Effects of memory biases on variability of temperature reconstruction

Citation for published version:
Lücke, L, Hegerl, G, Schurer, A & Wilson, R 2019, 'Effects of memory biases on variability of temperature reconstruction', *Journal of Climate*. https://doi.org/10.1175/JCLI-D-19-0184.1

Digital Object Identifier (DOI):
10.1175/JCLI-D-19-0184.1

Published In:
Journal of Climate

Effects of memory biases on variability of temperature reconstructions

Lucie Lücke*

School of Geosciences, The University of Edinburgh

Gabriele Hegerl

School of Geosciences, The University of Edinburgh

Andrew Schurer

School of Geosciences, The University of Edinburgh

Rob Wilson

School of Earth & Environmental Sciences, The University of St Andrews

**Corresponding author address:* School of Geosciences, Grant Institue, The King's Buildings, West Mains Road, Edinburgh EH9 3JW, United Kindom.

E-mail: lucie.luecke@ed.ac.uk

ABSTRACT

Quantifying past climate variation and attributing its causes improves our understanding of the natural variability of the climate system. Tree-ring based proxies have provided skilfull and highly resolved reconstructions of temperature and hydroclimate of the last Millennium. However, like all proxies, they are subject to uncertainties, arising from varying data quality, coverage and reconstruction methodology. Previous studies have suggested that biological-based memory processes could cause spectral biases in climate reconstructions. This study determines the effects of such biases on reconstructed temperature variability and the resultant implications for detection and attribution studies. We find that introducing persistent memory, reflecting the spectral properties of tree-ring data, can change the variability of pseudo proxy reconstructions compared to the surrogate climate and resolve model-proxy-discrepancies. This is especially the case for proxies based on ring-width data. Such memory inflates the difference between the Medieval Climate Anomaly and the Little Ice Age, and suppresses and extends the cooling in response to volcanic eruptions. When accounting for memory effects climate model data can reproduce long-term cooling after volcanic eruptions as seen in proxy reconstructions. Results of detection and attribution studies show that signals in reconstructions as well as residual unforced variability are consistent with those in climate models when the model fingerprints reflect autoregressive memory as found in tree-rings.

1. Introduction

Long-term climate reconstructions from natural climate archives provide the basis for quantifying the full amount of natural climate variability and attributing variations to external forcings or chaotic internal fluctuations. While tree-rings provide annually resolved and precisely dated climate signal (Stokes and Smiley (1968)) and correlate well with observed temperature and precipitation records (Fritts (1976)), they are subject to a wide range of uncertainties (e.g. Fritts (1976); Esper et al. (2004); Jones et al. (2009); Cook and Pederson (2010); Frank et al. (2010a)). Here we focus on investigating the impacts of spectral biases on temperature reconstructions from tree-rings, specifically impacts on low-frequency variability and response to volcanic forcing, and their implications for detection and attribution studies.

It is well known that physiological processes within a tree can affect the climate signal and induce a biological-based memory signal (Fritts (1976); Schulman et al. (1956); Matalas (1962); Vaganov et al. (2010)). Fritts (1976) suggests that the storage of sugar and hormones as well as the growth of leaves (needles), roots and fruits could affect the persistence of the climate signal from one year to the next. Many studies have found that data based on ring width (RW) as a proxy for past temperature and precipitation contains more autocorrelation and long-term memory than data derived from maximum latewood density (MXD) (Esper et al. (2015); Franke et al. (2013); Zhang et al. (2015b); Anchukaitis et al. (2012); Krakauer and Randerson (2003); Helama et al. (2009)). It should, however, be noted that it is not clear why MXD data do not portray similar persistent properties as RW. It was observed that RW underestimates and temporally extends the response to volcanic eruptions compared to MXD (Frank et al. (2010a); D'Arrigo et al. (2013); Anchukaitis et al. (2012); Esper et al. (2015)). Franke et al. (2013) found that RW temperature records are strongly red biased compared to observations, whereas the spectral characteristics of MXD data are in better agreement with observations, although they still seem biased regarding their ratio of low- to high-frequency variability. Furthermore, they found that these biases propagate into climate field reconstructions, which display significantly more memory than observations. Zhang et al. (2015b) conducted pseudo proxy experiments in which they increased the memory in precipitation data from climate models for China. They observed that increased local scale memory propagated into the pseudo proxy reconstruction. This modified the climate variability, with additional trends at certain intervals and an overall changed frequency spectrum.

Detection and attribution studies aim to quantify the response to external forcings in reconstructions and have shown that particularly volcanism, but also greenhouse gases have a detectable influence on climate reconstructions of the last Millennium (Hegerl et al. (2007); Schurer et al. (2013a,b)). However, previous studies have not taken reconstruction method, data availability or specific proxy biases into account. Here we use pseudo proxy methods to derive fingerprints of external forcings accounting for spectral biases in the proxy reconstructions.

Pseudo proxy experiments (PPEs, Smerdon (2011)) have provided valuable insight on effects of reconstruction methods, calibration, coverage and noise properties on proxy reconstructions. Such experiments involve proxy-network-like data sampling from climate model output and applying proxy methods to derive reconstructions which can be tested in the virtual reality of the model climate. Many pseudo proxy studies have addressed data coverage, location, calibration method and influences of different noise models (e.g. Von Storch (2004); Bürger et al. (2006); Hegerl et al. (2007); Von Storch et al. (2008); Lee et al. (2008); Christiansen et al. (2009); Neukom et al. (2014)). It was found that the addition of noise is one of the most important factors influencing the performance of the different reconstruction methods. Von Storch et al. (2008) showed that adding noise to pseudo proxy data can suppress low-frequency variance of temperature anomalies in the pseudo proxy reconstructions as a consequence of regression during calibration.

In this article, we investigate potential biases in large-scale temperature reconstructions that are related to biological effects in tree-ring proxies. First we introduce our temperature datasets (section 2), followed by methods for pseudo proxy experiments, data analyses and detection and attribution in section 3. Our results are shown in section 4, where we compare the spectral properties of observational and proxy data to find a suitable statistical model for pseudo proxy experiments. Based on this we focus on suitable memory models and evaluate the performance of pseudo proxy reconstructions. Lastly, we analyze their implications on detection and attribution analyses. We discuss our results in section 5.

2. Data

a. Tree-ring data

We use tree-ring data provided by the Northern Hemisphere Tree-Ring Network Development (N-TREND) consortium as published by Wilson et al. (2016); Anchukaitis et al. (2017). This consortium is the result of a collective strategy by the dendroclimatology community to improve large-scale summer temperature reconstructions. The dataset consists of 54 tree-ring chronologies and local reconstructions, which are selected from previously published reconstructions (Table S1). Thus, the data includes informed judgments of the original authors for the most robust temperature estimates for each particular location. The individual records use different tree-ring parameters as temperature proxies, including 11 records derived from ring width (RW), 18 records from maximum latewood density (MXD) and 25 mixed records (MIX). The mixed records consist of combinations of local, regional and grid point reconstructions derived from RW, MXD and blue intensity (BI) data. BI is a relatively new method to dendroclimatology and provides similar proxy climate information to MXD (see Campbell et al. (2007); Björklund et al. (2014); Rydval et al. (2014) for more information). A detailed table showing the details of the included proxy records is given in the supplementary material.

The records cover the mid-latitudinal band between 40° N and 75° N, following the recommendation of Wilson et al. (2016), as trees further south are more sensitive to multiple climate influences (Fritts (1976); St. George (2014); St. George and Ault (2014); Osborn et al. (2000); Franke et al. (2013)). The target area is further divided into three continental scale regions (North America, Western Eurasia and Eastern Eurasia). Each region has available data covering more than 1000 years, with 23 records extending back to at least 978 A.D. All records cover the period 1710 to 1988. However the number of available records decreases markedly towards the beginning of the last Millennium, and North America relies on only three records before 1100 A.D. The individual proxy locations are shown in figure 1a.

To understand the effects of different proxy types, we slightly modify the original N-TREND dataset. We distinguish three datasets, consisting of the full network (referred to as N-TREND FULL), RW data only (N-TREND RW) and MXD records only (N-TREND MXD). Given the small number of BI data in the mixed records we exclude BI-specific biases from our analysis by removing BI data from six mixed records for which the individual records were available. From those mixed records we additionally recover the original RW and MXD chronologies and include them into N-TREND RW and N-TREND MXD to increase the size of the datasets. Table S2 lists the affected sites and which data type was extracted for the different proxy datasets. The N-TREND MXD dataset consists hence of 22 tree-ring records in total, while N-TREND RW consists of 17 records.

b. Instrumental data

The CRUTEM4 dataset (Osborn (2013)) provides instrumental data over the period 1850 to 2013. CRUTEM4 is a gridded dataset of global historical near-surface air temperature anomalies over land with a resolution of 5°. The coverage of the reconstruction target area varies and is highly depended on the location (figure 1b). Prior to 1880 coverage is largely restricted to western Europe and lower latitudes of eastern North America. In addition to poor coverage, warm biases might arise from poorly shielded instruments for early instrumental data prior to the widespread use of the Stenvenson screen (Parker (1994); Böhm et al. (2009); Frank et al. (2007)). Given the greater uncertainty (Brohan et al. (2006)) and poor data coverage, data prior to 1880 was excluded from the analysis. Even at later times the hemispheric reconstruction is clearly biased towards Europe, where we find many of the grid points covering the full calibration period. North America is well covered at lower latitudes in this period, but lacks data at higher latitudes. Coverage is worst for Asia, where most grid points do not start before 1950. This makes the early instrumental record for Asia particularly prone to biases and shifts the hemispheric record heavily to Europe and North America.

c. Climate model data

We used the Community Earth System Model Last Millennium Ensemble Project (Otto-Bliesner et al. (2016)), referred to as CESM-LME, for all model-proxy comparisons and pseudo proxy experiments. The CESM-LME uses a version of CESM-CAM5_CN (1.9x2.5_gx1v6), with a resolution of $\sim 2^{\circ}$ in atmosphere and land components and $\sim 1^{\circ}$ resolution in ocean and sea ice components. External forcings include volcanic, solar, orbital, changes in land use/land cover and greenhouse gas forcing. Forcing reconstructions follow the recommendations by the Paleoclimate Intercomparison Project Phase III (PMIP3, Braconnot et al. (2012); Schmidt et al. (2011, 2012)) and are the same as used in the last Millennium simulation of the Community Climate System Model version 4 (CCSM4, Landrum et al. (2013)). The CESM-LME provides a large range of different experiments, including all transient forcings as well as ensembles of individual forcings and control runs, covering the period 850 to 2006. For our analyses, we use an ensemble of 13 climate simulations including all forcings, 5 simulations including volcanic forcing only and 2 control simulations. To improve like-for-like comparison of model and proxy data, we use only May to August (MJJA) surface temperature data over land and within the N-TREND target area of 40 to 75° N.

3. Methods

a. Reconstruction method

Our reconstruction method follows mostly the method introduced along with the original tree-ring dataset (Wilson et al. (2016, 2007); D'Arrigo et al. (2006)), targeting northern hemispheric (NH) mid-latitudinal summer (May-August: MJJA) land surface temperature. We first standardise all data to z-scores (mean $\mu = 0$, variance $\sigma^2 = 1$) over the period 1750-1950, then apply a nesting approach to ensure that the variance is independent of the number of available records (Cook et al. (2002); Meko (1997)). Next we classify the data into forward and backward nests of common data availability. We define the most replicated nest (NEST1), which includes all records and covers the period 1710-1988. We then find the other nests by going backward/forward in time and iteratively remove shorter records. A detailed list of the forward and backward nests is given in the supplementary material.

For each nest, we calculate regionally averaged time series. To ensure even contribution from all regions we restandardise the regional timeseries over the period 1750-1950. The regions are defined as longitudinal slices of the hemispheric band as shown in figure 1, providing a time series for North America (170° W - 10° W), Western Eurasia (10° W - 80° E) and Eastern Eurasia (80° E - 170° W). This approach slightly differs from the original method, in which North America had been additionally divided along the meridian at 100° W. By doing so, we ensure that more data is available for each region. This is important when constructing timeseries for RW or MXD only, which further reduces the number of available proxy records.

We derive a hemispheric mean series $z_i(t)$ for each nest i by averaging over the regional timeseries and calibrate the result for NEST1 $z_1(t)$ to the instrumental data $T_{obs}(t)$. The calibration covers the period 1880-1988. We choose the start date to exclude poor instrumental coverage and the end date to ensure full coverage by the tree-ring network. Calibration includes matching of variance and mean (Esper et al. (2005)) of instrumental and proxy data:

$$T_1(t) = \left(z_1(t) - \mu_{z_1}\right) \cdot \frac{\sigma^2_{\text{obs}}}{\sigma^2_{z_1}} + \mu_{\text{obs}}. \quad (1)$$

The hemispheric timeseries from all other nests are scaled to $T_1(t)$, the temperature timeseries obtained from NEST1, in the same way but each over the full period of NEST1. Ultimately, a homogeneous temperature reconstruction is derived by extracting the temperature for each year from the densest nest available. Comparing the different proxy datasets (figure 1c) we find that low and short term variability varies across the datasets, with FULL and RW displaying more low frequency variability throughout the last Millennium. This is highlighted in the average temperature difference between Medieval Climate Anomaly (MCA, 950-1250 Masson-Delmotte et al. (2013)) and Little Ice Age (LIA, 1450-1850 Masson-Delmotte et al. (2013)). MXD shows a smaller difference than RW and FULL. This can also be observed when comparing differences between 20th century warming and LIA, which is consistently higher in RW than in MXD data. As discussed by Wilson et al. (2016), the N-TREND reconstruction shows little divergence (Wilson et al. (2007);

D'Arrigo et al. (2008)) from the instrumental data during the late 20th century. However to exclude potential influences of the remaining divergence we use the period 1900-1980 representative for 20th century warming. All proxy reconstructions show a similar temperature difference between the LIA and this period.

b. Reconstruction uncertainty

Quantifying and including all forms of uncertainty in tree-ring (and other proxy) climate reconstructions is a significant challenge and beyond the scope of this article. However, we can model uncertainties caused specifically by coverage and calibration relatively easily using an ensemble approach (Frank et al. (2010b); Neukom et al. (2019)). In order to be able to replicate the same reconstruction method when conducting our pseudo proxy experiments, it was important to reduce computational time and thus keep the ensemble size relatively small. To address the coverage uncertainty we apply a bootstrapping approach to the proxy dataset, in which one proxy record is removed in turn before creating the reconstruction. Although this would ideally include the removal of each proxy record in the dataset in turn, we restrict the analysis to bootstrapping nine randomly selected long records in turn, extending back to at least 1150 A.D. Thus we estimate the coverage uncertainty specifically in the poorly covered periods. The chronologies which were in turn removed from N-TREND FULL were: AG12, AG4, FORF, AG2, ALT, AG5, AG1, AG11 and FIRT. For MXD: ALT, POLx, JAEM, ALPS, FORF, TYR, FIRT, ICE and SFIN. For RW: TAT, KOL, QUEw, OZN, GOA, ICE, YAM, IDA and TAY. Including the set consisting of all available records, we gain a total ensemble of ten sets of data for each N-TREND dataset, consisting of $1 \times 54 + 9 \times 53$ records for N-TREND FULL, $1 \times 22 + 9 \times 21$ for MXD and $1 \times 17 + 9 \times 16$ for RW.

To address the calibration uncertainty, we slice the calibration period into windows of lengths 60, 70 and 80 years similar to Frank et al. (2010b). For each window length we perform the calibration for an early, middle and late period (1880-1940, 1904-1964, 1928-1988, 1880-1950, 1899-1969, 1918-1988, 1880-1960, 1894-1974 and 1908-1988). Including the full period, we thus consider ten different implementations of calibration periods, gaining a total reconstruction ensemble of 100 reconstructions for each N-TREND dataset (Full, RW and MXD). This allows us to estimate the spread of our results depending on calibration and coverage uncertainty.

c. Pseudo proxy experiments

For our pseudo proxy experiments (PPEs) we generate sets of pseudo proxy data from climate model output and treat them in the same way as real proxy data. We sample from the CESM-LME ensemble at the grid cells closest to the proxy record to match spatial and temporal availability of the N-TREND dataset as in Neukom et al. (2018). For proxy records which represent an area larger than a single grid point, the average over all grid cells within the target area was calculated. The same was repeated for CRUTEM4 to generate a pseudo instrumental dataset. The pseudo proxy data was then processed in the same way as the real proxy reconstruction, including standardising ($\mu = 0$, $\sigma = 1$), nesting, regional averaging, calibrating to the pseudo-instrumental dataset and splicing of the nested data to obtain a hemispheric pseudo reconstruction. To account for calibration and coverage uncertainty, the calibration period was varied and longer records were bootstrapped in the same way as in the case of the real proxies. The same periods and chronologies as detailed in section b were used to create a total ensemble of 1300 PPEs from the 13 CESM LME simulations and 500 PPEs from the 5 volcanic forcing only simulations.

Thus, the pseudo proxy reconstruction represents the spatiotemporal availability of the proxy network and reconstruction methods, however it does not account for any proxy specific biases or non-climatic influences. This PPE serves as the baseline to represent characteristics of local climate model data without simulating tree-ring memory. It is referred to as PPE NoM. To simulate biological-based memory we manipulate the pseudo proxy records at the local scale. Two different memory models were distinguished: a short-range autoregressive model of order p (PPE AR), and a long-term memory model (PPE LTM). To concentrate on the effects of memory, we have not added additional non-climatic white noise to the pseudo proxies. An overview of the different experiments, their ensemble sizes and fitting parameters is given in table 1.

(i) PPE AR: This memory model is based on a linear decomposition of the tree-ring signal z into a climate term and an autoregressive memory term of order p. The tree-ring signal z_t of a given year t is impacted by the locally modelled climate signal x_t. This signal is subjected to a memory term, which integrates over the previous p year's signals $z_{t-1}, z_{t-2}, \ldots z_{t-p}$. The signal at time t can thus be written as

$$z_t = x_t + \sum_{k=1}^{p} \alpha_k z_{t-k} + \varepsilon_t \tag{2}$$

$$= \sum_{k=1}^{q} \gamma_k x_{t-k} + \sum_{k=1}^{p} \alpha_k z_{t-k} + \varepsilon_t, \tag{3}$$

where ε_t accounts for additional white noise. The set of parameters α determines the influence of the k previous years' climate on the proxy signal and represents the memory term. The first term represents the climate forcing, which accounts for the autoregressive structure of the climate signal x_t itself. The autoregressive character of the climate is parametrised by the coefficients γ and its order q. If x_t represents a zero mean white noise process, equation (3) represents an auto-regressive moving average process (ARMA(p,q)). This is an autoregressive process of order p forced by a moving average process of order q (Box (2016); Von Storch and Zwiers (2002)). Assuming the climate signal of the model simulations perfectly match the real world, the climate signal x_t is given by the model data, averaged over the proxy target area. With the starting points of the time series fixed up to x_p, $z_{i>p}$ can be iteratively calculated if the memory parameters α_j are known. Instead of fitting an ARMA(p,q) process with $p+q+2$ degrees of freedom on the proxy data, we apply an empirical approach for fitting the memory. We use the knowledge of the model climate signal x and the proxy signal z to find an estimate for α_k, which produces pseudo proxies with a similar memory as seen in the proxy records.

To identify the autoregressive structure in proxy records z and model x, the partial autocorrelation function (PACF) was calculated. The PACF ϕ_k of a timeseries y at lag k determines the correlation between y_t and y_{t-k}, which is not accounted for by $y(t-1),...y(t-k+1)$. Given that the partial autocorrelation of an AR(p) process decays to zero beyond lag p we can use it to identify the order p. The coefficients ϕ_i can be calculated from the Yule-Walker equations (Box (2016)). An initial estimate for the memory coefficients α was obtained by using

$$\alpha_k = \phi_k(z) - \phi_k(x) \tag{4}$$

with the PACF $\phi_k(z)$ and $\phi_k(x)$ at lag k for the proxy record z and the targeted model data x. This was found to be a good estimate for all lags higher than lag 1. For lag 1 α was systematically overestimated by equation (4), therefore an optimization algorithm was implemented to fit the PPE to the proxy target value.

A set of fitting parameters was derived for each proxy record z in the target dataset, and the associated pseudo proxy record $\tilde{z}$ was fitted using equation (2). We set $\varepsilon = 0$, concentrating on the effects of pure memory addition. To determine whether the results are spatially robust, we randomly re-distributed the parameters α over the pseudo proxy locations. We found that the spread of results is minimal compared to the spread caused by the variation of the calibration period and bootstrapping. In order to keep the ensemble number at a reasonable size we therefore did not include this uncertainty into the final ensemble of PPEs.

(ii) PPE LTM: This method involves a manipulation of the time series in its Fourier space, which is based on a previously published study by Zhang et al. (2015a). For a timeseries possessing long-term memory (LTM) its power spectral density will decay with

$$S(f) \sim f^{-\beta}. \tag{5}$$

The parameter β is a measure of the long-term memory. For white noise processes $\beta \approx 0$, whereas for red noise $\beta = 2$. A robust estimate for β can be obtained from a detrended fluctuation analysis of the second order (DFA-2) (Peng et al. (1994); Bryce and Sprague (2012)). For a timeseries $x(t)$ with zero mean $\langle x \rangle$ the cumulative sum $X_t = \sum_{i=1}^{t}(x_i - \langle x \rangle)$ is divided into N segments with window length n. The local trend Y_t for each segment is derived from a least-squares quadratic fit of X_t. The root-mean-square deviation of X_t from the local trend for any window-length n gives the fluctuation function

$$F(n) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(X_t - Y_t)^2}. \tag{6}$$

If $F(n)$ follows a power law scaling $F(n) \sim n^{\alpha}$, the spectral density will satisfy equation (5) and

$$\beta = 2\alpha - 1. \tag{7}$$

A double logarithmic plot of the fluctuation function can provide information about the amount of LTM in a timeseries and a robust estimate for α can be calculated from a linear fit.

It was shown in previous studies that surface temperature follows a slight LTM process on both hemispheric and regional scales (e.g. Rypdal and Rypdal (2014)), with $\beta \approx 0.2$ at regional scale and $\beta \approx 0.4$ over land (Fredriksen and Rypdal (2016)). Assuming that biological tree-ring memory $y(t)$ can be represented by an LTM process which is superposed on the climate signal $x(t)$, its spectral energy can be approximated as

$$S_z(f) = S_0(f) \cdot f^{\beta_z} \approx S_x(f) \cdot f^{\beta_y} = S_0(f) \cdot f^{\beta_x + \beta_y}. \tag{8}$$

The factor $S_0(f)$ accounts for the remaining signal and represents a white noise process. Equation (8) is linear in β, which can be used to estimate the additional memory β_y and fit the pseudo proxy records

$$\tilde{S}(f) = S(f) \cdot \beta_y \qquad \beta_y = \beta_z - \beta_x. \tag{9}$$

This way a pseudo proxy record with energy spectral density $S(f)$ is fitted such that its LTM is increased to proxy level. The inverse Fourier transform of the manipulated record $\tilde{S}(f)$ gives the pseudo proxy record $\tilde{z}(t)$.

d. Superposed epoch analysis

A superposed epoch analysis is used to reveal the response to volcanic forcing evident in last Millennium temperature reconstructions (e.g. Lough and Fritts (1987); Mass and Portman (1989); Hegerl et al. (2003); D'Arrigo et al. (2013); Masson-Delmotte et al. (2013); Esper et al. (2015); Wilson et al. (2016); Neukom et al. (2018)). We average over the temperature response to a set of volcanic eruptions, using a window of maximally 30 years, considering temperature anomalies with respect to ten years preceding a volcanic eruption. Any subsequent years within the recovery time of an event which are affected by major eruptions are excluded from the epoch analysis.

We assume that the latest reconstruction of atmospheric sulfate injection (eVolv2k) as published by Toohey and Sigl (2017) minimises the dating error for the proxy reconstructions. The volcanic forcing dataset implemented in the CESM-LME is based on the IVI2 reconstruction by Gao et al. (2008). Both datasets are based on ice core data and provide a measure of aerosol optical depth (AOD) and stratospheric sulfate injection. However, dating and magnitude of volcanic eruptions in IVI2 differ in many cases from eVolv2k. In order to perform a like-for-like comparison, we therefore use eruption dates as given in eVolv2k for the proxy data, while using IVI2 dates for the model/PPE data. To increase the number of events while minimising the error induced by dating uncertainty, we consider only events which appear within three years of difference in both datasets. An overview of the volcanic forcing during the Last Millennium shown by both reconstructions of sulfur injection is given in figure 6. The 16 events included in the epoch analysis have been marked. Note that the eruptions in 1761/2 and 1783 (Laki) were excluded from the analysis despite matching dating. As noted in Stevenson et al. (2017) in the CESM-LME Laki is wrongly dated at 1761 instead of 1783, which makes both dates unsuitable for our comparison. A table showing all eruptions is given in the supplement. It should also be noted that the dating of volcanic eruptions in the climate model/PPEs follows exactly IVI2 and thus has no dating uncertainty. However due to the uncertainty in the ice core based reconstructions of volcanic forcing, some degree of dating uncertainty remains in the analysis. Nevertheless, we assume that with our approach we have kept the dating uncertainty minimal.

e. Detection and attribution studies

To quantify the influence of forced variability in the proxy reconstructions, we perform detection and attribution using a Total Least Squares (TLS) regression following (Stott et al. (2001); Allen and Tett (1999)). The proxy reconstruction $Y(t)$ is regressed onto the fingerprint of volcanic forcing $X_1(t)$ and all other forcings $X_2(t)$, following

$$Y(t) = \beta_1 \cdot (X_1(t) - \nu_1(t)) + \beta_2 \cdot (X_2(t) - \nu_2(t)) + \nu_0(t). \tag{10}$$

The fingerprints of external forcing are given by the simulations of the CESM-LME. A TLS regression allows regressor $X(t)$ and regressand $Y(t)$ to be influenced by a similar amount of noise, which is given by their respective implementation of internal variability $\nu_0(t)$. The amount of internal variability in the fingerprints $X(t)$ can be reduced by averaging over multiple ensemble members. The scaling factors β_i indicate the magnitude of the fingerprints in the reconstruction.

The response to a forcing is considered detectable ($p < 0.05$) when the scaling factor is significantly positive. A scaling factor of 1 indicates perfect agreement between models and proxy reconstruction (Hegerl and Zwiers (2011)). The residual ε gives an estimate of internal variability in the proxies. To account for the uncertainty due to internal variability and to get a distribution for the scaling factors, we follow the method introduced by (Schurer et al. (2013a,b)). We repeated our calculations 100 times with different samples of internal variability superimposed on the noise-reduced observations and model fingerprints $\tilde{Z} = [Y(t) - \nu_0(t), X_i(t) - \nu_i(t)]$. In order to investigate the effects of autocorrelation in proxy data on detection and attribution results, we further repeated our analyses using pseudo proxy fingerprints.

4. Results

a. Spectral properties of observations and model simulations compared to tree-ring data

We compare the spectral characteristics of the proxy datasets to a set of local instrumental and model records over the period 1880-1988. This period provides the maximum availability for the proxy data and is well covered by the instrumental dataset.

For the PACF at local scale (figure 2a) the biggest differences can be noted at lag 1, where RW displays a higher correlation than all other datasets. At all lags, correlation is highest for RW, followed by MXD, replicating the findings of Esper et al. (2015). Model and instrumental data agree well, with observational data showing a slightly higher correlation at all lags. The medians of the PACF at lag 1 are offset by $\Delta\alpha \approx 0.4$ for RW and MXD, which remains relatively constant during the period of common data availability (figure 2b). N-TREND MXD is slightly higher than the CESM-LME ensemble but is consistent within its 5-95% range. MXD also agrees well with the observations within the short period in which instrumental data is available. We compute the detrended fluctuation function for each record (figure 2c) to obtain an estimate for the long-term memory at local scale using equation (7). Results for all datasets are relatively widely spread but overlap at the 5-95% range. The median of MXD, observations and CESM-LME agree with $\beta \approx 0.5$, while RW proxies have slightly more memory ($\beta \approx 0.8$).

Results at hemispheric scale are similar and show that the features observed on local scale propagate into the reconstructions. The PACF (figure 2d) is still highest for RW at lag 1 while MXD is more persistent at lag 2 and 3. Modelled and observed temperatures have less PACF at these lags. Note that at lag 4 the PACF is just above the significance level for observational data and some model simulations. It is not clear whether this is a real climatic feature or sampling noise. The magnitude of the lag 1 PACF of the MXD reconstruction agrees well with the model mean (figure 2e) but RW correlation is still significantly higher during most of the period of common data availability. The magnitude of fluctuation (figure 2f) is similar for RW and MXD, however RW has more memory with $\beta \approx 0.9$ compared to $\beta \approx 0.7$ for MXD. MXD agrees well with model and instrumental data ($\beta \approx 0.7$).

Our results suggest that an autoregressive process around order 3 can be fitted to the proxy data. Given that observational and model data seem to follow mainly an order 1 process we conclude that the third order process is caused by non-climatic noise such as biological memory processes.

b. Spectral properties of pseudo proxy data compared to real proxy data

We generated pseudo proxy data for different memory models, concentrating on an autoregressive process of order 3 (PPE AR3) and a long-term memory fit (PPE LTM). We compare the partial autocorrelation of different pseudo proxy experiments with real proxy data targeting the full network, MXD only and RW only. On local scale (figure 3 a-c) correlations of PPE NoM are significantly below the range of the correlation for all targets. All pseudo proxy records including memory match the real proxy range at lag 1. At higher lags PPE LTM decays quickly below the proxy range while PPE AR3 matches the proxy records even at higher lags. At the hemispheric scale (figure 3d-f) differences between PPE AR3 and PPE LTM are smaller but PPE AR3 still performs better. Throughout the last Millennium the lag 1 partial correlation for the pseudo proxies is shifted up to proxy level (figure 3g-i) but otherwise barely deviate from PPE NoM.

All the targeted proxy reconstructions have more power at low frequencies than at high frequencies (figure 4 a-c). The power spectral density follows approximately a power-law decay for multidecadal frequencies, observed as a linear decrease in the double logarithmic plot. However the gradient flattens towards decadal frequencies, indicating a deviation from the power-law. This is particularly prominent in case of RW but can also be observed in the other datasets. The multidecadal gradient is matched by the pseudo proxy reconstructions when accounting for memory, while PPE NoM has a much smaller gradient. PPE AR3 performs well for all targets. It overlaps well with the proxy ensemble within the 5 to 95% range and its median shows the distinctive flattening of the gradient towards its high frequency end. While PPE LTM also overlaps well with the proxy ensemble within the uncertainty range, the median decreases monotonically. Note that the spectral density of MXD is particularly noisy at low frequencies (fig. S5). Since this is specific to the MXD dataset, it could be caused by local influences but could also originate from data processing.

The detrended fluctuation analysis (DFA, figure 4d-e) confirms that PPE NoM has less long-term memory than the proxies, holding particularly for RW ($\beta \approx 0.3$ vs. $\beta \approx 0.9$) and FULL ($\beta \approx 0.4$ vs. $\beta \approx 0.8$), while the difference is smaller in case of MXD ($\beta \approx 0.3$ vs. $\beta \approx 0.6$). PPE AR3 and PPE LTM both replicate the gradient of the proxy targets. While for RW and FULL the average of PPE AR3 and the proxy target overlap roughly for most time steps, the magnitude of the fluctuation of the proxies is consistently lower than the PPEs.

We conclude that PPE AR3 and PPE LTM both reproduce spectral features characteristic to proxy data, such as increased autocorrelation at lag 1, inflation (suppression) of low-frequency (high-frequency) variability and more long-term memory. PPE AR3 performs best for all target datasets as it matches the partial autocorrelation at higher lags and reproduces the deviation of the spectral density from the power-law decay at high frequencies.

c. Effects of memory on temperature variability of pseudo proxy reconstructions

The ensemble mean and range of the millennial-length timeseries for the proxy and pseudo proxy reconstructions are shown in figure 5a-c. Long term deviations from the mean are inflated for memory PPEs compared to PPE NoM. As a result, the MCA is warmer for PPE AR3 and PPE LTM, while the LIA is slightly colder. This trend can be observed in all three target datasets, but is particularly strong for FULL and RW.

To quantify the effects of this inflation, we calculate the average temperatures of MCA and LIA. The temperature difference between those periods ranges around $\Delta T = 0.2$ for FULL and RW, but is less than half for MXD (figure 5e). However, the uncertainty on the exact value is relatively high due to the small number of available records at early times. Schneider et al. (2015) found that the MCA is less pronounced in MXD data, suggesting varying seasonal or spatial coverage as a reason. However PPE NoM shows a clear warming in the MCA for the MXD locations. For all target datasets, the median of ΔT is increased when implementing memory in the pseudo proxies. For PPE AR3 the median shifts towards the proxy value in case of FULL and RW targets. The temperature difference increases further for higher memory, with PPE LTM consistently being highest. The increase of ΔT with memory order is a robust feature, which can also be seen when comparing average temperatures of the LIA and the 20th century between 1900-1980 (figure 5g-i). Note that 20th century warming is slightly underestimated in the CESM-LME, likely due to strong indirect aerosol forcing (Otto-Bliesner et al. (2016)). This could be a reason for a small temperature difference compared to the proxy value, and could suppress stronger increase for memory PPEs.

To analyze the effects of biological memory on the magnitude and timescales of cooling in response to volcanic eruptions, we perform a superposed epoch analysis (figure 7a-c) including 16 well-dated volcanic eruptions. Schneider et al. (2015) compared the volcanic response in a density only reconstruction to ring width dominated reconstructions for the eruptions in 1257, 1452 and 1815. They found that the former shows a greater response amplitude, while the latter show a temporally extended cooling and thus longer recovery period. The same observations hold for our epoch analysis. Here, MXD responds strongly and recovers fast, with a slightly prolonged cooling around year three to five. RW has a smaller amplitude along with a prolonged cooling up to post-eruption year ten. While the magnitude of the PPE NoM amplitude varies slightly across the target datasets, it recovers much quicker than the proxies. Both magnitude and recovery time are affected by autoregressive memory, most prominent for RW, while long-term memory mainly dampens the amplitude. PPE AR3 shows a prolonged cooling, which is mostly consistent with the timescale of the proxy data. The median of the peak response of the PPE AR3 ensemble is much dampened compared to PPE NoM, and even slightly lower than N-TREND. However, it is consistent with N-TREND within the 5 to 95% range.

Comparing the residuals of proxy and PPE epoch analysis (figure S2 a-c), we note that the residuals increase particularly between year three to five after the eruption. This observation holds for all PPE's and for all target datasets. To increase our understanding, we compare an ensemble member of the CESM showing a particularly prolonged recovery and persistent cooling in year four after the eruption (figure 7d-f) and one with a particularly quick and steadily decreasing recovery (7g-h). In the former case, PPE AR3 reproduces the recovery time, the peak cooling and overlaps with N-TREND for all datasets within its uncertainty range. The residuals are negligibly small five years after the eruption (figure S2d-f). In the latter case, even though the cooling is more prolonged for PPE AR3 compared to PPE NoM neither its recovery time nor its amplitude match the proxy amplitude. The residuals are near constant up to year 15 (figure S2g-h). We conclude that model and proxy output can be consistent when taking memory effects into account. Memory can explain the long recovery time observed in proxy reconstructions but requires persistent cooling on a timescale between three to five years. This short-term persistence could be caused by internal variability, but also by missing short-term feedback mechanisms in the model, e.g. changes in the North-Atlantic Oscillation (Zanchettin et al. (2013); Driscoll et al. (2012); Timmreck (2012)).

d. Effects of memory in pseudo proxies on detection and attribution

We perform detection and attribution studies for the period 1300-1710 in order to evaluate if the previously observed low amplitude of fingerprints in proxies might be due to memory effects. We chose the upper end of this period to exclude an overlap with the fitting period (1710-1988) and the lower end to ensure reasonable data quality and coverage. Additional sensitivity tests were performed for the slightly longer period 1300-1850. The proxy reconstructions served as the regression targets, while the fingerprints of external forcing were PPE versions of the all forcings and volcanic forcing only simulations (figure 8). Neither the proxy reconstruction nor fingerprints were smoothed prior to the regression. The fingerprints are most affected for the RW version of volcanic forcing only, where the temperature anomalies deviate strongly from the PPE NoM reference at certain periods.

All target datasets show increased volcanic scaling factors for PPE AR3 and PPE LTM compared to PPE NoM (figure 9a-c). This indicates that the addition of memory to the fingerprints makes the model consistent with the proxy data in case of the longer period. The highest difference between the memory PPEs and PPE NoM can be observed in the RW reconstruction. For this dataset the scaling factors for volcanic forcing are increased up to the median value $\beta = 1.5$. The scaling factors also increase with memory for FULL and MXD, however the difference to the reference PPE NoM is smaller. These observations are consistent with the results of the epoch analysis, which showed that cooling amplitudes in response to volcanic forcing are reduced. Two main observations can be made from plotting the scaled fingerprints relative to their proxy targets (figure 9d-f), which are clearly present in FULL and RW, but only weakly present in MXD. The big drop of NH temperature following the eruption in the mid-15th century is matched much better by the memory PPEs in both magnitude and length, and the same applies to the eruptions in 1600 and 1640. Low frequency variability is increased for the memory fingerprints, resulting in a better fit for RW and FULL reconstructions, which show a substantial low frequency variability between 1450 and 1600. When targeting the period 1300-1850 (figure 10) the scaling factors are slightly reduced and in all cases are consistent with one. This could be explained by overfitting the peak warmth in the 16th century in the shorter analysis (compare figures 9 and 10). Note that the longer period is also influenced by the wrong dating of Laki (1761 instead of 1783) in the CESM-LME, which could influence the results and dampen the scaling factors.

The residual variability in reconstructions not explained by the fingerprints (figure 11a-c) shows a slight decrease when accounting for memory, which is particularly prominent in the RW case. Even though the proxy uncertainty is relatively high, the ensemble median shows a clear decrease when accounting for memory. Simultaneously, the variance of the PPE control runs decreases and approaches the proxy value. Thus, the residual variability becomes consistent with the control variability for PPE AR3 and higher memory in case of FULL and RW, while for MXD it is consistent for all memory PPEs.

We conclude that models and proxy reconstructions are consistent when accounting for memory effects in RW data. This indicates better correspondence between signal amplitudes in fingerprints and reconstructions.

5. Discussion and Conclusion

The implementation of memory improved the agreement between proxy and pseudo proxy reconstructions. Ring width only reconstructions have particularly benefited, but results for the full network reconstruction including both width and density proxies were also improved. Although it has long been well known that ring width data can be successfully fitted by an autoregressive memory model (Cook et al. (2002); Meko (1997)), we find, for the first time, that implementing autoregressive memory in climate model data can introduce almost identical spectral behaviour in model data and resolve proxy-model discrepancies such as the low signal amplitude of the volcanic signal in detection and attribution studies. An autoregressive process of third order performs best out of all our memory models considered. The remarkable agreement between the spectral density of RW only proxy reconstruction and PPE AR3 suggests that even though RW has a clear spectral bias, it is sensitive to the full range of the climate signal. A similarly good agreement was found for the full network, in particular for multi-decadal timescales, when the ensemble mean agrees well with PPE AR3. As a consequence of memory biases low frequency variability is inflated while high frequency variability is suppressed. This could lead to an overestimation of the magnitude of long-term anomalies, especially for RW data. This phenomenon is robust for all three datasets, where it leads to a warmer MCA, a cooler LIA and increased warming during the 20th century in the PPEs when including memory. The effect on the amplitude of the MCA is particularly high, which could be caused by poor data coverage further exacerbating the bias. Without considering memory, MXD reconstructions are most consistent with model simulations.

MXD data shows little autocorrelation and long-term memory compared to RW and improvements when fitting memory to the PPEs are small. However, reconstructions using density only still show more autocorrelation and long term memory than observations and model simulations. It remains unclear from our results if the deviations between MXD and observations/simulations arise from biases in the signal of density proxies or in the simulation of persistence of climate signal in the CESM.

The year to year memory causes a dampened amplitude in response to volcanic forcing along with a slower recovery, particularly affecting ring width reconstructions. This confirms earlier studies (Esper et al. (2015); Franke et al. (2013); Schneider et al. (2015); Stoffel et al. (2015)). Our results from the epoch analysis tie in with Neukom et al. (2018), who found that the addition of autoregressive AR(1) noise in pseudo proxy reconstructions would slightly dampen the amplitude, but not cause a prolonged cooling. We have, for the first time, provided a memory model which can explain the dampening and the prolonged cooling in proxy reconstructions and resolve the divergence between proxy and climate model response. We have shown that autoregressive memory processes cause a significant reduction of post-eruption temperatures for several years. A particular mismatch between PPEs and proxy targets is present in all datasets after around 5 years. This could be explained by internal variability or potentially a lack of short-term feedbacks in the climate model and can be resolved by PPE AR3 for specific ensemble members.

Our results from detection and attribution studies indicate that model simulations and proxy reconstructions agree better when accounting for biological-based memory. While the scaling factors are increased, the residuals are reduced to an extent which is consistent with the model implementation of internal variability. Residuals are smallest for the full network, which is likely a result of higher data coverage, including more than twice the amount of proxy records as MXD/RW only reconstructions. Our results indicate that for both periods the influence of internal variability is low compared to forced variability. When the fingerprints account for memory effects, more forced variability can be detected in the proxy reconstructions, this concerns particularly the variability related to volcanic forcing. The magnitude of the resulting scaling factors varies across the target datasets, with smallest values in case of MXD and highest values in case of RW. This observation holds for both analysed periods. For the period 1300-1710 the scaling factor for volcanic forcing obtained from the RW target dataset is significantly higher than one, and the low-frequency variability trend during the 16th century is extremely well fitted by the scaled PPE AR3 fingerprints. This indicates a potential overfit and does not occur when extending the analysis to 1850. However the longer period includes wrongly dated volcanos in the model and thus results are not fully reliable. The persistence of the climate signal due to biological memory processes introduces a degree of smoothing to the proxy reconstructions. This could explain previous observations that using smoothed fingerprints for detection and attribution studies results in higher scaling factors than using unsmoothed fingerprints (Schurer et al. (2013a,b)).

We conclude that it would be beneficial to include ring width into proxy reconstructions, as they agree well with the climate model signal. However spectral biases have to be considered when comparing model and proxy data. While we have been focusing on tree-ring data in this analysis, it is likely that memory biases of this kind will similarly affect other biological proxy archives, and thus propagate into multi-proxy studies. It is beyond the scope of this article to analyze the exact implications on calibration of proxy data. However, our results suggest that it is beneficial for the quality of RW data to invert autoregressive models to extract the real underlying climate signal. Given the sensitivity of low frequency variability to statistical processing, we conclude that the MCA-LIA difference is not a robust measure for model performance. When comparing model and proxies, spectral biases should be taken into account. Particularly for TLS-like calculations, where model and proxy reconstructions are assumed to have a similar noise structure, it would be beneficial to take into account that certain types of proxy data might not capture high frequency variability and is subject to inflated low frequency variability.

6. Data availability

The datasets and code generated during and/or analyzed during the current study are available from the corresponding author on request.

Acknowledgments. L.L. was supported by a studentship from the Natural Environment Research Council (NERC) E3 Doctoral training partnership [grant number NE/L002558/1]. A.S. and G.H. were supported by NERC under the Belmont forum, Grant PacMedy [NE/P006752/1]. G.H. was supported by NCAS [R8/H12/83/029]. G.H. was further funded by the Wolfson Foundation and the Royal Society as a Royal Society Wolfson Research Merit Award [WM130060] holder. We acknowledge the National Center for Atmospheric Research (NCAR) for producing and making publicly available their model output. We acknowledge the Northern Hemisphere Tree-Ring Network Development (N-TREND) for providing publicly available data.

The authors declare no conflict of interests.

References

Allen, M. R., and S. F. B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Climate Dynamics*, **15 (6)**, 419–434, doi:10.1007/s003820050291.

Anchukaitis, K., and Coauthors, 2012: Tree rings and volcanic cooling. *Nature Geoscience*, **5 (12)**, 836–837, doi:10.1038/ngeo1645.

Anchukaitis, K., and Coauthors, 2017: Last millennium northern hemisphere summer temperatures from tree rings: Part ii, spatially resolved reconstructions. *Quaternary Science Reviews*,

163, 1–22, doi:10.1016/j.quascirev.2017.02.020.

Björklund, J. A., B. E. Gunnarson, K. Seftigen, J. Esper, and H. W. Linderholm, 2014: Blue intensity and density from northern fennoscandian tree rings, exploring the potential to improve summer temperature reconstructions with earlywood information. *Climate of the Past*, **10 (2)**, 877–885, doi:10.5194/cp-10-877-2014.

Böhm, R., P. D. Jones, J. Hiebl, D. Frank, M. Brunetti, and M. Maugeri, 2009: The early instrumental warm-bias: a solution for long central european temperature series 1760–2007. *Climatic Change*, **101 (1-2)**, 41–67, doi:10.1007/s10584-009-9649-4.

Box, G. E. P., 2016: *Time series analysis : forecasting and control*. Fifth edition.. ed., John Wiley & Sons, Inc., Hoboken, New Jersey.

Braconnot, P., S. P. Harrison, M. Kageyama, P. J. Bartlein, V. Masson-Delmotte, A. Abe-Ouchi, B. Otto-Bliesner, and Y. Zhao, 2012: Evaluation of climate models using palaeoclimatic data. *Nature Climate Change*, **2 (6)**, 417–424, doi:10.1038/nclimate1456.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research*, **111 (D12)**, doi:10.1029/2005jd006548.

Bryce, R. M., and K. B. Sprague, 2012: Revisiting detrended fluctuation analysis. *Scientific Reports*, **2 (1)**, doi:10.1038/srep00315.

Bürger, G., I. Fast, and U. Cubasch, 2006: Climate reconstruction by regression–32 variations on a theme. *Tellus A: Dynamic Meteorology and Oceanography*, **58 (2)**, 227–235.

Campbell, R., D. McCarroll, N. J. Loader, H. Grudd, I. Robertson, and R. Jalkanen, 2007: Blue intensity in pinus sylvestris tree-rings: developing a new palaeoclimate proxy. *The Holocene*, **17 (6)**, 821–828, doi:10.1177/0959683607080523.

Christiansen, B., T. Schmith, and P. Thejll, 2009: A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness. *Journal of Climate*, **22 (4)**, 951–976.

Cook, E. R., R. D. D'Arrigo, and M. E. Mann, 2002: A well-verified, multiproxy reconstruction of the winter north atlantic oscillation index since a.d.1400∗. *Journal of Climate*, **15 (13)**, 1754–1764.

Cook, E. R., and N. Pederson, 2010: Uncertainty, emergence, and statistics in dendrochronology. *Dendroclimatology*, Springer Netherlands, 77–112, doi:10.1007/978-1-4020-5725-0_4.

D'Arrigo, R., R. Wilson, and K. J. Anchukaitis, 2013: Volcanic cooling signal in tree ring temperature records for the past millennium. *Journal of Geophysical Research: Atmospheres*, **118 (16)**, 9000–9010, doi:10.1002/jgrd.50692.

D'Arrigo, R., R. Wilson, and G. Jacoby, 2006: On the long-term context for late twentieth century warming. *Journal of Geophysical Research*, **111 (D3)**, doi:10.1029/2005jd006352.

D'Arrigo, R., R. Wilson, B. Liepert, and P. Cherubini, 2008: On the 'divergence problem' in northern forests: A review of the tree-ring evidence and possible causes. *Global and Planetary Change*, **60 (3-4)**, 289–305, doi:10.1016/j.gloplacha.2007.03.004.

Driscoll, S., A. Bozzo, L. J. Gray, A. Robock, and G. Stenchikov, 2012: Coupled model intercomparison project 5 (CMIP5) simulations of climate following volcanic eruptions. *Journal of Geophysical Research: Atmospheres*, **117 (D17)**, n/a–n/a, doi:10.1029/2012jd017607.

Esper, J., D. C. Frank, and R. J. S. Wilson, 2004: Climate reconstructions: Low-frequency ambition and high-frequency ratification. *Eos, Transactions American Geophysical Union*, **85 (12)**, 113, doi:10.1029/2004eo120002.

Esper, J., L. Schneider, J. E. Smerdon, B. R. Schöne, and U. Büntgen, 2015: Signals and memory in tree-ring width and density data. *Dendrochronologia*, **35**, 62–70, doi:10.1016/j.dendro.2015.07.001.

Esper, J., R. J. Wilson, D. C. Frank, A. Moberg, H. Wanner, and J. Luterbacher, 2005: Climate: past ranges and future changes. *Quaternary Science Reviews*, **24 (20-21)**, 2164–2166, doi:10.1016/j.quascirev.2005.07.001.

Frank, D., U. Büntgen, R. Böhm, M. Maugeri, and J. Esper, 2007: Warmer early instrumental measurements versus colder reconstructed temperatures: shooting at a moving target. *Quaternary Science Reviews*, **26 (25-28)**, 3298–3310, doi:10.1016/j.quascirev.2007.08.002.

Frank, D., J. Esper, E. Zorita, and R. Wilson, 2010a: A noodle, hockey stick, and spaghetti plate: a perspective on high-resolution paleoclimatology. *Wiley Interdisciplinary Reviews: Climate Change*, **1 (4)**, 507–516, doi:10.1002/wcc.53.

Frank, D. C., J. Esper, C. C. Raible, U. Büntgen, V. Trouet, B. Stocker, and F. Joos, 2010b: Ensemble reconstruction constraints on the global carbon cycle sensitivity to climate. *Nature*, **463 (7280)**, 527–530, doi:10.1038/nature08769.

Franke, J., D. Frank, C. C. Raible, J. Esper, and S. Brönnimann, 2013: Spectral biases in tree-ring climate proxies. *Nature Climate Change*, **3 (4)**, 360–364, doi:10.1038/nclimate1816.

Fredriksen, H.-B., and K. Rypdal, 2016: Spectral characteristics of instrumental and climate model surface temperatures. *Journal of Climate*, **29 (4)**, 1253–1268, doi:10.1175/jcli-d-15-0457.1.

Fritts, H. C., 1976: *Tree rings and climate*. Academic Press, London.

Gao, C., A. Robock, and C. Ammann, 2008: Volcanic forcing of climate over the past 1500 years: An improved ice core-based index for climate models. *Journal of Geophysical Research*, **113 (D23)**, doi:10.1029/2008jd010239.

Hegerl, G., and F. Zwiers, 2011: Use of models in detection and attribution of climate change. *Wiley Interdisciplinary Reviews: Climate Change*, **2 (4)**, 570–591, doi:10.1002/wcc.121.

Hegerl, G. C., T. J. Crowley, M. Allen, W. T. Hyde, H. N. Pollack, J. Smerdon, and E. Zorita, 2007: Detection of human influence on a new, validated 1500-year temperature reconstruction. *Journal of Climate*, **20 (4)**, 650–666, doi:10.1175/jcli4011.1.

Hegerl, G. C., T. J. Crowley, S. K. Baum, K.-Y. Kim, and W. T. Hyde, 2003: Detection of volcanic, solar and greenhouse gas signals in paleo-reconstructions of northern hemispheric temperature. *Geophysical Research Letters*, **30 (5)**, n/a–n/a, doi:10.1029/2002gl016635.

Helama, S., N. G. Makarenko, L. M. Karimova, O. A. Kruglun, M. Timonen, J. Holopainen, J. Meriláinen, and M. Eronen, 2009: Dendroclimatic transfer functions revisited: Little ice age and medieval warm period summer temperatures reconstructed using artificial neural networks and linear algorithms. *Annales Geophysicae*, **27 (3)**, 1097–1111, doi:10.5194/angeo-27-1097-2009.

Jones, P., and Coauthors, 2009: High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. *The Holocene*, **19 (1)**, 3–49, doi:10.1177/0959683608098952.

Krakauer, N. Y., and J. T. Randerson, 2003: Do volcanic eruptions enhance or diminish net primary production? evidence from tree rings. *Global Biogeochemical Cycles*, **17 (4)**, n/a–n/a, doi: 10.1029/2003gb002076.

Landrum, L., B. L. Otto-Bliesner, E. R. Wahl, A. Conley, P. J. Lawrence, N. Rosenbloom, and H. Teng, 2013: Last millennium climate and its variability in CCSM4. *Journal of Climate*, **26 (4)**, 1085–1111, doi:10.1175/jcli-d-11-00326.1.

Lee, T. C., F. W. Zwiers, and M. Tsao, 2008: Evaluation of proxy-based millennial reconstruction methods. *Climate Dynamics*, **31 (2-3)**, 263–281.

Lough, J. M., and H. C. Fritts, 1987: An assessment of the possible effects of volcanic eruptions on north american climate using tree-ring data, 1602 to 1900 a.d. *Climatic Change*, **10 (3)**, 219–239, doi:10.1007/bf00143903.

Mass, C. F., and D. A. Portman, 1989: Major volcanic eruptions and climate: A critical evaluation. *Journal of Climate*, **2 (6)**, 566–593, doi:10.1175/1520-0442(1989)002⟨0566:mveaca⟩2.0.co;2.

Masson-Delmotte, V., and Coauthors, 2013: Information from paleoclimate archives. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley, Eds., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Matalas, N. C., 1962: Statistical properties of tree ring data. *International Association of Scientific Hydrology. Bulletin*, **7 (2)**, 39–47, doi:10.1080/02626666209493254.

Meko, D., 1997: Dendroclimatic reconstruction with time varying predictor subsets of tree indices. *Journal of Climate*, **10 (4)**, 687–696.

Neukom, R., A. P. Schurer, N. J. Steiger, and G. C. Hegerl, 2018: Possible causes of data model discrepancy in the temperature history of the last millennium. *Scientific Reports*, **8 (1)**, doi: 10.1038/s41598-018-25862-2.

Neukom, R., and Coauthors, 2014: Inter-hemispheric temperature variability over the past millennium. *Nature Climate Change*, **4 (5)**, 362–367, doi:10.1038/nclimate2174.

Neukom, R., and Coauthors, 2019: Consistent multidecadal variability in global temperature reconstructions and simulations over the common era. *Nature Geoscience*, doi:10.1038/s41561-019-0400-0.

Osborn, T., 2013: Crutem4.2.0.0-2013-03: Climatic research unit (cru) gridded dataset of global historical near-surface air temperature anomalies over land (version 4.2.0.0 jan. 1850 - mar.2013). NERC British Atmospheric Data Centre, doi: 10.5285/eecba94f-62f9-4b7c-88d3-482f2c93c468.

Osborn, T. J., K. R. Briffa, and Coauthors, 2000: Revisiting timescale-dependent reconstruction of climate from tree-ring chronologies. *Dendrochronologia*, **18**, 9–25.

Otto-Bliesner, B. L., and Coauthors, 2016: Climate variability and change since 850 CE: An ensemble approach with the community earth system model. *Bulletin of the American Meteorological Society*, **97 (5)**, 735–754, doi:10.1175/bams-d-14-00233.1.

Parker, D. E., 1994: Effects of changing exposure of thermometers at land stations. *International Journal of Climatology*, **14 (1)**, 1–31, doi:10.1002/joc.3370140102.

Peng, C.-K., S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, 1994: Mosaic organization of DNA nucleotides. *Physical Review E*, **49 (2)**, 1685–1689, doi:10.1103/physreve.49.1685.

Rydval, M., L.-Å. Larsson, L. McGlynn, B. E. Gunnarson, N. J. Loader, G. H. Young, and R. Wilson, 2014: Blue intensity for dendroclimatology: Should we have the blues? experiments from scotland. *Dendrochronologia*, **32 (3)**, 191–204, doi:10.1016/j.dendro.2014.04.003.

Rypdal, M., and K. Rypdal, 2014: Long-memory effects in linear response models of earth's temperature and implications for future global warming. *Journal of Climate*, **27 (14)**, 5240–5258, doi:10.1175/jcli-d-13-00296.1.

Schmidt, G. A., and Coauthors, 2011: Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.0). *Geoscientific Model Development*, **4 (1)**, 33–45, doi:10.5194/gmd-4-33-2011.

Schmidt, G. A., and Coauthors, 2012: Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.1). *Geoscientific Model Development*, **5 (1)**, 185–191, doi:10.5194/gmd-5-185-2012.

Schneider, L., J. E. Smerdon, U. Büntgen, R. J. S. Wilson, V. S. Myglan, A. V. Kirdyanov, and J. Esper, 2015: Revising midlatitude summer temperatures back to a.d. 600 based on a wood density network. *Geophysical Research Letters*, **42 (11)**, 4556–4562, doi:10.1002/2015gl063956.

Schulman, E., and Coauthors, 1956: Dendroclimatic changes in semiarid america. *Dendroclimatic changes in semiarid America.*

Schurer, A. P., G. C. Hegerl, M. E. Mann, S. F. B. Tett, and S. J. Phipps, 2013a: Separating forced from chaotic climate variability over the past millennium. *Journal of Climate*, **26 (18)**, 6954–6973, doi:10.1175/jcli-d-12-00826.1.

Schurer, A. P., S. F. B. Tett, and G. C. Hegerl, 2013b: Small influence of solar variability on climate over the past millennium. *Nature Geoscience*, **7 (2)**, 104–108, doi:10.1038/ngeo2040.

Smerdon, J. E., 2011: Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments. *Wiley Interdisciplinary Reviews: Climate Change*, **3 (1)**, 63–77, doi: 10.1002/wcc.149.

St. George, S., 2014: An overview of tree-ring width records across the northern hemisphere. *Quaternary Science Reviews*, **95**, 132–150, doi:10.1016/j.quascirev.2014.04.029.

St. George, S., and T. R. Ault, 2014: The imprint of climate within northern hemisphere trees. *Quaternary Science Reviews*, **89**, 1–4, doi:10.1016/j.quascirev.2014.01.007.

Stevenson, S., J. T. Fasullo, B. L. Otto-Bliesner, R. A. Tomas, and C. Gao, 2017: Role of eruption season in reconciling model and proxy responses to tropical volcanism. *Proceedings of the National Academy of Sciences*, **114 (8)**, 1822–1826, doi:10.1073/pnas.1612505114.

Stoffel, M., and Coauthors, 2015: Estimates of volcanic-induces cooling in the nothern hemisphere over the past 1500 years. *Nature Geoscience*, **8 (10)**, 784–788, doi:10.1038/ngeo2526.

Stokes, M. A., and T. L. Smiley, 1968: Tree-ring dating. *Tree-ring dating*.

Stott, P. A., S. F. B. Tett, G. S. Jones, M. R. Allen, W. J. Ingram, and J. F. B. Mitchell, 2001: Attribution of twentieth century temperature change to natural and anthropogenic causes. *Climate Dynamics*, **17 (1)**, 1–21, doi:10.1007/pl00007924.

Timmreck, C., 2012: Modeling the climatic effects of large explosive volcanic eruptions. *Wiley Interdisciplinary Reviews: Climate Change*, **3 (6)**, 545–564, doi:10.1002/wcc.192.

Toohey, M., and M. Sigl, 2017: Volcanic stratospheric sulfur injections and aerosol optical depth from 500 BCE to 1900 CE. *Earth System Science Data*, **9 (2)**, 809–831, doi: 10.5194/essd-9-809-2017.

Vaganov, E. A., K. J. Anchukaitis, and M. N. Evans, 2010: How well understood are the processes that create dendroclimatic records? a mechanistic model of the climatic control on conifer tree-ring growth dynamics. *Dendroclimatology*, Springer Netherlands, 37–75, doi: 10.1007/978-1-4020-5725-0_3.

Von Storch, H., 2004: Reconstructing past climate from noisy data. *Science*, **306 (5696)**, 679–682, doi:10.1126/science.1096109.

Von Storch, H., E. Zorita, and F. González-Rouco, 2008: Assessment of three temperature reconstruction methods in the virtual reality of a climate simulation. *International Journal of Earth Sciences*, **98 (1)**, 67–82, doi:10.1007/s00531-008-0349-5.

Von Storch, H., and F. W. Zwiers, 2002: *Statistical Analysis in Climate Research*. CAMBRIDGE UNIVERSITY PRESS.

Wilson, R., and Coauthors, 2007: A matter of divergence: Tracking recent warming at hemispheric scales using tree ring data. *Journal of Geophysical Research*, **112 (D17)**, doi:10.1029/2006jd008318.

Wilson, R., and Coauthors, 2016: Last millennium northern hemisphere summer temperatures from tree rings: Part i: The long term context. *Quaternary Science Reviews*, **134**, 1–18, doi: 10.1016/j.quascirev.2015.12.005.

Zanchettin, D., C. Timmreck, O. Bothe, S. J. Lorenz, G. Hegerl, H.-F. Graf, J. Luterbacher, and J. H. Jungclaus, 2013: Delayed winter warming: A robust decadal response to strong tropical volcanic eruptions? *Geophysical Research Letters*, **40 (1)**, 204–209, doi:10.1029/2012gl054403.

Zhang, H., N. Yuan, J. Esper, J. Werner, E. Xoplaki, U. Büntgen, K. Treydte, and J. Luterbacher, 2015a: Modified climate with long term memory in tree ring proxies. *Environmental Research Letters*, **10 (8)**, 084 020, doi:10.1088/1748-9326/10/8/084020.

Zhang, P., H. W. Linderholm, B. E. Gunnarson, J. Björklund, and D. Chen, 2015b: 1200 years of warm-season temperature variability in central fennoscandia inferred from tree-ring density. *Climate of the Past Discussions*, **11 (1)**, 489–519, doi:10.5194/cpd-11-489-2015.

LIST OF TABLES

TABLE 1. Ensemble sizes for N-TREND and PPEs, each applying to the FULL, RW and MXD target dataset.

Name	fitting parameter	calibration	coverage	simulations	total
N-TREND	-	1+9	1+9	-	100
PPE NoM	-	1+9	1+9	13	1300
PPE AR3	α_1, α_2, α_3	1+9	1+9	13	1300
PPE LTM	β	1+9	1+9	13	1300
PPE NoM- VOLC	-	1+9	1+9	5	500
PPE AR3- VOLC	α_1, α_2, α_3	1+9	1+9	13	500
PPE LTM- VOLC	β	1+9	1+9	13	500
PPE NoM- CTRL	-	1+9	1+9	2	200
PPE AR3- CTRL	α_1, α_2, α_3	1+9	1+9	2	200
PPE LTM- CTRL	β	1+9	1+9	2	200

LIST OF FIGURES

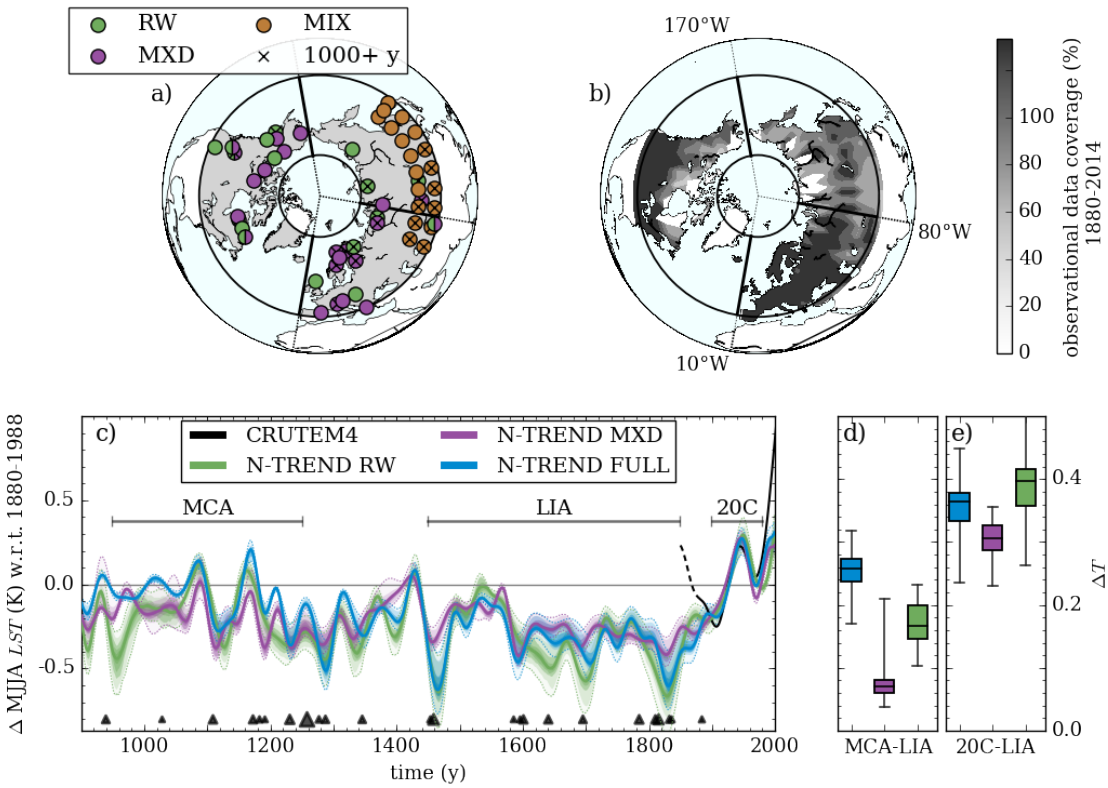


FIG. 1. a) N-TREND2015 dataset. b) Percentage of instrumental data coverage between 1880-2014 within the reconstruction target area. c) FULL, RW and MXD reconstruction ensembles. The median is shown as a solid line, with the 5th to 95th percentile indicated by a thin dotted line. Shading indicates the percentiles (55th to 95th in steps of 10). Instrumental data prior to 1880 is excluded from the analysis due to high uncertainty (dashed). All timeseries were smoothed by a 20 years smoothing spline for visualisation purposes. Triangles indicate years of volcanic activity and are scaled according to eruption magnitude (Toohey and Sigl (2017)). d) Difference of average temperature of Medieval Climate Anomaly (MCA: 950-1250) and Little Ice Age (LIA: 1450-1850) and e) 20th century (20C: 1900-1980) and LIA. Boxes range from upper to lower quartile, whiskers indicate the 5th to 95th percentile, solid line the median.

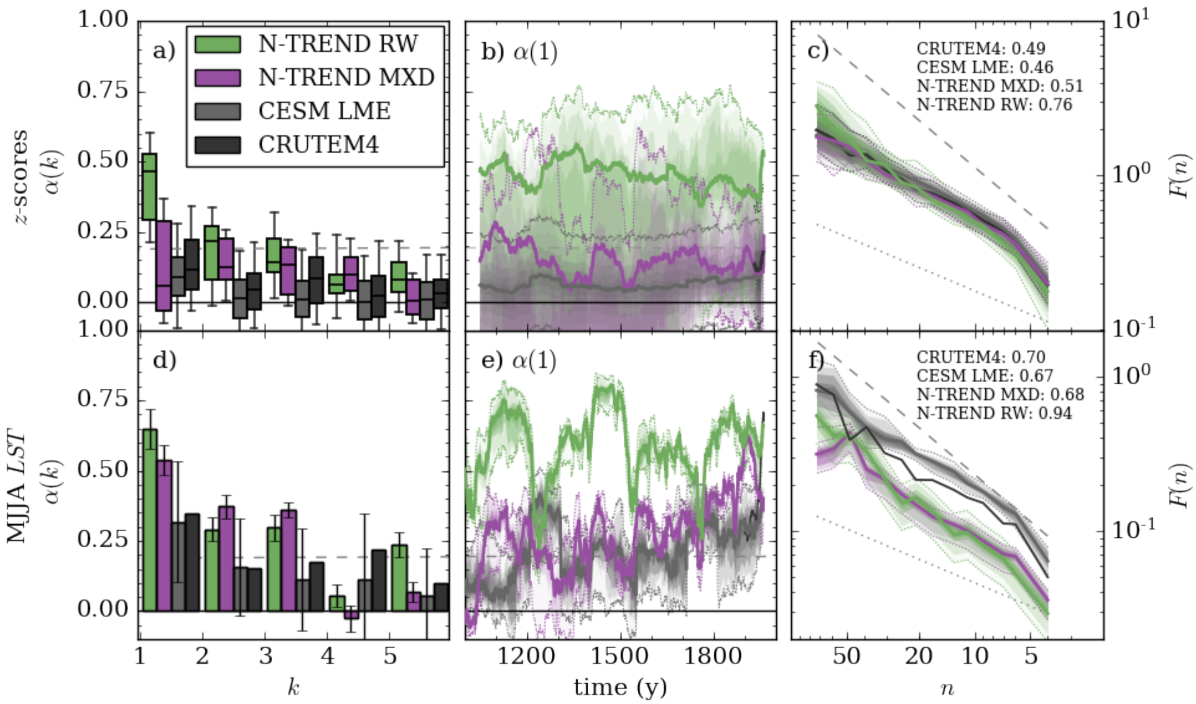
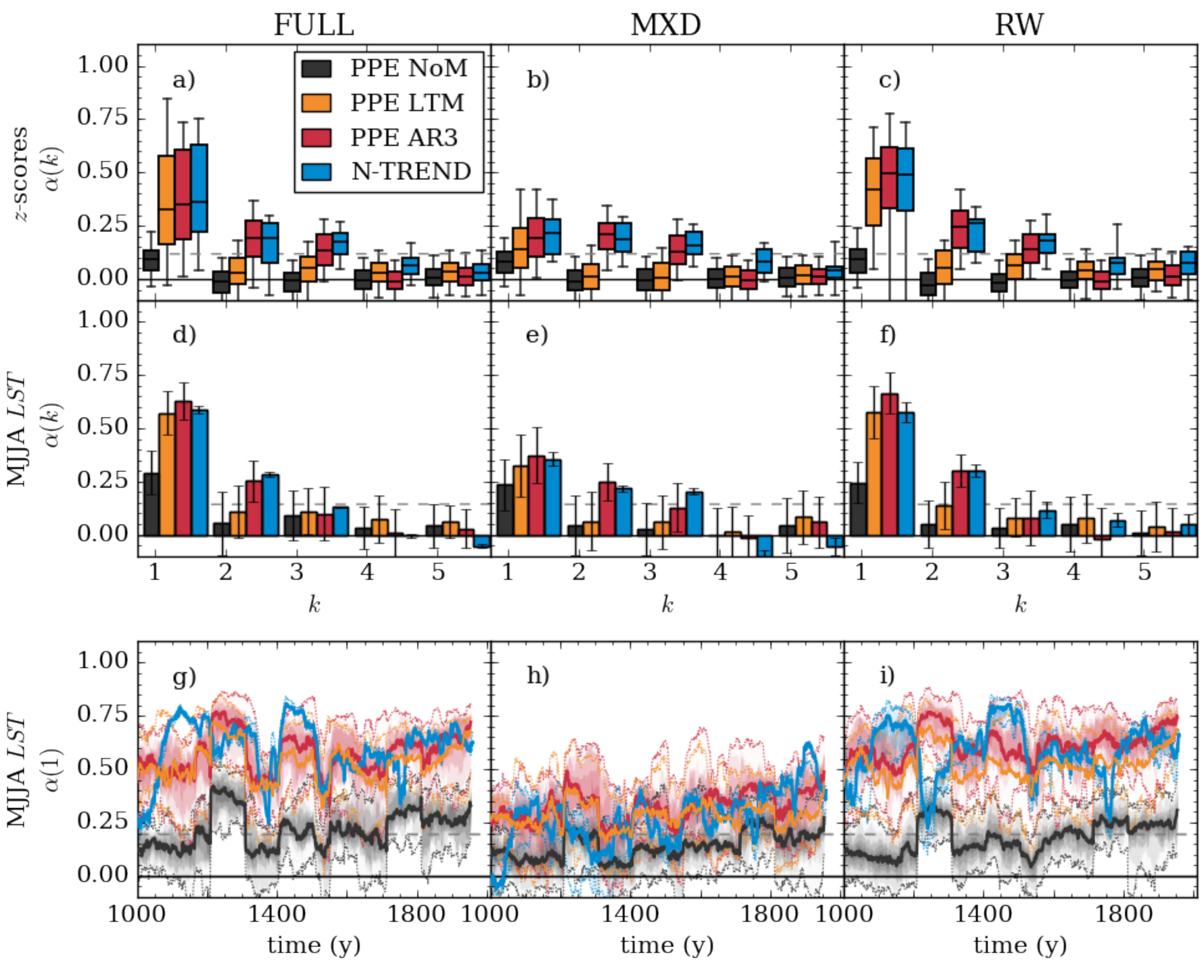


FIG. 2. a) Partial autocorrelation (PACF) $\alpha(k)$ during the calibration period (1880-1988) for local standardised records (z-scores). b) Median of PACF at lag 1 and percentile range (shaded) of the z-scores, calculated over a centered 100 years sliding window during the last Millennium (1000-2000). c) Detrended fluctuation analysis of the z-scores during the calibration period. Dotted (dashed) lines indicate the gradient displayed by white (pink) noise. d-e) as a-c) but for mean of hemispheric temperature reconstructions (bars indicate the 5th to 95th percentile of ensembles in d- note that the CESM includes 13 simulations and has a much higher spread accordingly).



863 FIG. 3. a-c) PACF between 1000-1900 A.D. for proxy z-scores (blue) and pseudo proxy experiments (PPEs) on local scale. d-f) PACF of hemispheric temperature reconstruction for the same period. g-i) 100y running mean of PACF at lag 1.

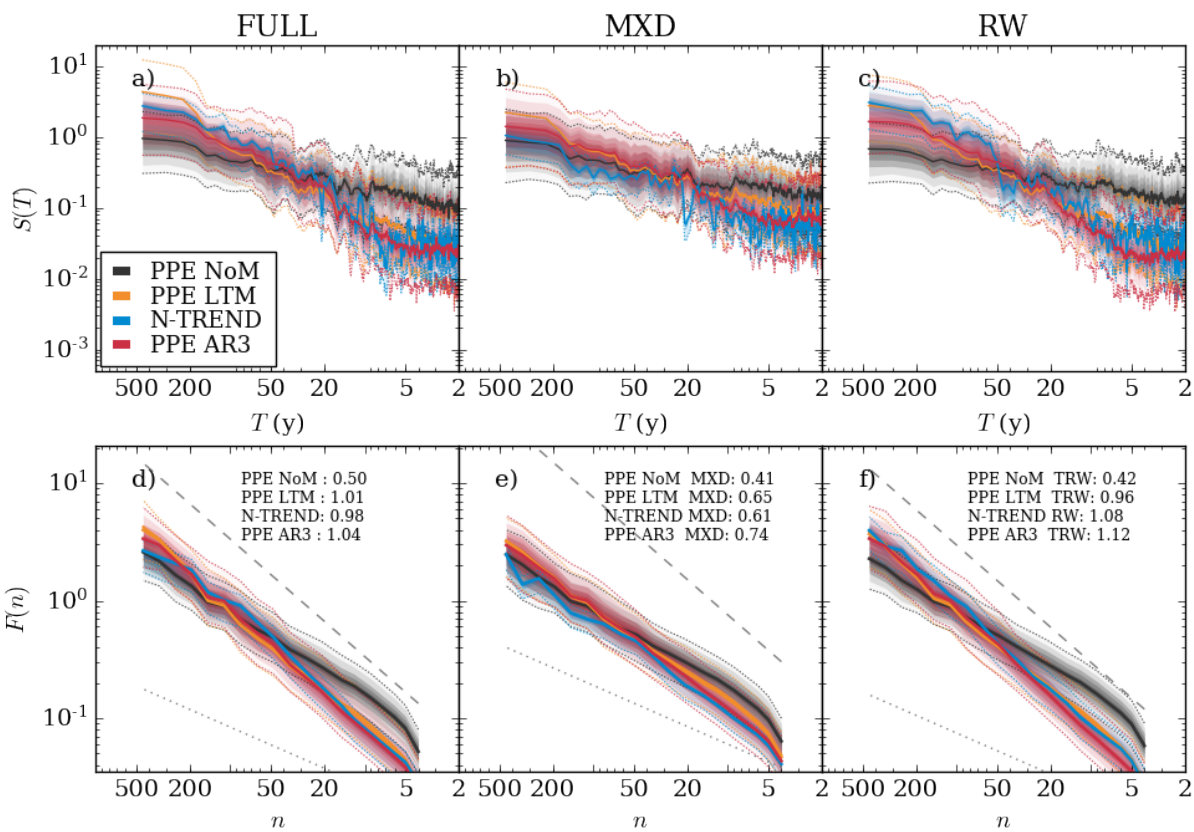
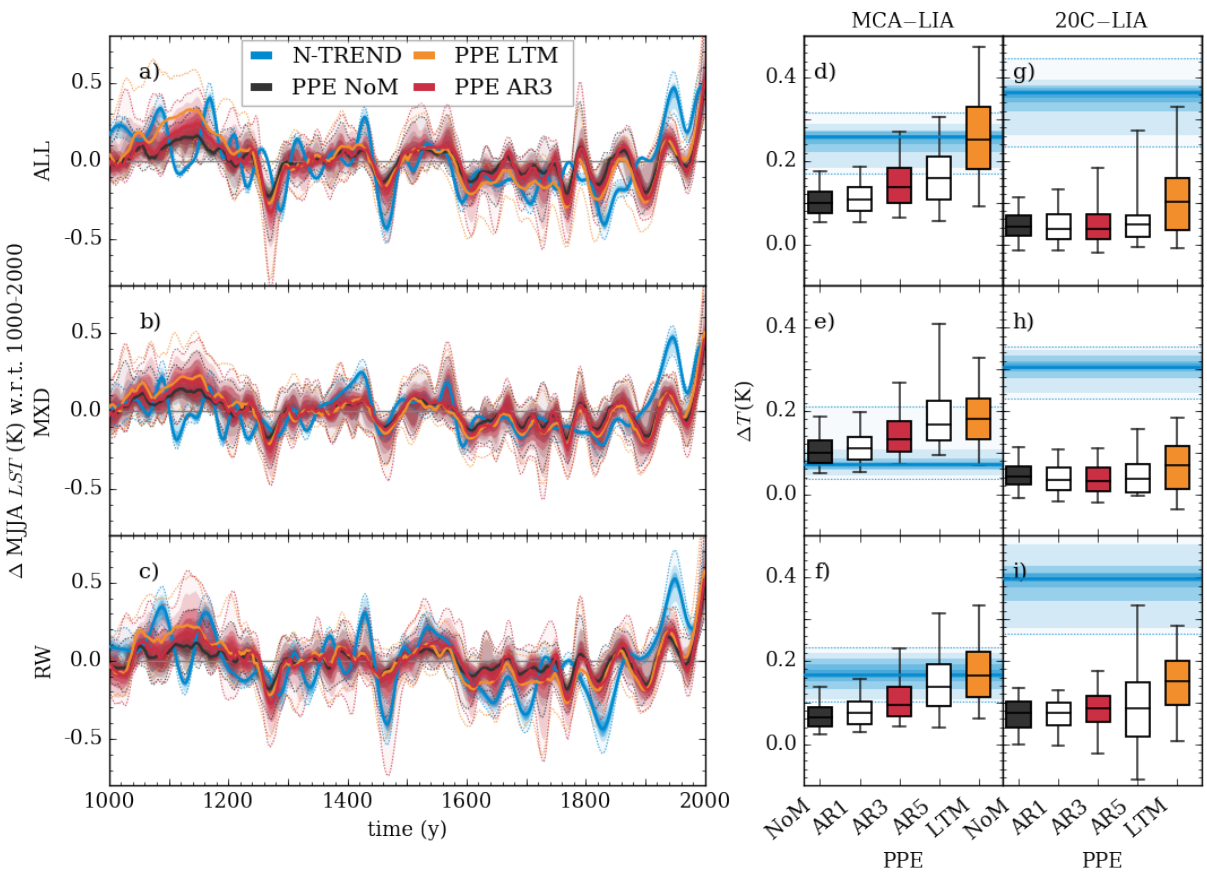


FIG. 4. a-c) Median and percentile range of the power spectral density $S(T)$ of proxy reconstructions compared to the PPEs, with ensemble range for PPE NoM and PPE AR3. The spectrum has been smoothed using a 7 year running mean filter to increase the visibility of the trend. d-e) Detrended fluctuation analysis $F(n)$ for proxy and pseudo proxy reconstructions. Dotted and dashed lines indicate the gradient displayed by white ($\beta = 0$) and pink noise ($\beta = 1$).



871 FIG. 5. a-c) Reconstructions of temperature anomalies during the Last Millennium displayed by real proxies and PPEs. Shading as in previous figures. d-f) Difference between average temperature of Medieval Climate Anomaly (MCA, 950-1250) and Little Ice Age (LIA, 1450-1850). g-h) Difference between average temperature of Little Ice Age and 20th century (20C, 1900-1980). Blue horizontal lines and shading indicate median and percentiles of the proxy reconstruction. Boxplots as in previous figures.

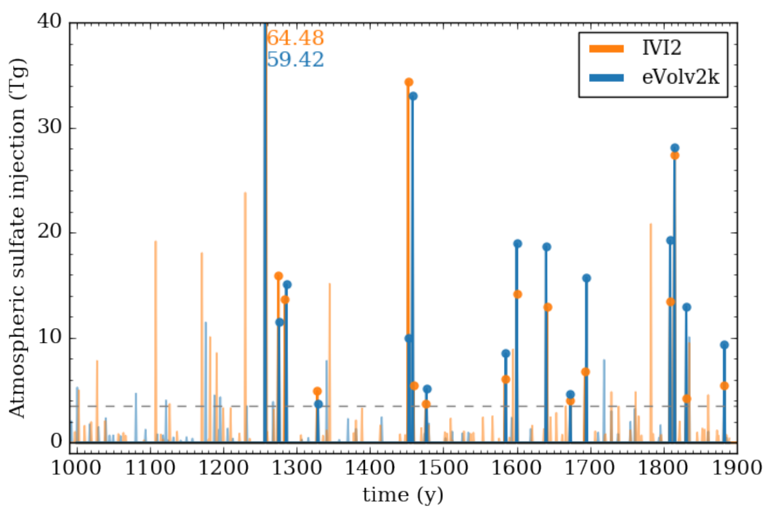


FIG. 6. Overview over atmospheric sulfate injection as in IVI2 (Gao et al. (2008)) and eVolv2k (Toohey and Sigl (2017)). Events chosen for the proxy (PPE) epoch analysis are highlighted and marked by a blue (orange) dot.

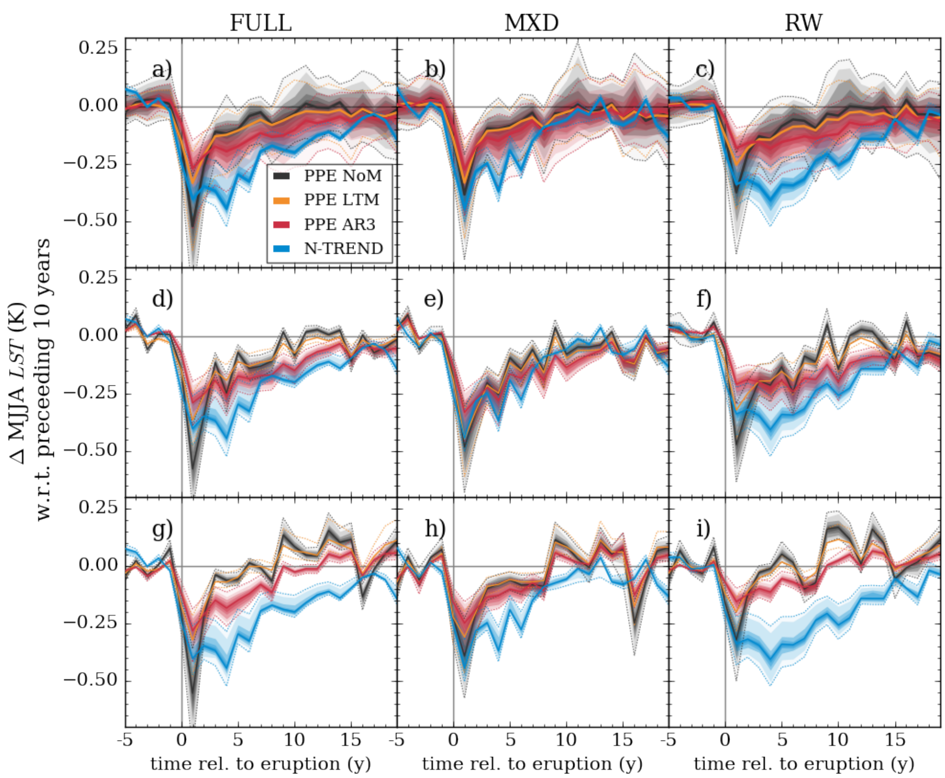


FIG. 7. Superposed epoch analysis for 16 well-dated volcanic eruptions between 1000-1900. a-c) Full ensemble range. Shading as in previous figures. d-f) Best matching ensemble member including reconstruction uncertainty (shaded). g-i) Poorly matching ensemble member.

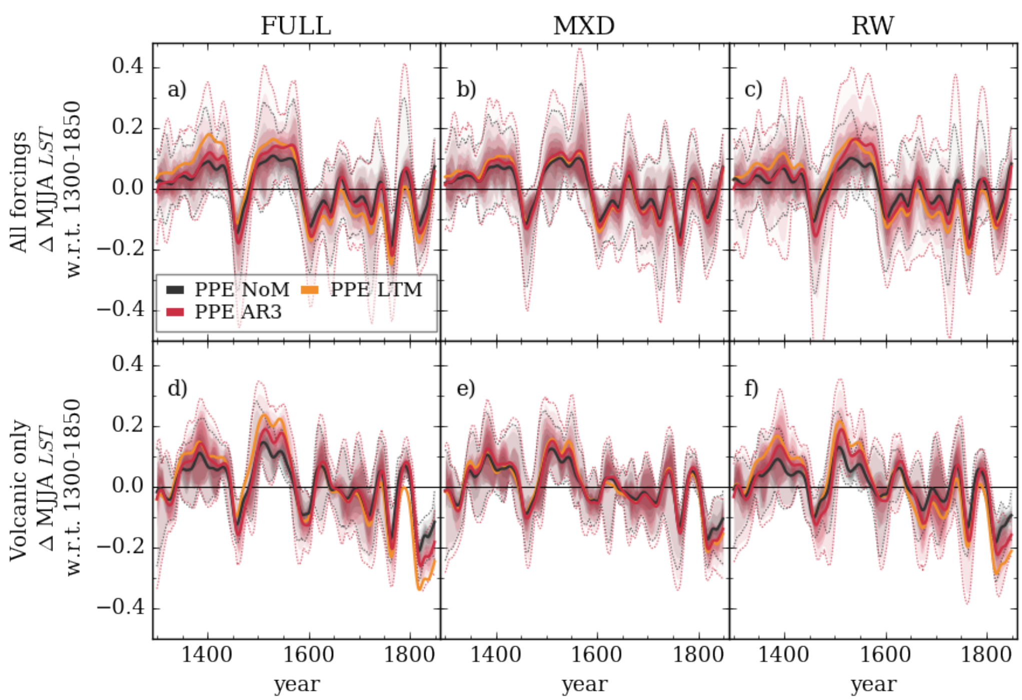


FIG. 8. Pseudoproxy fingerprints of external forcings for the PPE ensembles targeting the full, MXD and RW only network. Red/black shading indicates the percentiles of the PPE NoM and PPE AR3 ensemble. Fingerprints are smoothed using a 20 years running mean filter for visualisation purposes.

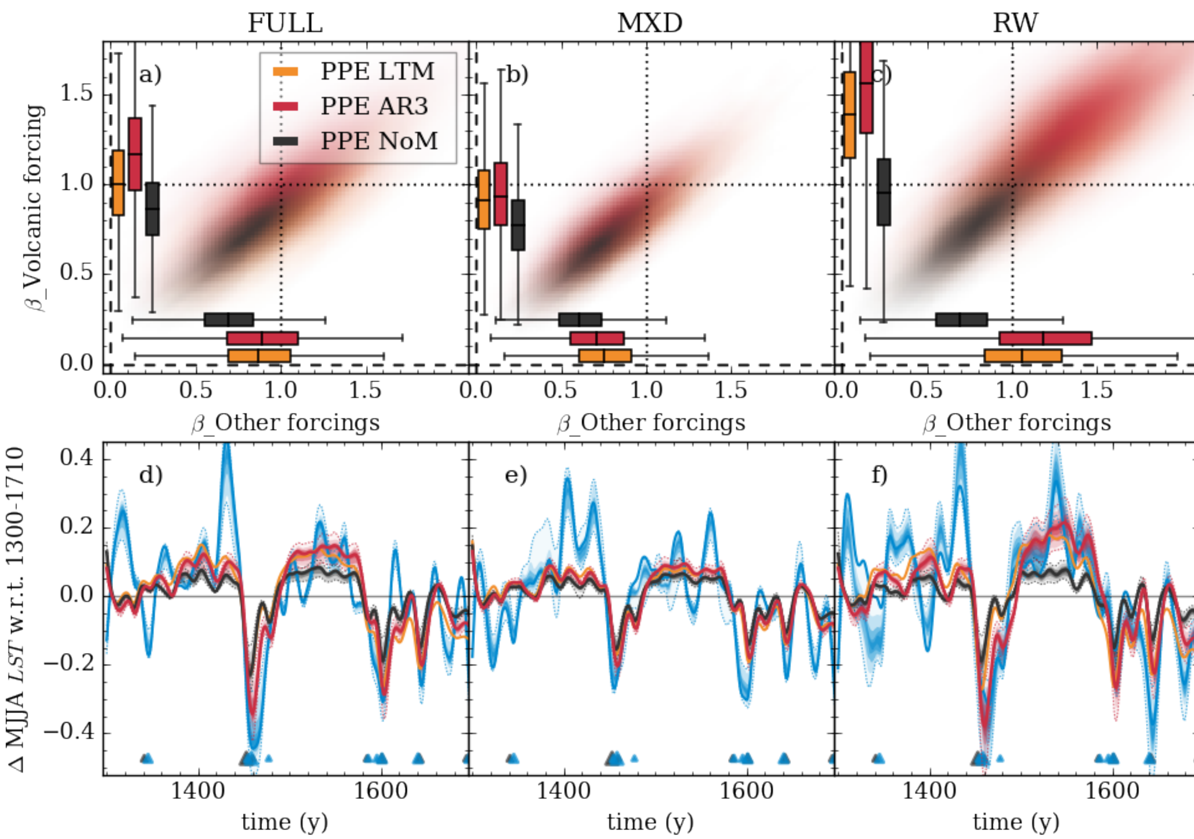


FIG. 9. Results for D&A targeting the period 1300-1710. a-c) Scaling factors. Boxplots indicate the distribution of the scaling factors (box: lower and upper quartile, line: median, whiskers: 5th to 95th percentile). d-f) Scaled PPE fingerprints against targeted proxy reconstruction (blue) during the regression period smoothed with a 15y lowpass filter.

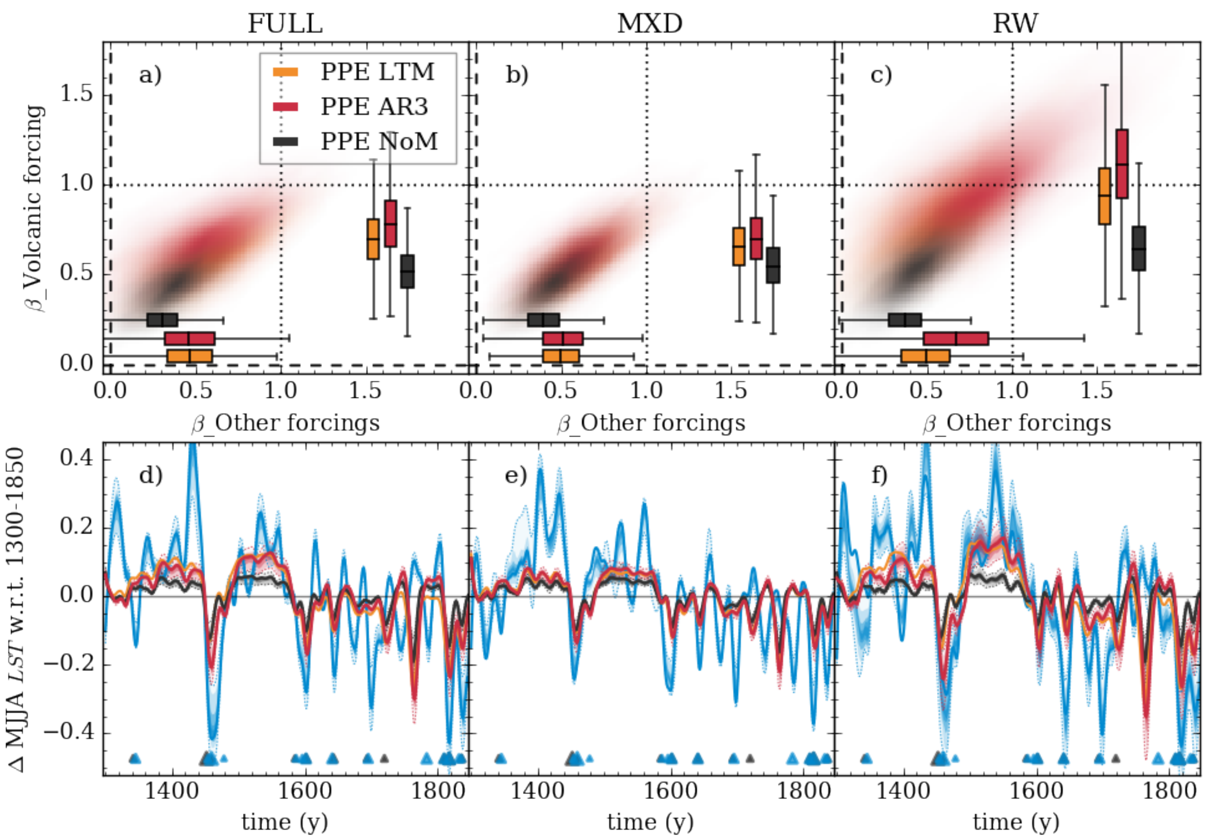
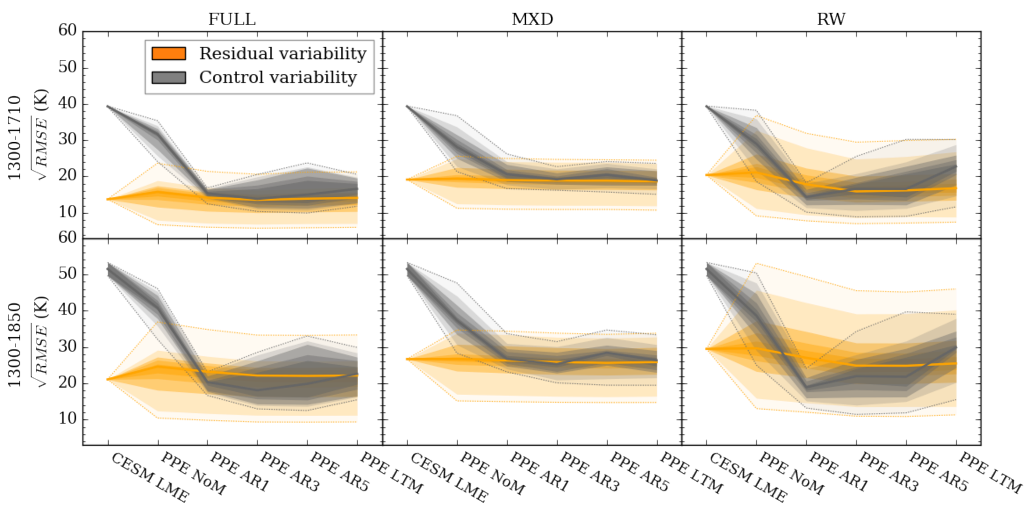


FIG. 10. As figure 9 but for the period 1300-1850.



889 FIG. 11. Unexplained residual variability of the TLS (orange) and square root of sum of squares of equivalent time slice of control variability shown in PPE versions of the CESM LME control simulation (gray).