)**

**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.

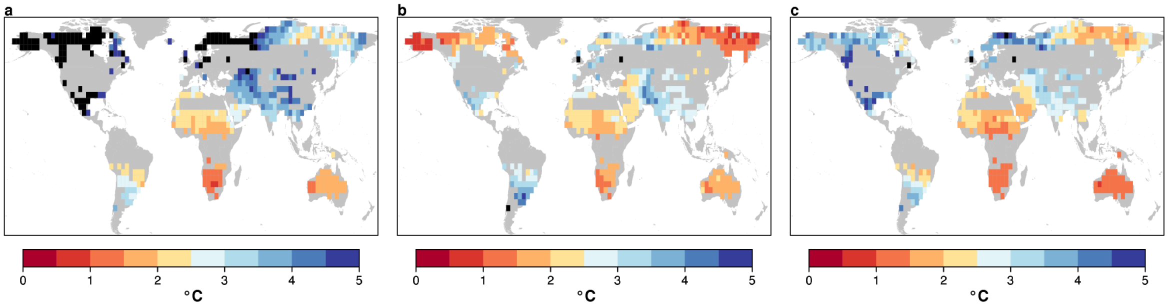


*Figure R5: RMSEP for 10-fold-leave-group-out cross validation of MTCO (a), MTWA (b) and MAT (c). Black gridcells indicate a RMSEP larger than 5C.*

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.

**Implemented changes in the revised manuscript:**

- **new Supplementary Figure SFig. 6, discussed in the manuscript text on p. 10 l. 7-9.**
- **enhanced Table 1 & improved linkage to the text**
- **masking of grid points with low transfer function performance (Figs. 4 & 5 in the manuscript)**
- **new discussion of uncertainties (for Fig. 8 & 9) on p. 14 l. 6-11.**

# Assessing performance and seasonal bias of pollen-based climate reconstructions in a perfect model world

Kira Rehfeld[1], Mathias Trachsel[2], Richard J. Telford[2,3], and Thomas Laepple[1]

[1]Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 14473 Potsdam, Germany
[2]Department of Biology, University of Bergen, Postboks 7803, N-5020 Bergen
[3]Bjerknes Center for Climate Research, Allégaten 55, N-5007 Bergen, Norway

*Correspondence to:* K. Rehfeld (kira.rehfeld@awi.de) and M. Trachsel (mathias.trachsel@uib.no)

**Abstract.** Reconstructions of summer, winter or annual mean temperatures based on the species composition of bio-indicators such as pollen, foraminifera or chironomids are routinely used in climate model-proxy data comparison studies. Most reconstruction algorithms exploit the joint distribution of modern spatial climate and species distribution for the development of the reconstructions. They rely on the assumption of 'uniformitarianism', which implies that environmental variables other than those reconstructed should not be important, or that their relationship with the reconstructed variable(s) should be the same in the past as in the modern spatial calibration dataset.

Here we test the implications of ~~uniformitarianism on such~~ this assumption on climate reconstructions in an ideal model world, in which climate and vegetation are known at all times. The alternate reality is a climate simulation of the last 6000 years with dynamic vegetation. Transient changes of plant functional types are considered as surrogate pollen counts, and allow to establish, apply and evaluate transfer functions in the modeled world.

We find that in our model experiments the transfer function cross-validation $r^2$ is of limited use to identify reconstructible climate variables, as it only relies on the modern spatial climate/vegetation relationship. However, ordination approaches that assess the amount of fossil vegetation variance explained by the reconstructions are promising. We furthermore show that correlations between climate variables in the modern climate/vegetation relationship are systematically extended into the reconstructions. Summer temperatures, the most prominent driving variable for modeled vegetation change in the Northern Hemisphere, are accurately reconstructed. However, the amplitude of the model winter and mean annual temperature cooling between the mid-Holocene and present day is overestimated, and similar to the summer trend in magnitude.

This effect occurs, because temporal changes of a dominant climate variable, such as summer temperature, are imprinted on a less important variable, leading to reconstructions biased towards the dominant variable's trends. Our results indicate that reconstructions of multiple climate variables from the same bio-indicator dataset should be treated with caution. Expert knowledge on the eco-physiological drivers of the proxies, and statistical methods that go beyond the cross-validation on modern calibration datasets are crucial to avoid misinterpretation.

## 1 Introduction

Continental-scale climate reconstructions (Bartlein et al., 2011; Davis et al., 2003; Mauri et al., 2014) are frequently used as a paleo-data target to evaluate and benchmark climate models (e.g. Harrison et al., 2014; Fischer and Jungclaus, 2011). These efforts have to rely on the fidelity of the ~~paleo-climate~~ paleoclimate reconstruction and the associated uncertainty estimates.

To arrive at quantitative assessments of past climate changes from pollen assemblages, transfer function algorithms are used to establish a link between modern climate and vegetation composition across space. The derived relationships are then applied to fossil pollen percentages, counted in sediment archives. The main challenge for quantitative interpretations is the fundamental ~~assumption (“the law of uniformitarianism”, Scott (1963) )~~ “uniformitarian principle’ (Scott, 1963) in transfer functions. It states, that the same laws govern species, or vegetation, distribution along climatic and environmental gradients in space, as they did at individual sites through climatic changes (Juggins, 2013). A presumption for the establishment of ecological transfer functions for climate reconstruction is therefore that environmental variables other than those considered in the calibration are not important, or that their relationship with the reconstructed variable(s) is the same in the past as in the modern spatial calibration dataset (Birks and Seppä, 2005). ~~These assumptions have~~ This assumption has been discussed since the early days of quantitative reconstructions based on paleoecological data (see, e.g. Birks et al., 2010; Juggins, 2013, and references therein). However, without knowing the past climate evolution, it is difficult to estimate to what extent ~~this assumption might be~~ it has been violated, and what the potential implications for reconstructing the Holocene climate evolution are.

Here, we use a climate model simulation with interactive vegetation as a testbed for pollen transfer ~~functions in the Holocene~~function methods. In the model world, the modern spatial climate and its relationship to vegetation is known, along with the Holocene climate and vegetation evolution.

Our general approach bears some similarities to previous ‘pseudoproxy’ experiments, where climate model simulations were used to test calibrations for temperature reconstructions of the last ~~millenia~~ millennia (Mann and Rutherford, 2005; Küttel et al., 2007; von Storch et al., 2004). However, as these studies target proxy records for climate which are ~~largely without modern spatial calibrations (e.g. tree rings)they~~ calibrated temporally against meteorological data (such as tree ring parameters), they largely focus on the effect of ~~noise on temporal calibrations~~proxy noise on the reconstruction. We ignore these proxy imperfections ~~, as well as~~ and age uncertainty, and focus on the ~~assumption~~ implications of ‘uniformitarianism’, which ~~motivates~~ is the operational principle behind the use of spatial calibrations to reconstruct temporal changes.

Key questions are: (i) To what extent does ~~the assumption of~~ uniformitarianism, and aspects of the estimation processes, bias reconstructions of the Holocene temperature evolution? (ii) Are there statistical indicators that can inform us on actual reconstructability of climate variables?

To address these questions within the model world, we need to assume that model climate and vegetation changes are consistent with each other, and that modeled plant functional type (PFT) and land cover type changes (desert fraction) can be used as surrogates for pollen counts in sedimentary archives.

## 2 Methods

### 2.1 Climate model simulations

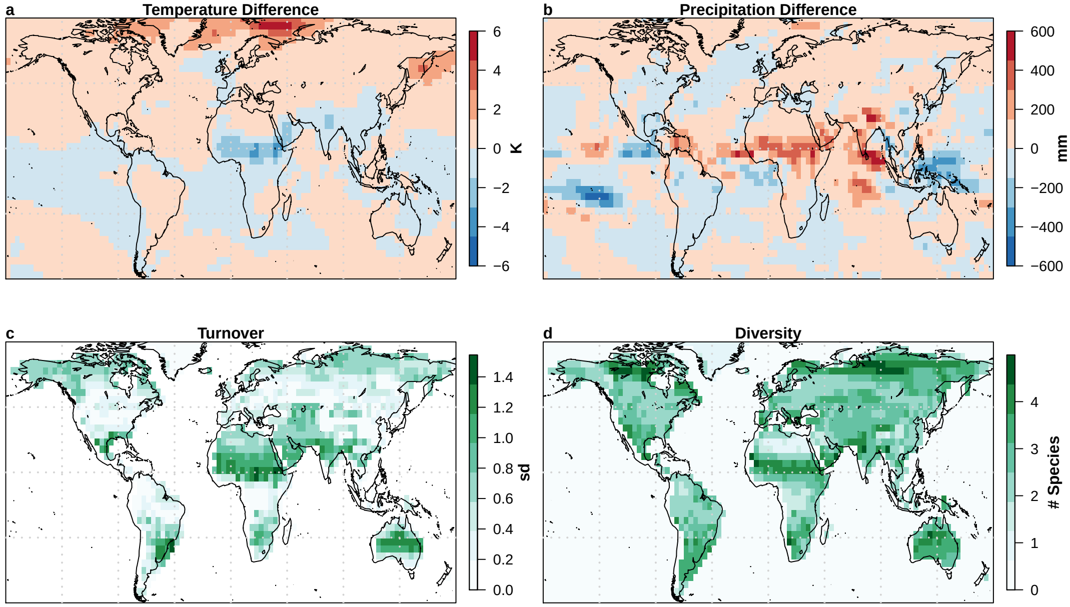


**Figure 1.** Temperature (a) and precipitation changes (b), vegetation turnover (c) and vegetation diversity as measured by the Hill's number ~~N2~~ $N_2$ (d) ~~across~~ between the ~~6k-run~~ 6k and 0k BP (Fischer and Jungclaus, 2011).

We use a 6000-year-long transient simulation of the coupled atmosphere-ocean climate model ECHAM5/MPIOM (Jungclaus et al., 2006) with a dynamic land surface and vegetation scheme provided by the JSBACH module (Raddatz et al., 2007; Brovkin et al., 2009) to investigate pollen-based climate reconstruction techniques. This simulation is described in (Fischer and Jungclaus, 2011) (hereafter *6k-run*) and is only forced by orbital changes over the last 6000 years. Environmental and atmospheric variables are available on a regular 3.75° × 3.75° latitude/longitude grid.

The vegetation module is described in Sitch et al. (2003) and Brovkin et al. (2009). The modeled climate-vegetation interaction through the growth, competition and mortality of the four tree, two shrub and two grass PFTs is nontrivial: Within each grid cell, plants compete for fractional cover, given their own net primary productivity, natural mortality as well as disturbance-driven mortality in response to climate (fire, heat and cold extremes, growing season length). Given a latitude, soil texture, $CO_2$ concentration, temperature and precipitation, processes changing water balance, photosynthesis, leaf cover and respiration are simulated on a daily or monthly time step. The turnover of wood, leaves and roots, decomposition, mortality and establishment is calculated annually, and the resulting vegetation cover is fed into the next year. Table 1 in the supplementary information lists the PFTs and their bioclimatic temperature limits.

The Holocene climate and vegetation evolution of this model simulation have been extensively used and characterized in paleoclimate model-data comparisons (Fischer and Jungclaus, 2011; Dallmeyer et al., 2011, 2013, 2015; Laepple and Huybers, 2014; Rehfeld and Laepple, 2016). While vegetation biases have been observed against present-day conditions in some areas (Brovkin et al., 2009; Dallmeyer et al., 2011), the overall patterns are consistent (Brovkin et al., 2009). Climate and vegetation changes from mid-Holocene to present day are substantial (Fig.1) and differ between the seasons (Fig.3, top row). We note that, although the resolution of the climate model, and thus the model world calibration dataset is coarse, its spatial and seasonal range is comparable to that of real-world calibration datasets (suppl. Fig. (SFig.) 1).

### 2.2 Reconstruction methods

Quantitative climate reconstruction (Juggins and Birks, 2012; Birks et al., 2010) based on a multivariate pollen count dataset requires algorithms that translate past vegetation changes into estimates of past climate changes. Most approaches use three datasets: A paired calibration set, and one downcore pollen record~~to be reconstructed~~. The calibration set combines modern pollen and climate data from recent, or modern, conditions taken from surface samples across ecological and climatic gradients. An example from the real-world would be pollen counts from lake sediment surfaces across Europe, paired with data from meteorological stations near these lakes. Several approaches for quantitative reconstructions based on ecological species counts have been established (see e.g. Birks et al., 2010, for a review). Here we focus on two popular techniques: Best Modern Analog methods (here: BMA, often also called Modern Analogue Technique), and the multivariate calibration method of Weighted Averaging (WA).

BMA methods directly match the species composition of fossil assemblages against the modern calibration set ~~.~~ (Overpeck et al., 1985). To obtain a reconstruction value for a fossil sample, $N$ analog modern samples with the lowest ecological distance (most commonly estimated using the Squared-Chord-Distance (Overpeck et al., 1985)) are selected. Their modern reference climate variables are averaged to obtain the past climate estimate. These approaches are expected to work well on samples with a low number of taxa~~, but estimates of calibration function errors may be biased low due to autocorrelation in climate and vegetation, as the method inherently favors nearby sites (Telford and Birks, 2005, 2009)~~ . In this study we use BMA with N = 5 and the Squared-Chord distance.

Multivariate calibrations, on the other hand, are based on the regression of modern vegetation onto estimates of a climate variable at many calibration sites, to establish one global parametric function between them. In WA calibration, climate optima for different taxa are derived by ~~performing~~ computing a weighted average of climate variable estimates at all sites at which a taxon is present. Weights are derived from the relative abundance of the taxon. The step from past vegetation composition to estimates of past climate then relies on a second weighting step, in which the climate optima of all taxa present in the fossil sample are averaged, again weighted by their relative abundance. We employ WA here to illustrate results that are common to reconstructions based on BMA and WA-related methods, which may therefore depend on properties of the dataset, or the general approach of reconstructing climate based on modern spatial climate calibrations. In this study ~~,~~ we use WA with square-root transformed scores and inverse deshrinking.

### 2.3 Estimates of reconstruction uncertainty

In a real-world situation, the true climate evolution is unknown and ~~the~~ a root mean square error of prediction (RMSEP) is estimated in the modern calibration set. In the following we use k-fold cross-validation with k=10 (1/k-th of the samples are used for verification) but note that even using ~~cross-validation~~leave-group-out-cross-validation, the RMSEP may be biased low due to autocorrelation in the modern data (Telford and Birks, 2005, 2009). As we know the true climate in the model world, we can additionally obtain the root mean square error of the reconstruction (RMSE) by comparing the reconstructed climate variable to its simulated counterpart.

We employ multivariate constrained ordination methods to test, which climate variables explain vegetation variance. While Redundancy Analysis (RDA) extends principal component analysis, Canonical Correspondence Analysis (CCA) is the equivalent method for frequency data (Borcard et al., 2011).

We evaluate the similarity between trend and correlation fields using a sign-test, similar to Kendall's rank correlation, defined as a fraction $\nu(X,Y) = \frac{S(X,Y)}{\text{\# reconstr. grid cells}}$ varying between -1 and +1. A grid cell counts into the sign sum $S(X,Y)$ as +1 if the signs in field X and field Y are the same, and as -1 if they are opposite. Summation goes over all grid cells where a reconstruction was performed. This sign test yields $\nu = 1$ if and only if all grid cells in field X and Y have the same sign, and $\nu = -1$ if all signs are opposing. $\nu = 0$ suggests, that there are as many grid cells with opposing signs as there are with the same signs, indicating that there is no underlying similarity between the fields.

### 2.4 Calibration and reconstruction workflow

We perform PFT-based calibrations and climate reconstructions at each grid point on land which displays enough diversity and temporal variations in the simulated vegetation. Therefore, we select all points for the reconstruction tests with an effective number of species ~~N2~~ $N_2$ larger than 2 (Hill, 1973)[1], and vegetation turnover larger than 0.5. Turnover is estimated from the length of the first detrended correspondence analysis axis in standard deviation units (Hill and Gauch, 1980).

The simulated vegetation history through time at a grid point forms the fossil vegetation dataset. The simulated modern surrounding vegetation and climate fields, averaged over the last 30 years, yield the matrices containing modern pollen and climate information for the modern training set. We select all surrounding land-points in a radius of 2500km and subsample them such, that the calibration set size is roughly equal for all sites and not latitude-dependent.

Pollen matrix columns contain the percentages of the ~~eight~~ nine PFTs (acronyms in Appendix A, details in Suppl. Table 1), ~~and~~ including the desert fraction as a virtual PFT. Each column in the modern climate matrix corresponds to a climate variable and we choose the warmest month, coldest month and annual mean temperatures (MTWA, MTCO, MAT) and precipitation (MPWA, MPCO, MAP) variables.

We note that large-scale PFT-based pollen reconstructions use roughly 2-3 times the number of PFTs (as e.g. in Davis et al., 2003; Mauri et and raw pollen spectra contain often more than 10 times the number of taxa. However, the effective number of species,

[1] The Hill's number $N_2$ is defined as $N_2 = \left(\sum_{i=1}^{N} p_i^2\right)^{-1}$, as the reciprocal of the weighted mean of the abundances $p$. If all taxa are equally abundant and $p_i = 1/N$, $N_2$ is equal to N. If only one taxon is present, and all others are zero, $N_2$ =1.

as estimated by Hill's $N_2$, is much lower than the number of taxa itself, and rare taxa do not have a large influence on reconstructions using BMA or WA. Our cutoff at $N_2 = 2$ is well within the range of $N_2$ for modern pollen spectra (SFig. 2). In general, a low number of PFTs or taxa may lead to a problem of multiple analogs, where a pollen assemblage is similar to several modern assemblages that are very different in their climatic setting (ter Braak et al., 1996) . However, supporting our cutoff choice at $N_2 = 2$, we do not find indications that this is a problem here. The overall high transfer function $r^2$ (Fig. 7) shows, that analogs are not picked at random from the training set. To pinpoint this further 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 (SFig. 3). The ratios are generally smaller than 0.5, thus illustrating that the analog sites are not randomly drawn from the training set.

In many conventional paleoecological studies one or two climate variables would be selected for reconstruction, which are expected to have influenced vegetation development significantly, and independently (Juggins, 2013; Telford and Birks, 2011). As we want to investigate, which variables can be skillfully reconstructed, we perform joint reconstructions of all six climate variables, both via BMA and WA. We note that jointly reconstructing several climate variables is done in several large-scale regional reconstructions (e.g. in Mauri et al., 2014; Bartlein et al., 2011; Davis et al., 2003) and come back to this later in the discussion.

Fig. 2 illustrates the whole calibration and reconstruction workflow for a BMA reconstruction at an example grid point selected from the Arctic (120°E,72°N). CCA analyses (Fig. 2d) suggest, that summer temperature is the main climate variable driving ~~vegetation development in the~~ modern vegetation around the site, whereas winter temperatures have little to no impact on the vegetation changes in the model~~(shown also in Suppl.Fig.3)~~. A summer temperature calibration based on BMA can explain considerable amounts of variance in the modern vegetation-climate relationship, it also shows a low RMSEP of $\sim 1.15°$ C. In the model world, we can compare reconstructed and the simulated true past model climate evolution (Fig. 2f) and find that summer temperatures (MTWA) are faithfully reconstructed, whereas the reconstructions of annual mean (MAT) and winter temperatures (MTCO) largely fail.

## 3 Results

### 3.1 Simulated and reconstructed Holocene temperature trends

The simulated mid-late Holocene temperature evolution shows a zonal structure characterized by warming trends around the Equator and across Asia and cooling trends in the mid-to-high latitudes (Fig. 3 top row). The seasonal insolation forcing caused by changes of the orbital configuration results in distinct temporal trends for summer and winter temperature, which differ in their strength and in some regions also in their signs. In the Arctic regions, the trends in the model simulation are strong ($\sim$-0.5K/kyr) for summer, and weaker ($\sim -0.1$K/kyr) for winter and the annual mean. The warming trends around the Equator appear strongest in the coldest month. Similar patterns occur in the mean annual precipitation, with drying in the Northern and wetting in the Southern Hemisphere. We focus here on temperature and refer the reader with interest in the precipitation changes to ~~SupplementaryFig~~SFig. ~~1.~~4.

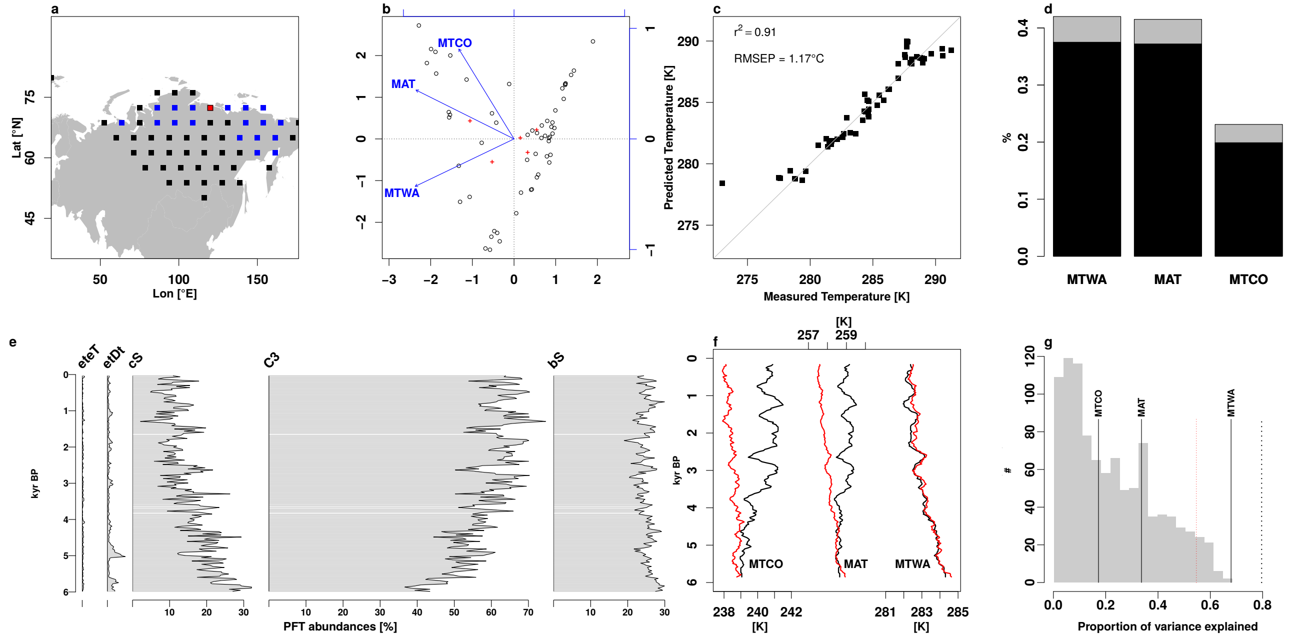


**Figure 2.** Exemplary calibration, BMA reconstruction and verification workflow for the grid point site in Siberia (120°E,~~70~~72°N) highlighted as a red square in (a). Surrounding grid points from which the modern analogs are drawn are shown as black dots, chosen analogs in blue. CCA analyses show, that MTWA explains most variance in modern vegetation (b), and performs sufficiently well in leave-one-out cross validation (c). The jointly reconstructed climate variables show considerable shared (black), and rather little independent variance (grey) in the modern calibration (d). Past vegetation changes, as shown in the percentage PFT diagram (e), appear to be correlated with (f) simulated and reconstructed climate. Red lines show the simulated 'true' past temperatures, black lines the reconstructions. The MTWA reconstruction explains most fossil vegetation variance in the `randomTF` significance test (Telford and Birks, 2011).

We now analyze the winter (MTCO), summer (MTWA) and annual mean (MAT) temperature patterns reconstructed using BMA and WA (Fig.~~3~~3 middle & bottom rows). Reconstructed winter trend patterns diverge from the simulated trends. In many regions the reconstructed trends are higher than ±1K/kyr in magnitude, and thus stronger than anywhere in the simulated model climate. Negative temperature trends in polar regions are not consistently captured, and an east-to-west warm-to-cold gradient appears for both reconstruction techniques WA and BMA.

In contrast, the reconstructed summer trends show broad similarities to the simulated temperature changes. Equatorial warming and polar cooling are captured by both WA and BMA. Differences exist in the magnitude of the changes, rather than the sign, except for in the Middle East, where warming is suggested by BMA and WA, and the true simulation trends showed a cooling, in particular around present-day Turkey.

Amongst the climate variables, MTWA appears to be most consistent between simulations and reconstructions. This is also supported by the results of the sign test (described in Sec. 2.3), which yields $\nu \approx 0.5$ for WA and BMA. MTCO is least

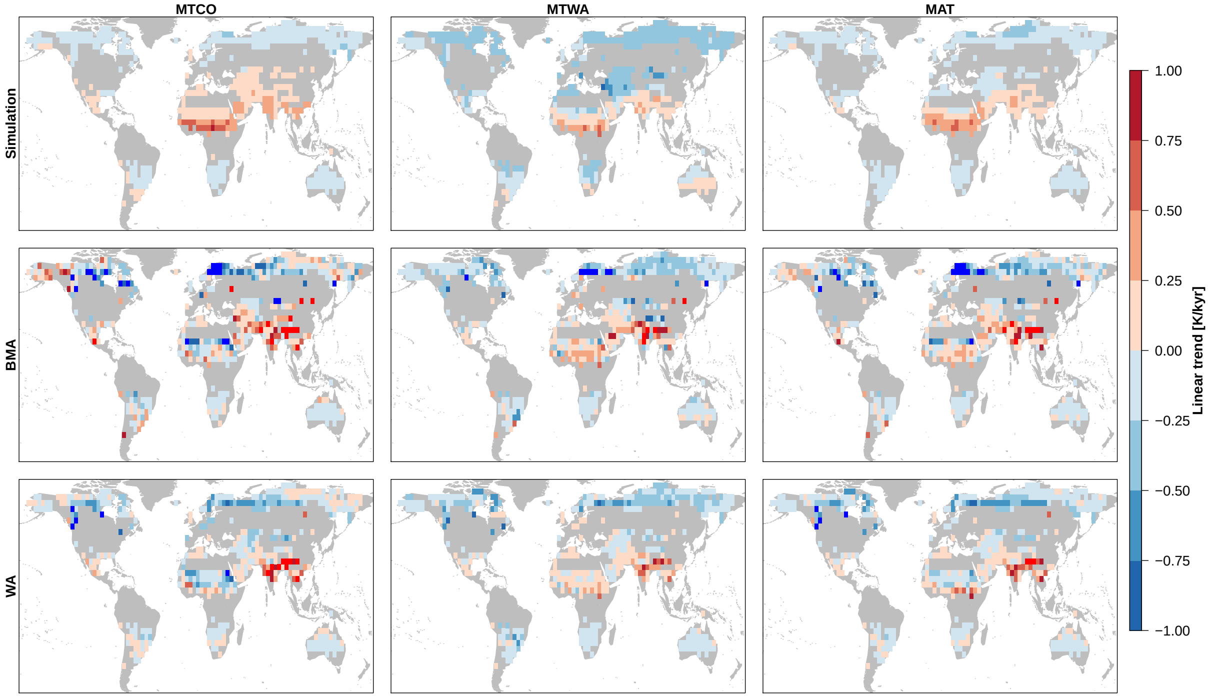


**Figure 3.** Linear trend in the simulated (top row) vs. the reconstructed temperature evolution between 6k and present day based on BMA (middle row) and WA (bottom row). Saturated red/blue colors indicate that the grid point's trends are stronger than 1K/kyr.

consistent ($\nu \approx 0.3$). Between WA and BMA, results appear more patchy for BMA than for WA (i.e. sign or magnitude vary less gradually across space), but this does not imply that either method captures correct degrees of change. This is further underlined by the temperature standard deviations taken across the trend fields, which are much larger for WA ($\overline{\text{sd}} = 1.8$K, bottom row in Fig. 3) and BMA ($\overline{\text{sd}} = 2.9$K, middle row) than for the simulation ($\overline{\text{sd}} = 1.2$K, top row). Thus, for both reconstruction methods reconstructed trends are spatially more heterogeneous than the simulated trends.

The spatial patterns and magnitudes of the reconstructed trends are very similar across all three seasons (compare panels across rows in Fig.3). Visually, they show a stronger similarity than the spatial patterns of the simulated seasonal trends (compare panels of the top row). This is due to the fact that grid cells with large positive or negative trends appear in the same positions across the seasons (i.e., row-wise), but not necessarily across methods (i.e., column-wise). The sign test shows slightly larger correspondences within each row/across seasons for the same method ($\overline{\nu} = 0.59$) than for the columns/same season across methods ($\overline{\nu} = 0.47$). Due to the influence of the strong trends in the same places, this discrepancy is stronger for Pearson correlations across the fields of Fig. 3 (by method $\overline{\rho} = 0.79$, by season ($\overline{\rho} = 0.46$). One explanation for this observation could be that all seasonal reconstructions are biased towards a single specific season.

### 3.2 Seasonal bias of temperature reconstructions

To further investigate this finding, we analyze the correlation between the different seasons in the simulations across modern space and across time and contrast them with the correlation through time between the reconstructed seasonal time series (Fig. 4). Ideally, the temporal correlation of the reconstructions should equal the temporal correlation of our ‘true’ (model simulated) climate evolution. Correlations across modern space are calculated over all the ~~grid-points~~ grid points relevant in the calibration and reconstruction process, thus for WA these are all ~~grid-boxes~~ grid boxes in a radius of 2500km whereas for BMA, only the sites picked as modern analog in the reconstruction are used (see Fig. 2a for an example). For simplicity, we perform the analysis for winter (MTCO) against summer (MTWA) temperature, but other variable combinations (e.g. temperature against precipitation) would lead to similar results.

Across modern space MTCO and MTWA are mostly positively correlated (Fig. 4a), as towards the poles temperatures get colder in summers as well as in winter. Exceptions are found around Eastern ~~Siberia~~ Russia and equatorial regions in Africa, where summer and winter temperatures are anti-correlated across space.

The temporal correlations of the WA-reconstructed MTCO and MTWA (Fig. 4b) show a very similar pattern of the correlation sign, although with stronger amplitudes of the correlation values. Indeed, the sign test yields $\nu = 0.76$, indicating that the large majority of the grid cells in Fig. 4a and Fig. 4b share the same sign. In contrast, the ‘true’ temporal MTCO/MTWA correlation over the late Holocene (Fig. 4e), which should ideally be similar to the reconstructed temporal correlation (Fig. 4b), shows a different picture ($\nu = 0.26$). This suggests that the modern spatial covariance has been directly propagated to the temporal covariance of the reconstructions.

Here, and in Fig. 5, we mask grid points for fossil reconstructions with low transfer function performance as measured by the cross-validation $r^2$, as we expect them to return less reliable results.

The same observation holds for the BMA-based results (Fig. 4d). The modern spatial MTCO/MTWA covariances at the sites picked as modern analogs, shown in Fig. 4c, are noisier than the covariances calculated over all ~~grid-boxes~~grid boxes, but show a similar pattern. The seasonal correlation in the BMA-reconstructions again directly follows the modern spatial MTCO/MTWA correlation ($\nu = 0.68$). In contrast, the similarity to the actual temporal covariance (Fig. 4e) is low, as the sign test underlines ($\nu = 0.03$).

### 3.3 Reconstruction skill

We showed that the ability to reconstruct Holocene temperature trends in our model world strongly depends on the analyzed season and region (Fig. 3). It is also important to quantify the reconstruction skill for the full Holocene evolution, including millennial variability and absolute temperature estimates. We analyze two metrics, (i) the temporal Pearson correlation between the ‘true’ past changes and the climate variable reconstructions (“correlation skill” , Fig. 5), and (ii) the RMSE deviation of the reconstructed from the ‘true’ climate.

Consistently high correlation skill values for the BMA reconstruction can be found across the Arctic for MTWA, and in

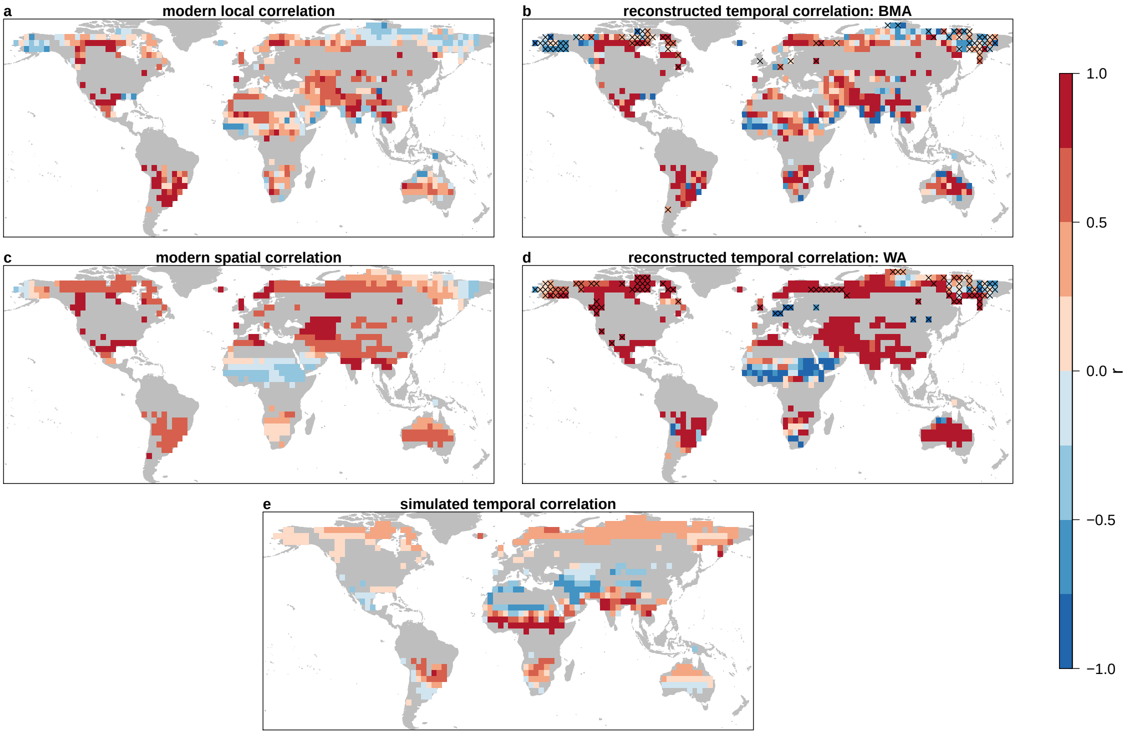

a
modern local correlation
b
reconstructed temporal correlation: BMA
c
modern spatial correlation
d
reconstructed temporal correlation: WA
e
simulated temporal correlation
1.0
0.5
0.0
−0.5
−1.0
r

**Figure 4.** Correlation of coldest and warmest month temperatures. The correlation patterns across modern calibration space (a) are similar to the temporal correlation pattern estimated from WA reconstructions (b). The correlations at the sites picked as modern analogs (c) are similar to those obtained in the final BMA reconstructions (d). In contrast, the 'true' temporal correlation pattern from the model temperatures differs considerably from the reconstructed temporal correlation fields. This demonstrates that the correlation in the reconstructions mainly depends on the modern calibration and not, as one would hope for, from the correlation of the Holocene temperature evolution. Crosses in (b) and (d) indicate gridboxes with a $r^2 < 0.5$ in cross-validation.

the Sahel for MAP. Simulated MAT changes are correlated with MTWA changes in the high latitudes, which explains the relatively weaker but positive correlation there. Winter ~~climate, and summer~~ precipitation reconstructions do not show good skill anywhere.

Most regions with high positive correlation skill show comparably low temporal RMSE (~~Fig.4 in the Suppl~~SFig. ~~Information~~5), whereas many regions with low RMSE do not show high correlation skill. In a real-world situation, the true past climate evolution is unknown and a root mean square error of prediction (RMSEP) is estimated from the modern calibration set (cf. Sec. 2.3). In our model world, the RMSEP is acceptable and below 3°C for MTWA and MAT, whereas it is generally high for winter temperature, in particular for North America. The low correlation skill for winter temperatures in the Arctic is also reflected by the temporal RMSE and the modern RMSEP (SFig. 5 & 6). A comparison of summer temperature downcore RMSE and modern spatial RMSEP, given in ~~Suppl.Fig~~SFig. 7, shows that modern RMSEP is higher than the actual reconstruction

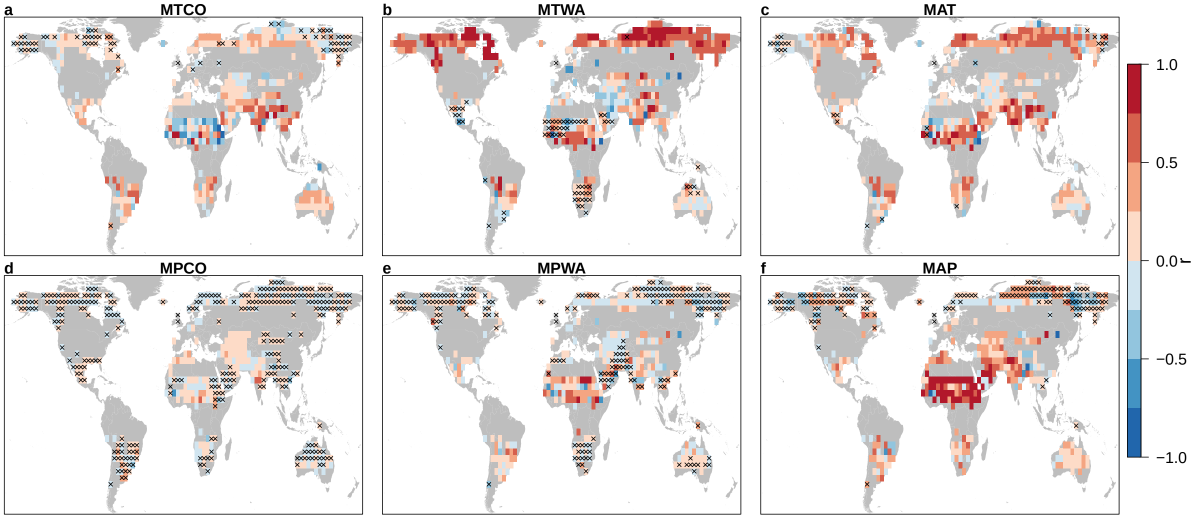


**Figure 5.** Performance of the BMA calibration models as evaluated by the correlation between the reconstructed and simulated climate variables (a-f) at each grid point. Crosses mask grid boxes with cross-validation $r^2$ <0.5.

error in many places, but there is little resemblance to the patterns of the estimated downcore RMSEP. ~~Furthermore, if the calibrations are performed with a smaller radius~~ If the calibration radius is reduced, the modern calibration error decreases (results not shown).

## 3.4 Testing for the predictability of reconstruction skill

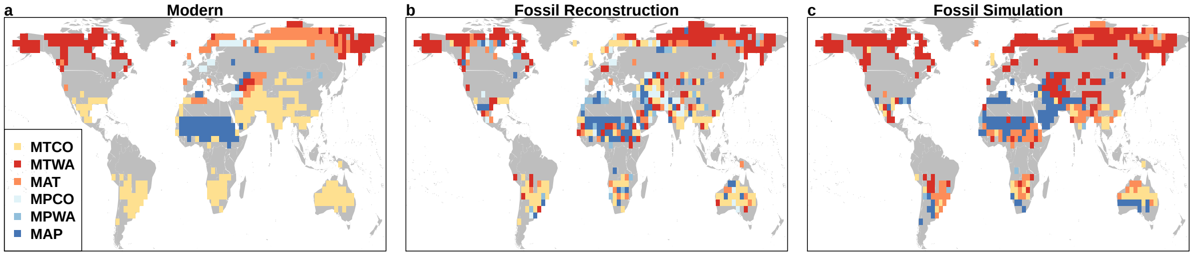


**Figure 6.** Climate variables explaining most variance in modern vegetation (a), between reconstructed climate and fossil vegetation (b) and simulated climate and fossil vegetation (c). Variable explaining most variance in the modern world (a) are not necessarily those explaining vegetation changes in the 'true' model past (c).

The inaccuracy of the covariance estimates (Fig. 4b), and the dependency of the reconstruction skill on the analyzed climate variable (Fig. 5) highlights, that it is important to determine which climate variables can be reconstructed in a given setting - and what other variables they are colinear with in the modern training set. We can discern two statistical approaches to identify the driving variable for climate-related vegetation changes: Those relying on the modern calibration set, and those which involve the fossil downcore record. In both, higher variance explained should be reflecting a higher environmental relevance (Juggins and Birks, 2012).

In the following, we compare the results of estimating the driving climate variable with both approaches (Fig. 6a,b), with the pattern of the 'true' climate variable explaining most simulated fossil vegetation change in our model simulation (Fig. 6c). The ordination fields underlying this summary figure are given in the SFigs. 8 to 10. For the modern spatial approach, we use CCA ordination of modern PFTs and climate to determine the climate variable which explains most vegetation variance across the modern calibration space (Fig. 6a). Temperature variables dominate the ordination results globally, except for the Sahel zone, which is dominated by precipitation-changes. MTWA explains most variance in arctic Canada and eastern Siberia, whereas MAT appears to dominate in Siberia and Northern Europe.

For the fossil downcore record approach, we identify which BMA-reconstructed climate variable explains most variance in the fossil vegetation set using constrained ordination (RDA). The results, as can be seen in Fig. 6b, are different and less smooth than those obtained for the modern spatial vegetation changes. Note that the patterns we observe here are highly similar to those identified from the ratio of the first two axes of the ordination (Juggins, 2013) (SFig. 11).

Finally, as we have access to the 'true' past vegetation and climate changes in the model world, we can assess, which climate variable explains most simulated fossil vegetation change. The RDA results, shown in Fig. 6c, confirm a strong summer temperature signal above the Arctic circle, and the potential existence of a precipitation signal in the Middle East and the Sahel zone.

Contemplating Fig.6a, b, and c we observe that the driving variables, identified by the fossil downcore approach (Fig. 6b) are closer to the true (Fig. 6c) driving variables than the driving variables estimated from the modern calibration dataset (Fig.6a). This suggests, that looking at the variance explained by downcore reconstructions may tell us more about what actually drove vegetation changes, than looking at the variance explained in modern vegetation.

Furthermore, analyzing the variance explained in the modern calibration dataset can suggest a high importance (by a high explained variance) for variables that are not necessarily relevant to vegetation development. This is due to the colinearity of the climate variables (c.f. Fig. 2b). This is demonstrated in Fig. 7, which shows the transfer function $r^2$ for all climate variables. In large parts of Siberia, MAT explained most variance (Fig. 6a). However, MTWA transfer function $r^2$ (Fig. 7b) is about as high as that of MAT (Fig. 7c) there, and dominates the rest of the Arctic. MAP appears well reconstructible in the Southern Hemisphere, in regions where MTCO also has a high transfer function $r^2$. Seasonal precipitation transfer functions do not perform well on inter-regional scales outside Africa. There, they appear to perform better, which is likely due to their colinearity with MAP (c.f. Fig. 6).~~Note that the patterns we observe here are highly similar to those identified from the ratio of the first two axes of the ordination Juggins (2013) , as can be seen in Suppl.Fig.2.~~

For the potentially more skillfull approach of using the downcore reconstruction to test for reconstruction skill, a formalized test (`randomTF`) has been proposed in Telford and Birks (2011). It relies on the comparison between the fossil variance explained by the actual reconstruction, and the variance explained by reconstructions based on surrogate modern climate (but using the same modern and fossil pollen assemblages). Above 50°N, where temperature changes occur over the course of the 6k-run,

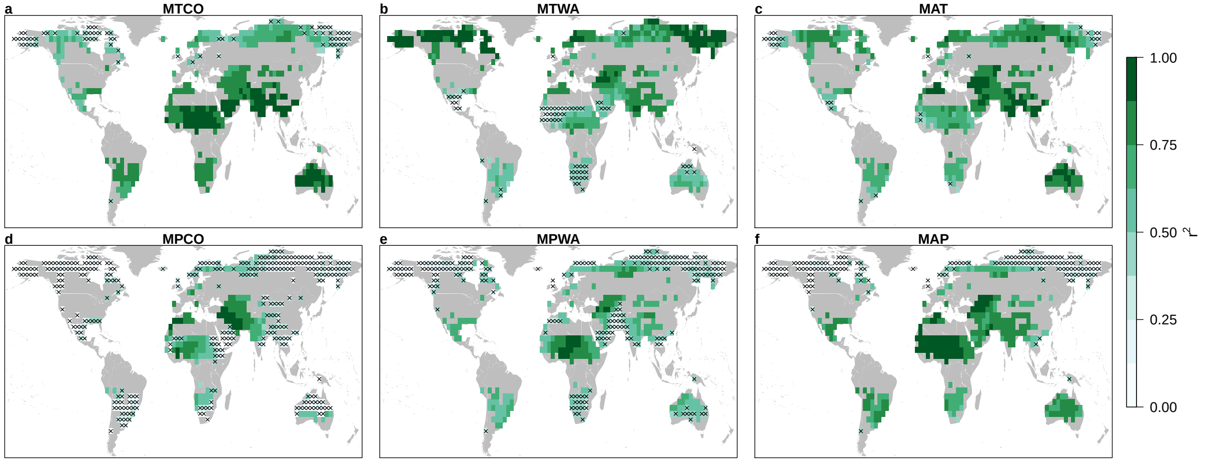


**Figure 7.** Spatial patterns of BMA transfer function $r^2$ in the modern calibration set (grid points with a distance of less than 2500km from the reconstruction site) of the six jointly reconstructed climate variables MTCO (a), MTWA (b), MAT (c), MPCO (d), MPWA (e), MAP (f). Points with a $r^2 < 0.5$ are crossed out. Transfer function performance appears good, although some variables had little impact on vegetation changes in the past.

84.7% of the grid cell vegetation changes are identified as most strongly related to MTWA (Table 1). If the `randomTF`-test has power, it should indicate a lower p-value for reconstructions of climate variables that were related to vegetation changes. Table 1 indicates a significant p-value ($\leq 0.1$) for MTWA in 68.9% of grid cells. MAT, picked as most relevant in 14% of the grid cells, appears reconstructible in 23% of the grid cells. MTCO, MAP, MPCO and MPWA – which have no or little relevance for vegetation development in the region – show up as significant in only 14-16% of the grid cells. Although our test approach does not meet the criteria of a formal statistical power assessment, these results suggest, that `randomTF` may have indicative power.

### 3.5 Influence of the modern climate background on the reconstructed climate

Following the principle of uniformitarianism, a reconstruction should not depend on the climate state in which the calibration set was taken. We test this in a case study, by ~~calibrating once, as throughout the manuscript so far,~~ comparing the calibration to the most recent time period (the last 30 years of the model run, equivalent to 0-30yrs BP) ~~, and once to~~ which we use throughout the manuscript, to one for the first period (5970-6000 yrs BP) in the simulation. We subsequently perform reconstructions for both calibration periods. Fig. 8 shows exemplary BMA results for a Siberian site.

Averaged across all reconstruction sites, MTWA reconstructions calibrated at 6k are .75K (-3.6,1.7K, 90% confidence interval) warmer than those based on calibrations at 0k~~(90confidence interval)~~. In particular, sites across the Northern Hemisphere are reconstructed with warmer temperatures. Inspection of the locations and temperatures around the analog sites chosen for the 0k and 6k calibrations suggests, that the warm bias may be caused by spatial autocorrelation in the vegetation, rather than

**Table 1.** Outcome of the significance test using `randomTF`. All 196 grid points above 50°N are considered, and p-values are estimated for all climate variables. Actual relevance is obtained by counting the number of times the variable is picked as the most relevant variable in the RDA of simulated climate and vegetation (Fig. 6) and dividing by the number of grid cells.~~Temperature units are in Kelvin, precipitation in mm/year.~~

| | | **`randomTF`: significant ($p < 0.1$)** | | | **`randomTF`: not significant ($p > 0.1$)** | | |
|---|---|---|---|---|---|---|---|
| | **Relevance [%]** | **RMSEP** | **r(rec,sim)** | **No. cells [%]** | **RMSEP** | **r(rec,sim)** | **No. cells [%]** |
| **MTCO [°C]** | 1.5 | 4.16 | 0.17 | 13.8 | 3.31 | 0.08 | 86.2 |
| **MTWA [°C]** | 84.7 | 0.92 | 0.71 | 68.9 | 2.00 | 0.37 | 31.1 |
| **MAT [°C]** | 13.8 | 2.43 | 0.56 | 23.5 | 2.13 | 0.26 | 76.5 |
| **MPCO [mm yr$^{-1}$]** | 0.0 | 180.80 | -0.03 | 9.2 | 113.63 | 0.00 | 90.8 |
| **MPWA [mm yr$^{-1}$]** | 0.0 | 237.44 | 0.06 | 16.3 | 184.9 | ~~0.0~~ 0.00 | 83.7 |
| **MAP [mm yr$^{-1}$]** | 0.0 | 150.52 | 0.21 | 15.8 | 123.76 | 0.04 | 84.2 |

climate, in addition to other local confounding factors. The 6k analog sites tend to lie further northward (in the Northern Hemisphere) than those for the 0k calibration. However, the 6k analog sites do not systematically cluster northward. Therefore, the northward migration of the analog sites does not compensate fully for the warmer background climate state, so that the overall reconstructed temperatures are warmer. This demonstrates that, at least in our experiment, the climatological and ecological similarity of the calibration period to the period for reconstruction influences the reconstruction outcome.

The question whether the detected differences in Fig. 8 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., 20 This assumption would result in a standard error of 0.13°C, thus considerably smaller than the differences we found. In the (unrealistic) extreme case of a complete dependency of errors, the differences would be not significant. In reality the true uncertainty likely lies between the two extremes assumed here but a more detailed analysis of the spatial and temporal covariance structure of the proxy uncertainty is required to provide better error estimates.

## 4 Discussion

Using a Holocene climate model simulation as a testbed for pollen based climate reconstructions allowed us to analyze the reconstruction skill and to understand potential seasonal biases of pollen based climate reconstruction methods.

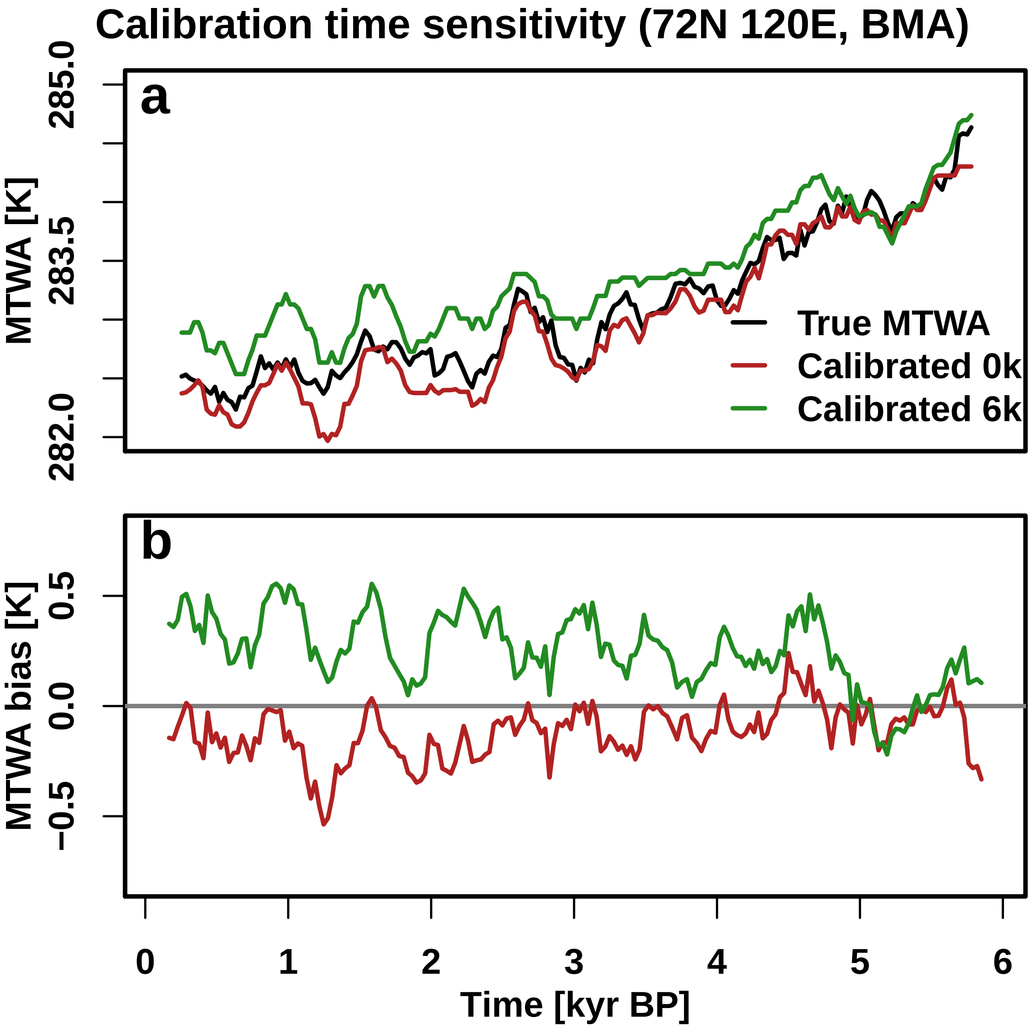


**Figure 8.** Reconstructions are sensitive to the calibration time period. Warmest month temperature trends for reconstructions based on a calibration for the last 30 years (0k) and first 30 years (6k) of the model run (A). 6k results are mostly warmer (B). All time series are based on 300-year running means.

### 4.1 Limitations

The ~~model world we have investigated does certainly not describe actual past climate and vegetation dynamics. However, for our calibration exercise, we only require vegetation and climate to be consistent, they need not necessarily be representative for the late Holocene. Furthermore, the methods we have tested are limited by~~ complexity of the vegetation representation in the model as well as the simulated climate evolution are a strong simplification of reality. Therefore, results on the Holocene evolution of specific PFTs, the actual spatial pattern of PFTs, or the reconstructability of a certain climate variable in a certain region should not be directly translated to the real world. On the other hand, conclusions on reconstruction methods and the relation of spatial calibration and downcore reconstruction only require a consistent dataset of climate and vegetation parameters in space and time and do not depend on details of the climate evolution or vegetation response, as long as the dataset is realistic enough that we can apply the real world reconstruction workflow. The major factor shaping our results is that the modern spatial relationships between climate variables is different from the changes in the relationships over time, which is a robust feature related to the transient insolation forcing (Laepple and Lohmann, 2009) .

One might be concerned that the low number of simulated plant functional types, ~~as large-scale PFT-based pollen reconstructions use roughly 2-3 times the number of PFTs (as e.g. in Davis et al., 2003; Mauri et al., 2014) .We have therefore~~ or the low spatial resolution of the model might bias our reconstruction efforts. However, we showed that the actual information contained in the plant functional types and the spatial climate field is not fundamentally different than in the PFTs (or taxa) and the climate calibration datasets used in real world reconstructions (SFigs. 1 & 2).

Given the design of our study, we have limited our analyses to ~~identify~~ identifying general features of the calibration vs. reconstruction relationship rather than interpreting the actual numbers of temperature changes or ~~recontruction biases. In our stude~~reconstruction biases. Furthermore, we assumed perfect proxy recording ~~,~~ and did not add any non-climatic noise. If these were added, tests which rely on the downcore record, such as `randomTF`, may become less powerful, and downcore RMSE could become higher.

### 4.2 Identification of climate variables driving vegetation evolution through time

Our study shows that in our model world, regardless of the reconstruction technique, the reconstructed climate evolution is very similar between the variables (Fig. 3). This strong covariance between the variables is determined by the modern spatial covariance and not, as one would hope, the temporal covariance of local climate (Fig. 4). This finding can be understood in a simple thought experiment. Let us assume that the ~~(model)~~ vegetation evolution at every grid point ~~is~~ would be driven by one single variable. This single variable could be one of the analyzed variables (e.g. summer temperature) or any other variable, such as the length of the growing season, cloudiness or soil moisture. All other variables have no direct influence on the vegetation, themselves, and are merely covarying with the driving variable. In this case, the reconstructed covariability is implicit in the transfer function and fully determined from the modern spatial relationship regardless of the true past relationship between the variables similar to what we found (Figs. 3 and 4).

Reconstruction skill will consequently depend on whether we reconstruct the driving variable, or, in case that we reconstruct a secondary variable, on the question whether the the relationship with the driving variable is the same across space and in time. The example of our model-world Arctic shows, that the latter is not always the case. Past vegetation changes there, as Fig. 6 shows, were predominantly driven by summer temperature and mean annual temperature change, yet the modern transfer function $r^2$ for MTCO is acceptable in most grid boxes (Fig. 7). Skill for winter temperature reconstructions are, however, low (Fig. 5a), particularly in regions where the modern spatial covariance between summer and winter temperatures (Fig. 4a,c) is negative, whereas the temporal covariance is positive (Fig. 4e).

Therefore, an important question is whether we can determine the variable driving vegetation changes. This would increase our confidence in the reconstruction. In the simplest case, vegetation patterns across modern space are only determined by the current climate. In this case, the climate variable maximizing the modern spatial correlation, information accessible in the real world, would be the driving variable (Fig. 6a). However, the variable explaining most of the modern spatial vegetation variance was, in our evaluation, not necessarily the one explaining most of the temporal vegetation evolution (compare Figs. 6a vs. 6c). Therefore, either other parameters ~~than just the~~ beyond modern climate play a role, or the driving variable was not included in our set of six variables. In the model world, and likely in reality, both occurs. Evolving parameters such as soil properties are partly determining the spatial vegetation distribution, but are constant over time in the model world. On the other hand, chances to identify the correct driving variable are also small, as, for example, the length of the growing season might have a stronger influence than summer temperature. What follows from this is that methods only relying on the modern spatial climate/vegetation relationship are insufficient to identify the driving variables across time. Here, inverse modeling reconstruction techniques which do not rely on modern spatial calibration sets (Guiot et al., 2009; Yu, 2013) may provide useful additional information. In addition to the downcore tests outlined in Sect. 3.4, a priori expert knowledge on regional ecology is helpful to identify variables of climatic and ecological relevance.

### 4.3 Seasonal bias on reconstructed trends in non-driving variables

In the Northern Hemisphere extratropics of our model world, summer temperature is the variable driving vegetation change across the mid-to-late Holocene. The modern spatial correlation between summer, winter and consequently also mean annual temperatures is positive. Since the modern spatial information determines the downcore temporal reconstruction for all variables, the reconstructions of winter/annual mean temperature changes are biased towards the trend in summer temperatures. What are the implications of such a bias on reconstructions of climate variables which are not primarily influencing vegetation? Fig. 9 shows the simulated and BMA-reconstructed summer and annual mean temperature ~~changes~~for the Northern hemisphere extratropics (all grid boxes north of 50°N). Patterns and magnitudes are highly similar for WA, as well as when only grid boxes with summer/annual mean temperature as dominant variables are picked (not shown). Mid-to-late Holocene summer temperatures are slightly overestimated, but the trend and magnitude are correct. In contrast, the annual mean cooling has the same magnitude as the reconstructed (and simulated) summer cooling – it is exaggerated due to the summer bias in the reconstruction.

Such a correlation bias on jointly reconstructed climate variables is hard to detect and prove for real-world data. However, the above considerations suggest that for non-driving variables physically implausible temperature reconstructions may arise due to correlations across modern space. Consequently, estimated temperature trends based on proxy data may appear larger than in the model world, or may have a different shape. One example is the reconstruction of the annual mean temperature evolution of the past 11000 years (Marcott et al., 2013). The reconstructed cooling trend in the mid-late Holocene was stronger than the cooling simulated by climate models, a mismatch potentially related to a seasonal bias of the reconstruction (Meyer et al., 2015; Liu et al., 2014). Another example is the comparison between pollen-proxy-based and climate model simulated winter temperature changes between the Last Glacial Maximum and present day, which are stronger in the reconstructions than in the model simulations (Braconnot et al., 2012). Given our above results, such findings could potentially be explained as changes that are overestimated in the proxy data due to confounding effects of third variables, for example summer ~~temperatures.~~ length or precipitation changes.

### 4.4 Implications and Outlook

~~While we~~ We have focused our ~~study on the seasonality of temperatures~~analysis on the seasonal evolution of temperatures. However, it is likely that similar biases also affect pollen-assemblage-based reconstructions of other climate variables, such as precipitation. In this light, the result of larger pollen-derived than model simulated precipitation changes between the mid-Holocene and present-day (Braconnot et al., 2012) might be influenced by a reconstruction bias as the linkage between temperature and precipitation (Trenberth, 2005), may differ across space, time, and timescales (Rehfeld and Laepple, 2016). Similarly, modern spatial relationships differing from past temporal relationships ~~will~~ might also affect other assemblage-based climate reconstructions. Examples include planktonic foraminifera counts which are used to reconstruct marine temperature changes; in this case, the climate variables include water temperatures at different seasons and water depths (Telford et al., 2013). Similar effects ~~are likely also found~~ might also be in place for other environmental or climate proxies such as chironomids, diatoms and dinoflagellates (Telford and Birks, 2011), which all rely on modern spatial calibration approaches. Consequently, it ~~could~~ would be interesting to study ecological, geographical and climatic effects on reconstruction results in other ecological models (e.g. FORAMCLIM Lombard et al., 2011). In the vegetation model used, the simulated PFTs have broad climatic tolerances (Suppl. Table 1). This might exaggerate the seasonal bias problem, as the winter sensitivity of the simulated vegetation might too be low. While this would strengthen our general conclusion that transfer function diagnostics based on modern calibration data alone are not sufficient to characterize reconstructability, it asks for a cautious interpretation of the magnitude of the reconstruction bias.

This study could be extended in several directions. Adding ~~recording~~ proxy noise and age uncertainty would allow a more in-depth comparison of spatial and temporal errors, and a more representative test of the `randomTF`-algorithm. ~~Using transient~~ Repeating this study with a dynamic vegetation model that simulates a larger number of PFTs (Sitch et al., 2003, e.g. LPJ-GUESS) could provide more insight. Transient paleoclimate model experiments with such a more complex land surface and biosphere scheme (i.e., with a larger number of PFTs) could be particularly useful to test, whether assemblage-based climate reconstruction methods allow for the accurate joint reconstruction of several climate variables. A first ~~warning about~~ estimate of potential biases in model-data comparison of multiple climate variables can be obtained through the comparison of simulated spatial and temporal covariances. If they are very different, caution is called for in the interpretation of joint proxy reconstructions of these variables.

## 5 Conclusions

Using a Holocene climate model simulation with interactive vegetation as a testbed, we analyzed the skill and potential biases in pollen-based climate reconstructions. We find that in our model experiments, transfer function reconstruction methods pull the spatial covariances between climate variables through into the downcore temporal reconstructions. As a consequence, temporal changes of a dominant climate variable (for the Northern Hemisphere: often summer temperature) are imprinted on a less important variable (here: often winter temperature), leading to reconstructions biased towards the dominant variable's trends. ~~Our analyses suggest~~

The principle of uniformitarianism underpinning transfer-function climate reconstructions assumes that environmental variables other than those considered in the calibration are not important, or that their relationship with the reconstructed variable(s) is the same in the past as in the modern spatial calibration dataset. In our model world, we have clearly shown that this assumption is violated, as the modern spatial relationship between climate variables, such as winter and summer temperatures and the past temporal relationship often differs. Translating this to real world reconstructions would imply that large-scale reconstructions of multiple climate variables need to be carefully considered, as reconstructions of climate variables which are not primarily influencing vegetation can be biased. ~~Spatial and temporal vegetation changes are not always caused by the same physical mechanisms, violating the law of uniformitarianism underpinning transfer-function climate reconstructions. Therefore, rather than from~~ It would also imply that the driving climate variables cannot be reliably determined by only analyzing the modern spatial climate-vegetation relationship. Therefore, climate variables which actually drove vegetation variability in the past are likely better identified using expert knowledge on ecology, and with statistical analyses involving the fossil vegetation record.

## Appendix A: Acronyms

**PFT** Plant functional type

- **teT** PFT: tropical evergreen trees
- **tdT** PFT: tropical deciduous trees
- **eteT** PFT: extratropical evergreen trees
- **etdT** PFT: extratropical deciduous trees
- **rS** PFT: raingreen shrubs
- **cS** PFT: cold shrubs
- **C3** PFT: C3 grass
- **C4** PFT: C4 grass

**BMA** Best modern analog method (in literature also: Modern Analog approach) ~~Weighted averaging (partial least squares)~~

**WA** Weighted averaging

**RDA** Redundancy analysis

**CCA** Canonical correspondence analysis

**RMSE(P)** Root mean square error (of prediction)

**MAT** Mean annual temperature

**MTWA** Mean temperature warmest month

**MTCO** Mean temperature coldest month

**PANN** Mean annual precipitation

**MPCO** Mean precipitation coldest month

**MPWA** Mean precipitation warmest month

## Appendix B: Used software

All analyses were carried out in the open source environment `R`, version 3.2.2. Reconstructions were performed using the `rioja` package (v. 0.9-5), `paleosig` (v. 1.1-3) and the `vegan` library (v. 2.3-0). The code is available on request.

*Acknowledgements.* We gratefully acknowledge Anne Dallmeyer and Johann Jungclaus for providing the ECHAM5/MPIOM model output, ~~and~~ Ulrike Herzschuh for discussion and two anonymous reviewers whose reviews helped us to improve the manuscript. We thank the Initiative and Networking Fund of the Helmholtz Association (grant VH-NG-900) for funding and the DAAD-PPP program (contract 57160457) for travel support.

## References

Bartlein, P. J., Harrison, S. P., Brewer, S., Connor, S., Davis, B. A. S., Gajewski, K., Guiot, J., Harrison-Prentice, T. I., Henderson, A., Peyron, O., Prentice, I. C., Scholze, M., Seppä, H., Shuman, B., Sugita, S., Thompson, R. S., Viau, A. E., Williams, J., and Wu, H.: Pollen-based continental climate reconstructions at 6 and 21 ka: A global synthesis, Clim. Dyn., 37, 775–802, doi:10.1007/s00382-010-0904-1, 2011.

Birks, H. J. B., Heiri, O., Seppä, H. and Bjune, A. E.: Strengths and Weaknesses of Quantitative Climate Reconstructions Based on Late-Quaternary Biological Proxies, doi:10.2174/1874213001003020068, 2011.

Birks, H. J. B. and Seppä, H.: Pollen-based reconstructions of late-Quaternary climate in Europe - progress, problems, and pitfalls., Acta Palaeobot., 44, 317–334, 2005.

Borcard, D., Gillet, F., and Legendre, P.: Numerical Ecology with R, doi:10.1007/978-1-4419-7976-6, 2011.

Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., and Zhao, Y.: Evaluation of climate models using palaeoclimatic data, Nat. Clim. Chang., 2, 417–424, doi:10.1038/nclimate1456, 2012.

Brovkin, V., Raddatz, T., Reick, C. H., Claussen, M., and Gayler, V.: Global biogeophysical interactions between forest and climate, Geophys. Res. Lett., 36, L07 405, doi:10.1029/2009GL037543, 2009.

Dallmeyer, A., Claussen, M., Herzschuh, U., and Fischer, N.: Holocene vegetation and biomass changes on the Tibetan Plateau – a model-pollen data comparison, Clim. Past, 7, 881–901, doi:10.5194/cp-7-881-2011, 2011.

Dallmeyer, A., Claussen, M., Wang, Y., and Herzschuh, U.: Spatial variability of Holocene changes in the annual precipitation pattern: A model-data synthesis for the Asian monsoon region, Clim. Dyn., 40, 2919–2936, doi:10.1007/s00382-012-1550-6, 2013.

Dallmeyer, A., Claussen, M., Fischer, N., Haberkorn, K., Wagner, S., Pfeiffer, M., Jin, L., Khon, V., Wang, Y., and Herzschuh, U.: The evolution of sub-monsoon systems in the Afro-Asian monsoon region during the Holocene - comparison of different transient climate model simulations, Clim. Past, 11, 305–326, doi:10.5194/cp-11-305-2015, 2015.

Davis, B. A. S., Brewer, S., Stevenson, A., and Guiot, J.: The temperature of Europe during the Holocene reconstructed from pollen data, Quat. Sci. Rev., 22, 1701–1716, doi:10.1016/S0277-3791(03)00173-2, 2003.

Fischer, N. and Jungclaus, J. H.: Evolution of the seasonal temperature cycle in a transient Holocene simulation: orbital forcing and sea-ice, Clim. Past, 7, 1139–1148, doi:10.5194/cp-7-1139-2011, 2011.

Guiot, J., Wu, H. B., Garreta, V., Hatté, C., and Magny, M.: A few prospective ideas on climate reconstruction: from a statistical single proxy approach towards a multi-proxy and dynamical approach, Clim. Past, 5, 571–583, doi:10.5194/cp-5-571-2009, 2009.

Harrison, S. P., Bartlein, P. J., Brewer, S., Prentice, I. C., Boyd, M., Hessler, I., Holmgren, K., Izumi, K., and Willis, K.: Climate model benchmarking with glacial and mid-Holocene climates, Clim. Dyn., 43, 671–688, doi:10.1007/s00382-013-1922-6, 2014.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A.: Very high resolution interpolated climate surfaces for global land areas,Int. J. Climatol., 25, 1965–1978, 2005.

Hill, M. O.: Diversity and evenness: a unifying notation and its consequences, Ecology, 54, 427–432, doi:10.2307/1934352, 1973.

Hill, M. O. and Gauch, H. G.: Detrended correspondence analysis: an improved ordination technique, Vegetatio, 42, 47–58, 1980.

Juggins, S.: Quantitative reconstructions in palaeolimnology: new paradigm or sick science?, Quat. Sci. Rev., 64, 20–32, doi:10.1016/j.quascirev.2012.12.014, 2013.

Juggins, S. and Birks, H. J. B.: Data handling and numerical techniques, in: Dev. Paleoenviron. Res. Track. Environ. Chang. Using Lake Sediments, edited by Birks, H. J. B., Lotter, A. F., Juggins, S., and Smol, J. P., chap. 14, p. 745, Springer, Berlin/Heidelberg, 2012.

Jungclaus, J. H., Keenlyside, N., Botzet, M., Haak, H., Luo, J.-J., Latif, M., Marotzke, J., Mikolajewicz, U., and Roeckner, E.: Ocean Circulation and Tropical Variability in the Coupled Model ECHAM5/MPI-OM, J. Clim., 19, 3952–3972, doi:10.1175/JCLI3827.1, 2006.

Küttel, M., Luterbacher, J., Zorita, E., Xoplaki, E., Riedwyl, N., and Wanner, H.: Testing a European winter surface temperature reconstruction in a surrogate climate, Geophys. Res. Lett., 34, 2–7, doi:10.1029/2006GL027907, 2007.

Laepple, T. and Lohmann, G.: Seasonal cycle as template for climate variability on astronomical timescales, Paleoceanography, doi:10.1029/2008PA001674, 2009.

Laepple, T. and Huybers, P.: Ocean surface temperature variability: Large model-data differences at decadal and longer periods., Proc. Natl. Acad. Sci. U. S. A., doi:10.1073/pnas.1412077111, 2014.

Liu, Z., Zhu, J., Rosenthal, Y., Zhang, X., Otto-Bliesner, B. L., Timmermann, A., Smith, R. S., Lohmann, G., Zheng, W., and Elison Timm, O.: The Holocene temperature conundrum, Proc. Natl. Acad. Sci., pp. 1–5, doi:10.1073/pnas.1407229111, 2014.

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, doi:10.5194/bg-8-853-2011, 2011.

Mann, M. and Rutherford, S.: Testing the fidelity of methods used in proxy-based reconstructions of past climate, J. Clim., pp. 4097–4107, 2005.

Marcott, S. A., Shakun, J. D., Clark, P. U., and Mix, A. C.: A reconstruction of regional and global temperature for the past 11,300 years., Science, 339, 1198–201, doi:10.1126/science.1228026, 2013.

Mauri, A., Davis, B. A. S., Collins, P. M., and Kaplan, J. O.: The influence of atmospheric circulation on the mid-Holocene climate of Europe: a data–model comparison, Clim. Past, pp. 1925–1938, doi:10.5194/cp-10-1925-2014, 2014.

Meyer, H., Opel, T., Laepple, T., Dereviagin, A. Y., Hoffmann, K., and Werner, M.: Long-term winter warming trend in the Siberian Arctic during the mid- to late Holocene, Nat. Geosci., 8, 122–125, doi:10.1038/ngeo2349, 2015.

Overpeck, J., Webb, T., and Prentice, I. C.: Quantitative interpretation of fossil pollen spectra: Dissimilarity coefficients and the method of modern analogs, doi:10.1016/0033-5894(85)90074-2, 1985.

Raddatz, T. J., Reick, C. H., Knorr, W., Kattge, J., Roeckner, E., Schnur, R., Schnitzler, K. G., Wetzel, P., and Jungclaus, J.: Will the tropical land biosphere dominate the climate-carbon cycle feedback during the twenty-first century?, Clim. Dyn., 29, 565–574, doi:10.1007/s00382-007-0247-8, 2007.

Rehfeld, K. and Laepple, T.: Warmer and wetter or warmer and dryer? Observed versus simulated covariability of Holocene temperature and rainfall in Asia, Earth Planet. Sci. Lett., 436, 1–9, doi:10.1016/j.epsl.2015.12.020, 2016.

Scott, G. H.: Uniformitarianism, the uniformity of nature, and paleoecology, New Zeal. J. Geol. Geophys., 6, 510–527, doi:10.1080/00288306.1963.10420063, 1963.

Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J. O., Levis, S., Lucht, W., Sykes, M. T., Thonicke, K., and Venevsky, S.: Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model, Glob. Chang. Biol., 9, 161–185, 2003.

Telford, R. J. and Birks, H. J. B.: The secret assumption of transfer functions: Problems with spatial autocorrelation in evaluating model performance, Quat. Sci. Rev., 24, 2173–2179, doi:10.1016/j.quascirev.2005.05.001, 2005.

Telford, R. J. and Birks, H. J. B.: Evaluation of transfer functions in spatially structured environments, Quat. Sci. Rev., 28, 1309–1316, doi:10.1016/j.quascirev.2008.12.020, 2009.

Telford, R. J. and Birks, H. J. B.: A novel method for assessing the statistical significance of quantitative reconstructions inferred from biotic assemblages, Quat. Sci. Rev., 30, 1272–1278, doi:10.1016/j.quascirev.2011.03.002, 2011.

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, doi:10.5194/cp-9-859-2013, 2013.

Ter Braak, C. J., van Dobben, H., and di Bella, G.: On inferring past environmental change from species composition data by nonlinear reduced rank models., In: van Houwelingen H.C. (eds), Invited Papers, the XIIIth International Biometric Conference, The Biometric Society, Amsterdam, pp. 65-70, 1996.

Trenberth, K. E.: Relationships between precipitation and surface temperature, Geophys. Res. Lett., 32, 2–5, doi:10.1029/2005GL022760, 2005.

von Storch, H., Zorita, E., Jones, J. M., Dimitriev, Y., González-Rouco, F., and Tett, S. F. B.: Reconstructing past climate from noisy data., Science, 306, 679–82, doi:10.1126/science.1096109, 2004.

Yu, S.-Y.: Quantitative reconstruction of mid- to late-Holocene climate in NE China from peat cellulose stable oxygen and carbon isotope records and mechanistic models, The Holocene, 23, 1507–1516, doi:10.1177/0959683613496292, 2013.

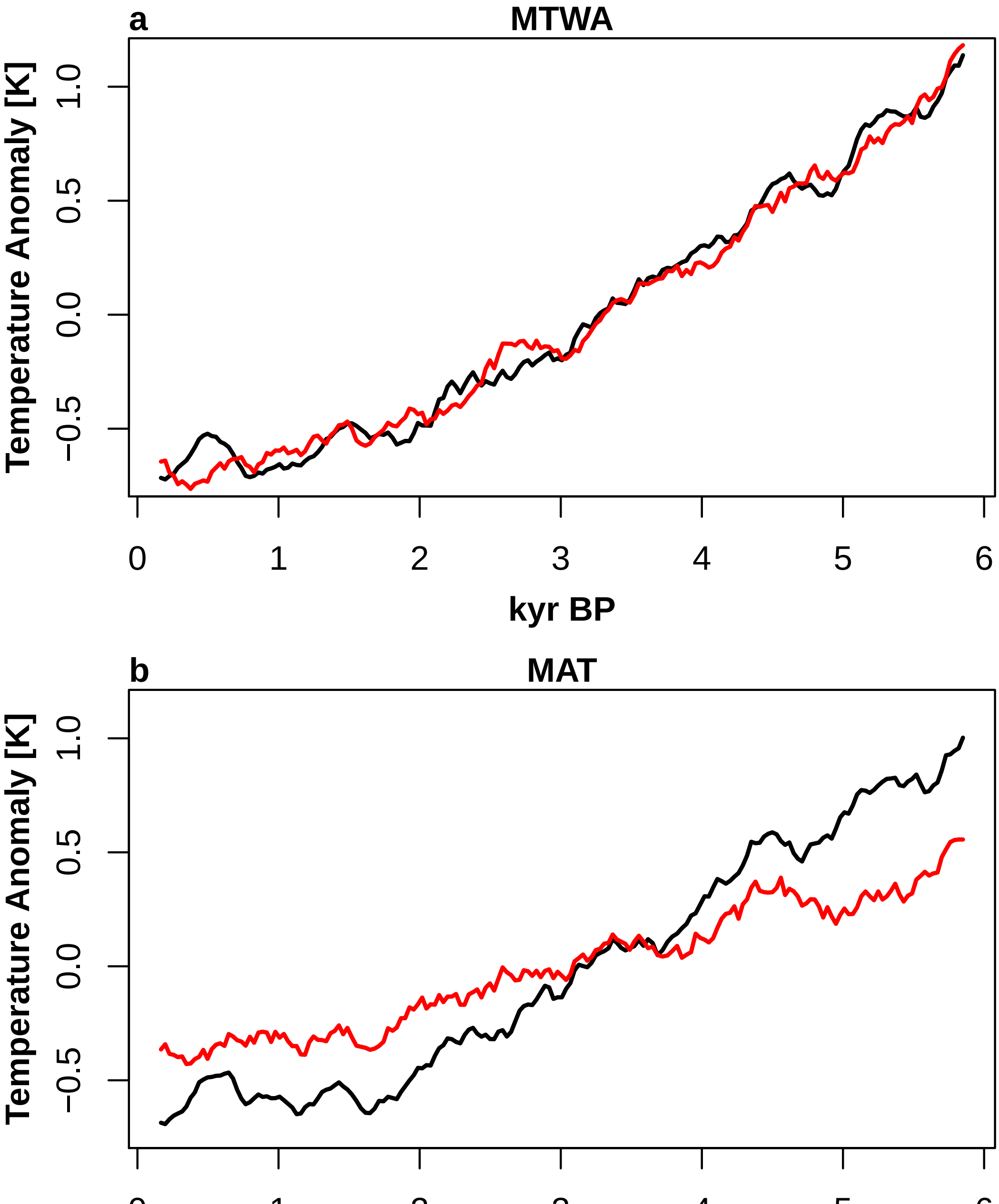


**Figure 9.** Simulated (red) and BMA-reconstructed (black) extratropical mean temperature changes over the 6k-run (BMA). The amplitude of the summer temperature trends (a) agree well, whereas the amplitude for the simulated mean annual temperature change (b) is overestimated in the reconstructions.