Supplementary information to "Consistent multi-decadal variability in global temperature reconstructions and simulations over the Common Era"

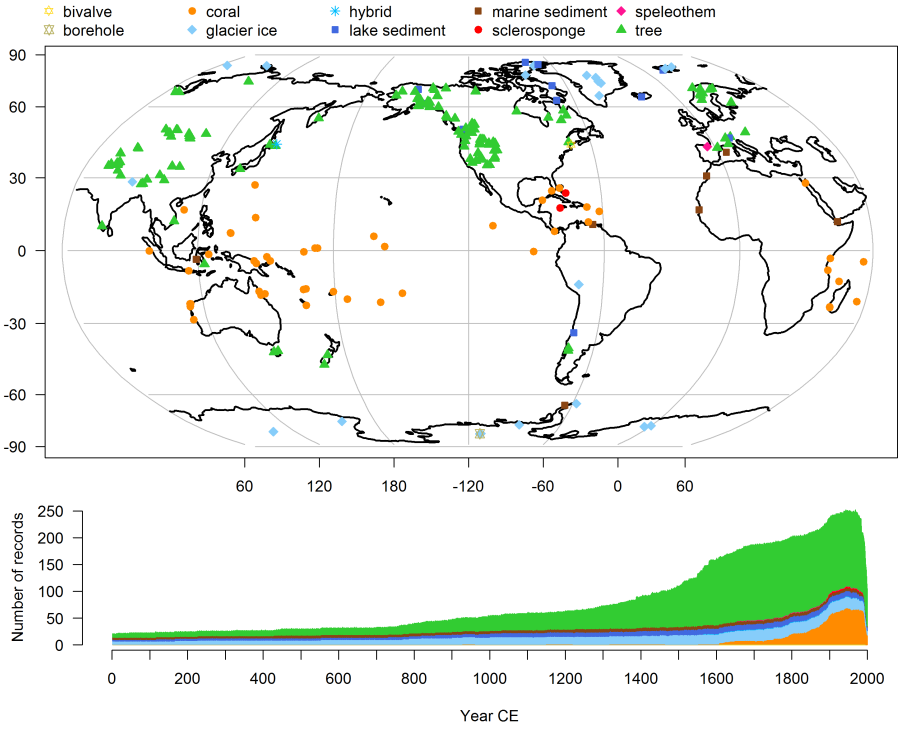
PAGES 2k Consortium*

May 24, 2019

*Corresponding author: Raphael Neukom (neukom@giub.unibe.ch)

Contents

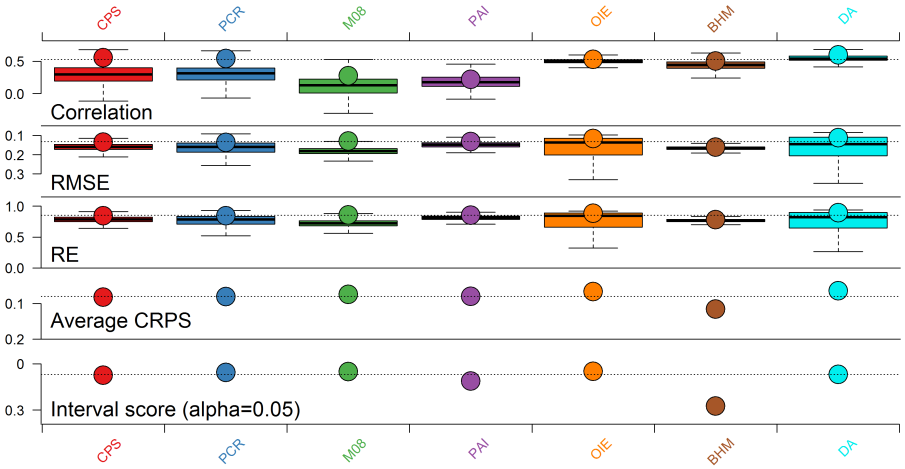
1 Spatio-temporal proxy data availability



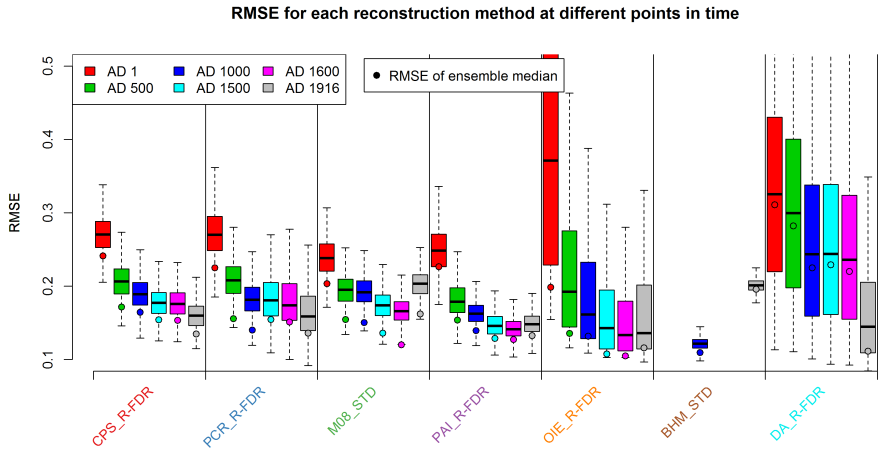
Supplementary Figure 1: Spatiotemporal distribution of proxy records in the R-FDR screened proxy network[1] used in the final reconstructions. Top: Proxy map by archive type, coded by color and shape. Bottom: Temporal availability of proxy data per archive, coded by color as in the top panel.

2 Evaluation of reconstructions

Supplementary Fig. **2** provides a comparison across methods for the evaluation metrics cross-correlation, root mean squared error (RMSE), reduction of error[2] (RE), continuous ranked probability score[3,4] (CRPS) and the interval score[3] at the 95% level. The 1881-1915 evaluation period is used for all methods. A comparison of the RMSE of different proxy subsets representing different points in time is shown in SupplementaryFig. **3**. It is based on reconstructions using only those proxies extending back least to the year 1600, 1500, 1000, 500 and 1, respectively. In general, the different methods yield a consistent picture with similar performance and no discernible best method can be identified based on these metrics. For instance, the DA method has very good performance for all metrics in the most recent period, but errors strongly increase back in time. For alternative ensemble median GMST time series based on evaluation skill see Supplementary Fig. **6**.



Supplementary Figure 2: Evaluation performance of reconstructions with the target over the 1881-1915 evaluation period. Boxplots represent ensemble members, circles the skill of the ensemble median. Dotted horizontal lines are the median of all methods. RMSE, CRPS and interval score have a reversed y-axis, so that performance increases towards the top of the figure for all metrics.



Supplementary Figure 3: RMSE of reconstructions with the target over the 1881-1915 evaluation period for different proxy subsets. Gray boxes represent the full network, purple represents proxy records that extend back to the year 1600 CE or beyond. Cyan: 1500 CE, blue: 1000 CE, green: 500 CE, red 1 CE. Boxplots represent ensemble members, circles the RMSE of the ensemble median. For the BHM method, only the 1000 CE set is available. Note that lower values indicate more skillful results.

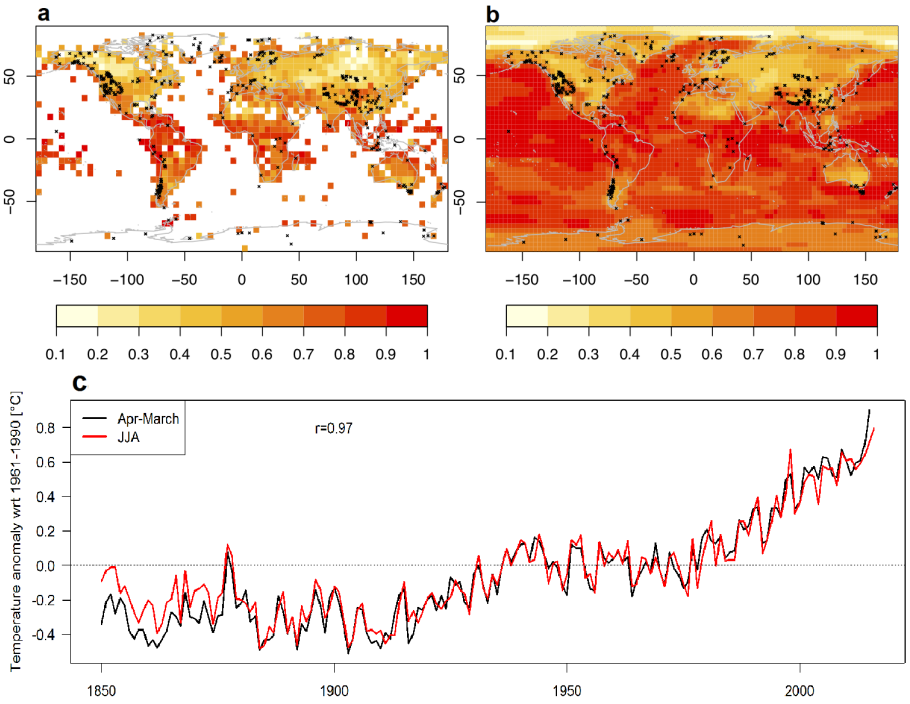
3 Seasonal sensitivity tests

3.1 Exploring possible impacts of treating seasonal proxies as if they represent an annual average

The reconstructions in this study are based on the assumption that the annually resolved proxies represent the annual mean temperature (except for the DA method[5]). This is a valid assumption in some cases but not for all proxy types and locations. For instance, a majority of the tree-ring based records from Northern Hemisphere continents respond more strongly to climatic conditions during the growing season than to the annual mean. To assess the uncertainties in space and time due to this assumption, we conduct two analyses: 1) the correlation between summer and annual temperature at each grid box in an instrumental data set and a model simulation; 2) long-term trends of global mean annual mean summer and annual average temperature over the past millennium.

3.1.1 Correlation between annual mean and boreal summer mean temperature

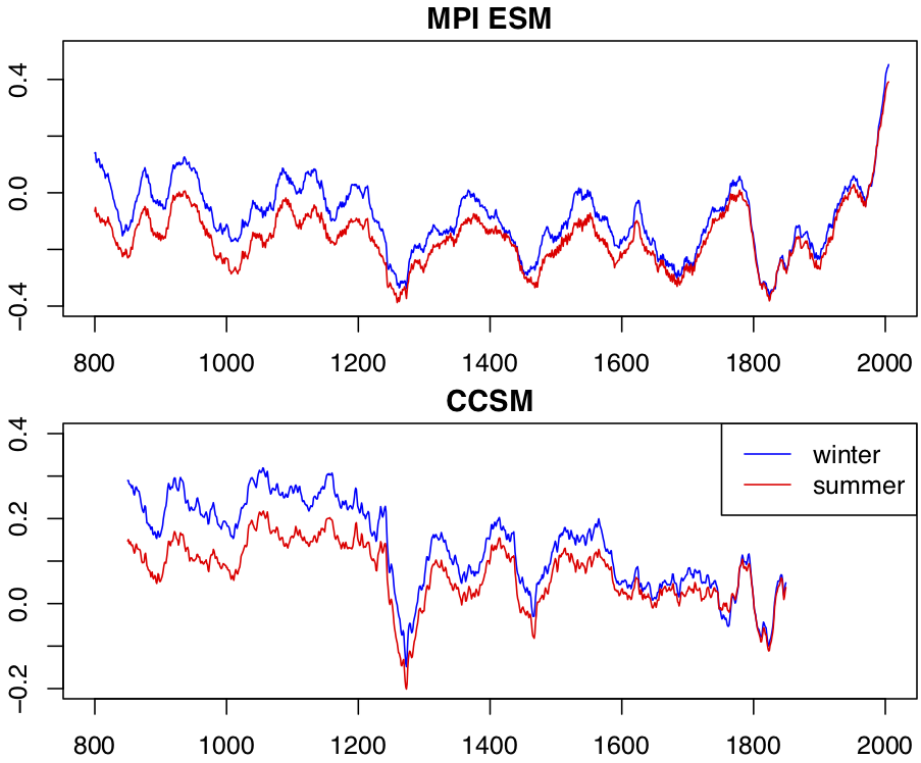
Here we calculate the correlation between annual average temperature and NH summer (JJA) temperature. First, the results are shown for the CRUTEM4 median data set[6] of gridded instrumental observations for the period 1851-2016 (Supplementary Fig. **4**a) (note that in many locations the observations start after 1851). Second, the results are shown for one ensemble member of the MPI ESM model simulations[7] in the period 850-2000 CE (Supplementary Fig. **4**b). Both observations and simulations suggest that correlations of summer temperatures and annual mean temperatures fall below 0.5 for large parts Northern Hemisphere land regions. For the PAGES 2k dataset, however, the majority of sites is located in regions with higher correlation and few time series are from regions with lowest correlation coefficients. Hence, there is still a clearly positive relationship between summer and annual mean temperature. Additionally, aggregation of the paleodata into a global average increases the correlation between annual and summer temperature, and our GMST April-March instrumental target correlates very strongly with JJA GMST ($R = 0.97$; Supplementary Fig. **4**c).



Supplementary Figure 4: **a** CRUTEM4 correlation between annual mean temperature and JJA mean temperature. **b** Same for one MPI ESM ensemble member. Symbols indicate the PAGES 2k paleodata locations in the year 1500 CE. **c** Comparison of the instrumental target used herein (Apr.-Mar. Cowtan & Way[8] GMST) with the boreal summer (JJA) seasonal average of the same dataset.

3.1.2 Long-term trend in summer and winter temperature

Furthermore, we looked at the long-term evolution of NH summer (Apr. to Sep.) and winter (Oct. to Mar.) temperature over the past 1000 years in simulations with the MPI ESM[7] and the CCSM model[9]. We find clear differences in the long-term trend between Apr. to Sep. and Oct. to Mar. season in a global mean temperature. In both models, global mean temperature decreases around 0.2 K per millennium more in the Oct. to Mar. season than in the Apr. to Sep. season (Supplementary Fig. **5**) due to orbital forcing. Given that our proxy database contains more boreal summer than winter proxies, our estimates of the long-term cooling trends are likely minimum values.



Supplementary Figure 5: Global mean 31-year running mean temperature anomalies (with respect to the last 100 years of the simulations). One ensemble member from the MPI ESM (top) and one ensemble member from the CCSM model (bottom). 'Summer' is Apr. to Sep.; 'Winter' is Oct. to Mar.

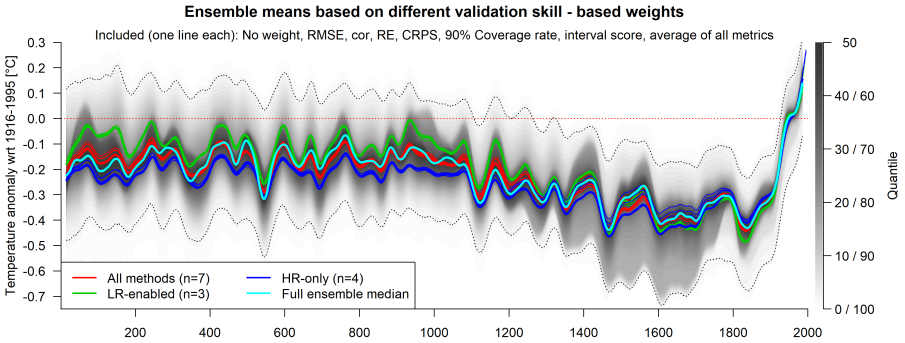
4 GMST best estimates

All seven reconstructions were combined to generate a consensus GMST reconstruction. We explored the sensitivity of the final reconstruction to the choice of weightings for each of the methods. Weighting the different methods and ensemble members based on their evaluation skill within the instrumental era has only a minor effect on the resulting best estimate time-series, because the methods generally perform similarly and the skill metrics do not clearly identify a "best" method (Supplementary Fig. **2**–**3**). Thus, results are extremely robust to using the simple average vs. skill-based weighted means (Supplementary Fig. **6**).

Separating methods that do or do not incorporate low-frequency proxy data has the strongest effect on the multi-method mean (Supplementary Fig. **6**). Methods that incorporate low- and high-frequency records (PAI, OIE and M08) yield best estimates with a larger pre-industrial trend than methods using only high-frequency records. However, due to the relatively short overlap between instrumental and proxy data, and because all but one method (DA) do not forward model the proxy data[10], it is not possible to compare and evaluate the different reconstructions in terms of their performance on time scales longer than multi-decadal.

Thus our data do not allow an objective judgment of which of the two method groups (high-resolution records only vs. all records included) yields the more reliable low-frequency temperature estimate. As stated in the main text, high-resolution paleoclimate archives (such as tree rings) often under-represent fluctuations on time scales longer than multi-decadal, and many are seasonally biased[1,11]. On the other hand, marine-based, low-resolution records seem to overestimate the true variance[12].

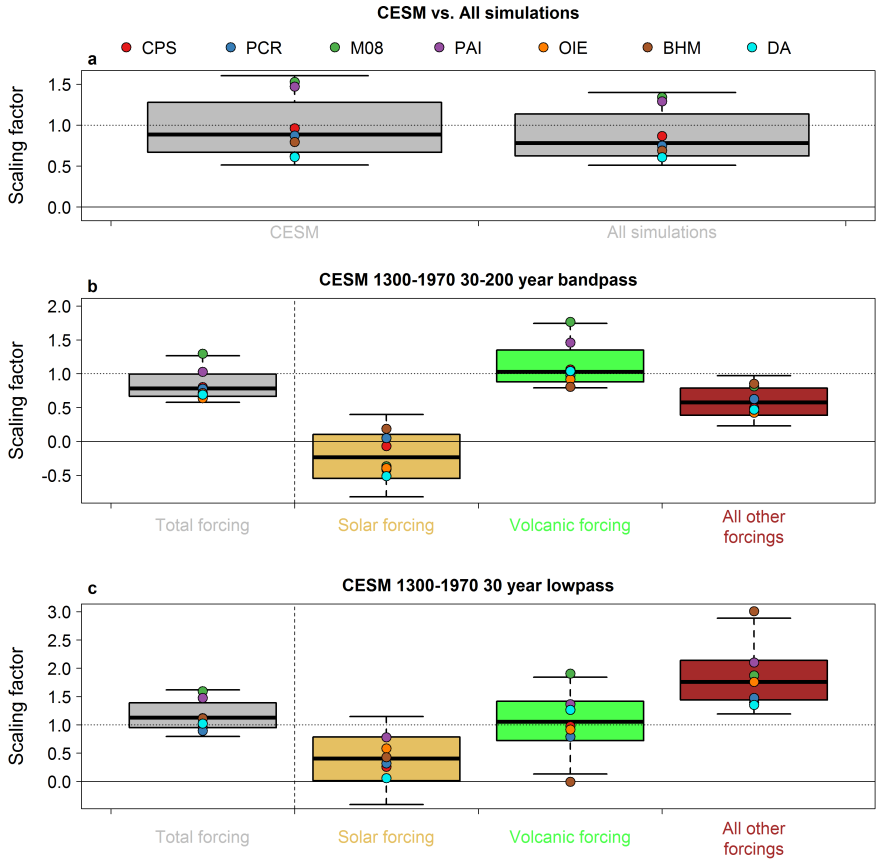
Given these findings, we strongly recommend that future users of this reconstruction product make use of our full 7000-member multi-method ensemble, which captures uncertainties arising from both the different reconstruction methods and sampling of the input data.



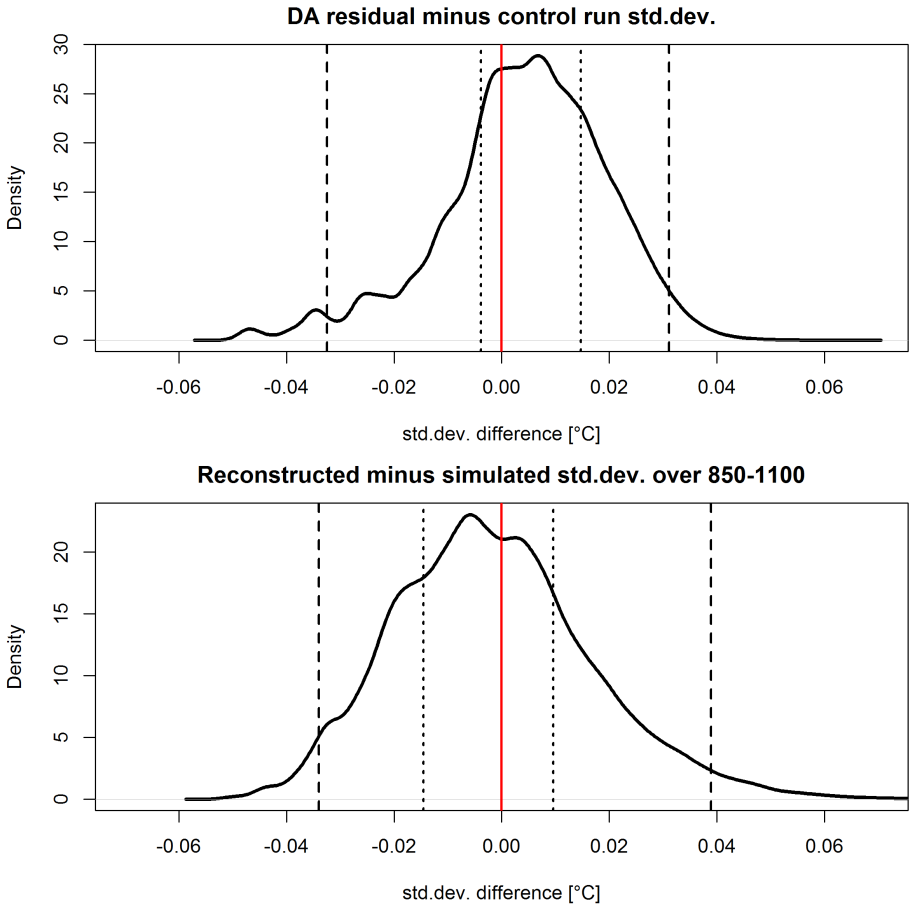
Supplementary Figure 6: Same as Fig. 1a in the main text, but showing alternative approaches to generating a consensus time series. Red: ensemble mean of all methods weighted based on evaluation skill, with one thin line representing each: no weight (full ensemble mean), RMSE, correlation, RE, CRPS, 90% coverage rate, interval score and an average of all metrics. Green: same but using only methods that incorporate low-resolution (>1 year) proxy time series (PAI, M08 and OIE methods). Blue: same but using only methods that only incorporate high-resolution proxy records (<=1 year; CPS, PCR, BHM and DA methods). Cyan: median across the full ensemble.

5 Response to forcing

5.1 Additional D&A plots



Supplementary Figure 7: **a** Scaling factors for total forcing for the CESM model (as shown in Fig. 3a, left). Right: same but using all 23 model simulations instead of only the CESM ensemble. **b** Same as Fig. 3a but over the period 1300-1970, thus including the industrial era. All other forcings are used in this experiment instead of GHG-forcing only to include the full spectrum of anthropogenic forcing, including aerosols. The 30-200 year bandpass filter used here removes the warming trend of the last 150 years. **c** Same as **b** but using 30 year lowpass filtered data. The recent warming trend is now retained in the data and the anthropogenic signal is clearly visible (right boxplot). In fact, the scaling factors are significantly above one for anthropogenic forcing, indicating an underestimation of anthropogenic forcing by the model relative to the reconstructions. This is probably caused by the strong aerosol forcing in CESM, leading to a relatively weak 20th century warming in this model compared to other simulations[13].



Supplementary Figure 8: Comparison of estimates of internal multi-decadal GMST variability. Top: Density plot of all possible comparisons between the estimates of unforced internal variability based on the D&A residuals (n = 7000) and model control simulations (n = 43) shown in Fig.3b (left) in the main text. Each value from the control simulations is subtracted from each value from the D&A residuals. Horizontal dashed (dotted) lines show the 95%-range (interquartile range). Bottom: Same but for the estimates based on the temperature variability during the Medieval Quiet Period from reconstructions (n = 7000) and model simulations (n = 23) shown in Fig.3b (right). In both cases, the value of zero is within the interquartile range, indicating consistency between the different lines of evidence.

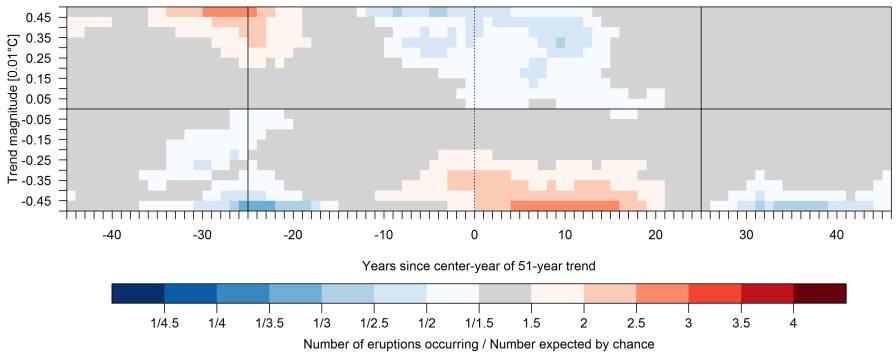
5.2 Volcanic forcing

Volcanic forcing is often interpreted as to alter temperature anomalies on time scales of months to a few years. Our results (Figs. 2-4 in the main text) do suggest a GMST response to volcanic forcing on multi-decadal time scales. Along with evidence from the literature[14, 15], the following analyses indicate the this multi-decadal response is not just an artifact from filtering, but reflects a physically consistent property of the climate system.

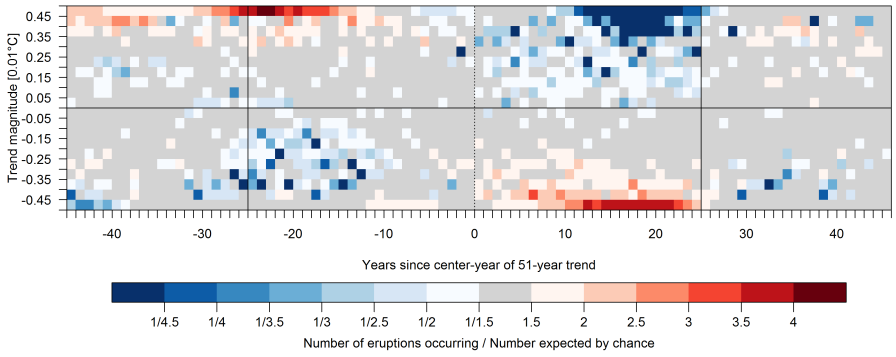
We calculate the probability that a volcanic eruption occurs n years before or after the center-year of a linear trend using all overlapping pre-industrial 51-year trends of all 7000 ensemble members. The results are shown for a window of $-45<n<45$ years in Supplementary Fig. **9**. There is high probability that strong cooling trends are centered around 10 years prior to a large volcanic eruptions (threshold aerosol optical depth AOD$>$0.15, 22 eruptions within 1-1850 CE). Strong warming trends, in turn, are probable to be centered around 25 years after an eruption, i.e. the beginning of the 51-year trend period is roughly at the time of the eruption. Due to this relatively fast beginning of the recovery from cooling, the strongest cooling trend are centered a few years prior to the eruption event. The figure also shows that strong warming (cooling) trends are unlikely to be centered around eruption events (around 25 years after an eruption).

Results for the model simulations are very similar (Supplementary Fig. **10**), indicating that the observed pattern is physically plausible. In the model data, the strong warming trend occur closer to the eruption date (i.e. are shifted to the right on the x-axis) compared to the reconstructions, which causes also the strong cooling events to be centered 3-5 years earlier relative to the eruption date. This faster recovery in the model world is also visible in Fig. 2 in the main text.

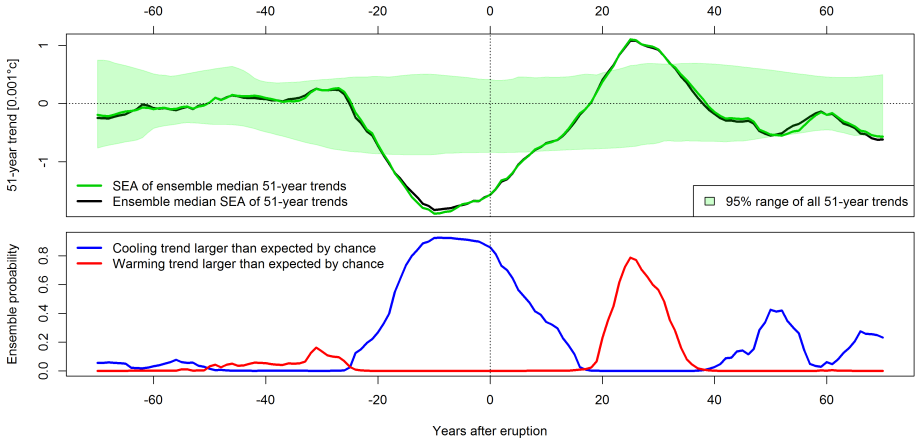
Superposed Epoch Analysis of the 51-year trends confirms these findings and shows that both the cooling around the eruption date and the subsequent strong warming are significant (Supplementary Fig. **11**).



Supplementary Figure 9: Probability of large volcanic eruptions to occur before, during or after a 51-year trend. The number of eruptions occurring in year n before or after the center year of the trend across the entire reconstruction ensemble is divided by the number of eruptions occurring by chance. The horizontal axis shows the lag n from the center year of the trend with the vertical line indicating the full range of the 51-year trend. The vertical axis indicates the magnitude of the trend with large positive trends at the top and large negative trends at the bottom.

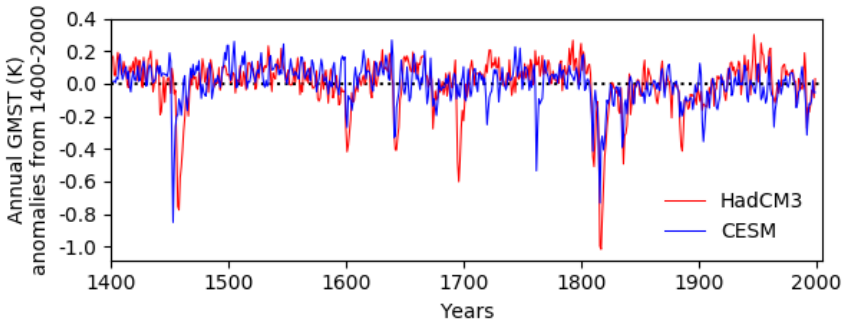


Supplementary Figure 10: Same as Supplementary Fig. **9**,but for the 23 climate model simulations used herein. The period covered is 850-1850.



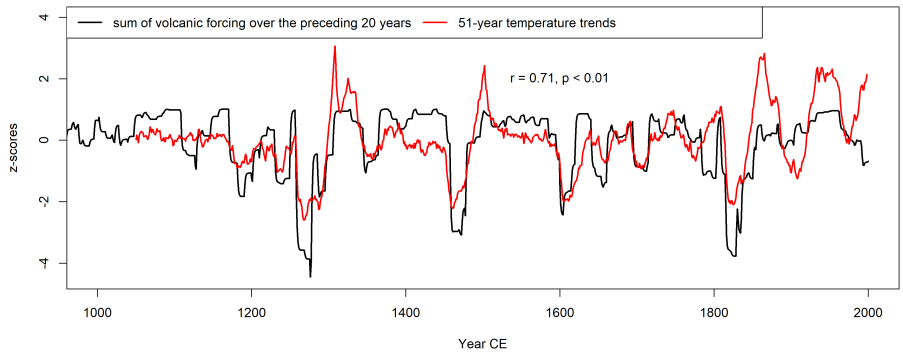
Supplementary Figure 11: Superposed Epoch Analysis of 51-year trends around large volcanic eruptions (AOD>0.15). Top: The green line shows the average 51-year trend of the reconstruction ensemble median between −70 to +70 years after the eruption date (anomalies wrt years -70 to -30). Green shading represents the 95% confidence interval based on all 51-year trends that are not overlapping with an eruption. The black line shows the median response of the individual ensemble members. Bottom: Ensemble probability for the average trend to be below (blue) or above (red) the confidence interval (of the individual ensemble member).

Next, we further assess the temperature response to volcanic forcing in model simulations, to test whether the observed multi-decadal signal in the temperature response also evidenced in physics-based simulations. Supplementary Fig. **12** shows the median temperature evolution in volcanic-only forced simulations of the HadCM3[16] and CESM_1[13] ensembles. This figure illustrates the long-term effect of volcanic cooling, with a clear multi-decadal pattern.



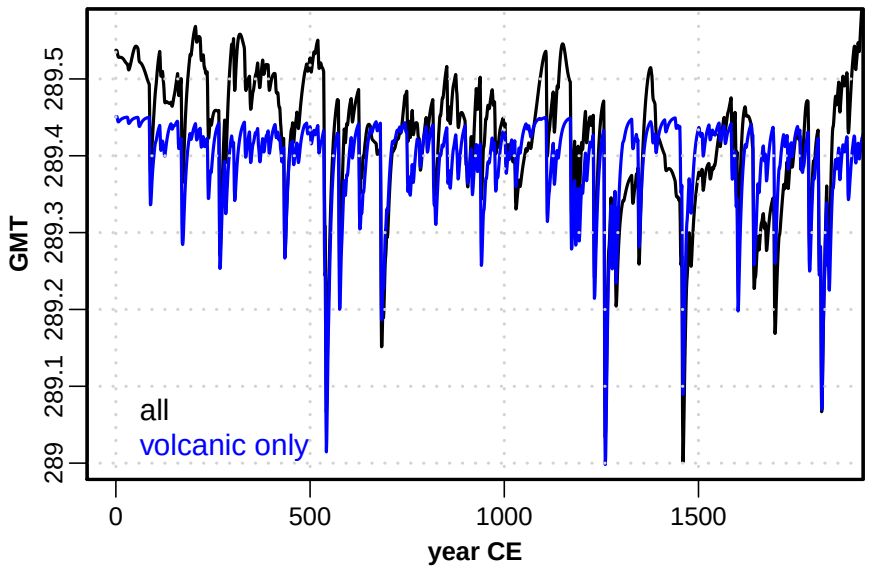
Supplementary Figure 12: Volcanically-only forced model simulations between 1400-2000 CE. Ensemble median (unfiltered) temperature evolution simulated by HadCM3[16] ($n = 3$) and CESM_1[13] ($n = 5$) using only volcanic forcing as external driver.

Supplementary Fig. **13** assess the cumulative effect of magnitude and frequency of volcanic eruptions on multi-decadal temperature trends in climate models. It shows AOD magnitudes integrated over 20 years vs. 51-year temperature trend in the model simulations. There is a strong relationship between these two curves ($r = 0.71, p < 0.01$, corrected for first-order autocorrelation), providing more evidence for an influence of the frequency and magnitude of volcanic forcing on multi-decadal temperature trends.



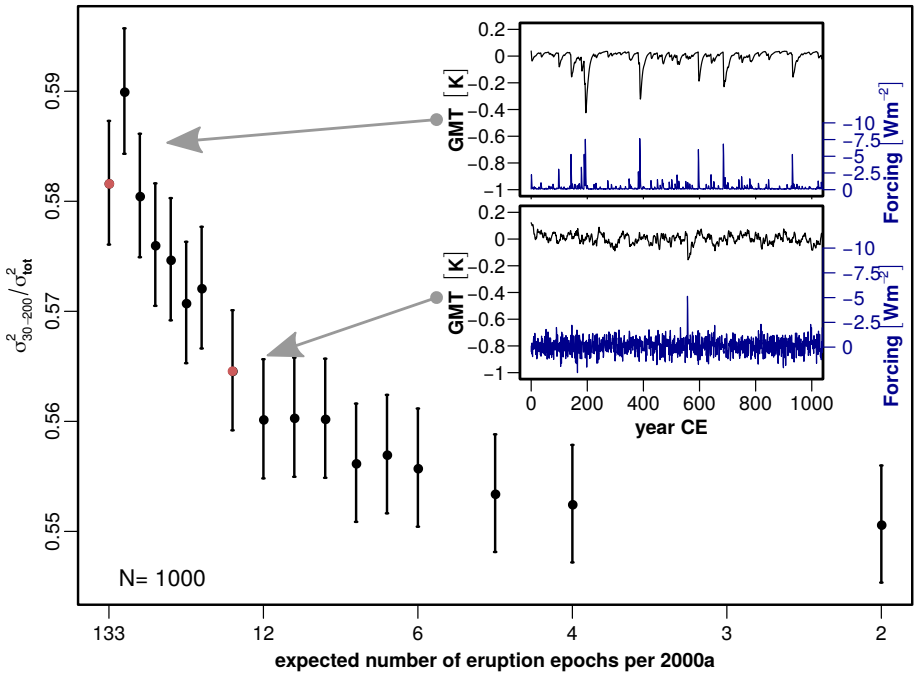
Supplementary Figure 13: 30-year cumulative volcanic forcing and multi-decadal GMST trends in model simulations. Black: 20-year sum of AOD, to represent the cumulative magnitude and frequency of volcanic eruptions. Red: ensemble median 51-year temperature trends across the 23 model simulations used herein. Years on the horizontal axis reflect the last year of each 51-year trend, to account for the lagged response of the trends to eruptions (Supplementary Fig. **9**). Correlation and p-value, corrected for lag-1 autocorrelation, are provided in the Figure.

In addition, we run a simple zero-dimensional Energy Balance Model[17] (EBM, see Methods) to test if changes in the frequency of volcanic eruptions can alter multi-decadal temperature variability. Indeed, a multi-decadal signal is evident in both the full-forced ($all = CO_2 + solar + volcanic$) and volcanic-only EBM (Supplementary Fig. **14**).



Supplementary Figure 14: Temperature variations over the CE based on an EBM using full (black) and volcanic-only (blue) forcing. The blue line corresponds to the green line in Fig. 4a in the main text.

To further test the influence of volcanic forcing on multi-decadal GMST variability, a stochastic volcanic forcing experiment is performed. For this, the 2000-year-long volcanic history[18,19] is split in 133 segments which start from a near-zero volcanic background flux before an eruption, and which contain an event, as well as the subsequent decay to a background flux ($-0.115W/m^2$). For volcanic surrogates, we randomly draw from these 133 volcanic epochs to construct 2000-year long forcing histories. The EBM is run for 1000 randomized forcing histories. A considerable proportion of the resulting GMST variability is contained in the multi-decadal frequency band (red dot at x=133 and upper inset in Supplementary Fig. **15**). Now, keeping the overall forcing variability fixed to ($var(\Delta F) = 0.6W/m^2 + 0.01$), we decrease the recurrence time for volcanic eruptions by mixing in non-volcanic background epochs, adding Gaussian distributed noise over the whole time period (Supplementary Fig. **15** lower inset) such that with fewer eruptions, white noise becomes stronger. We then evaluate the ratio of multi-decadal (30-200a) variance to total variance using power spectra (c.f. Ref. 20). We find that with the increase of the recurrence time of volcanic events, the relative proportion of variance in the multi-decadal to centennial band decreases.

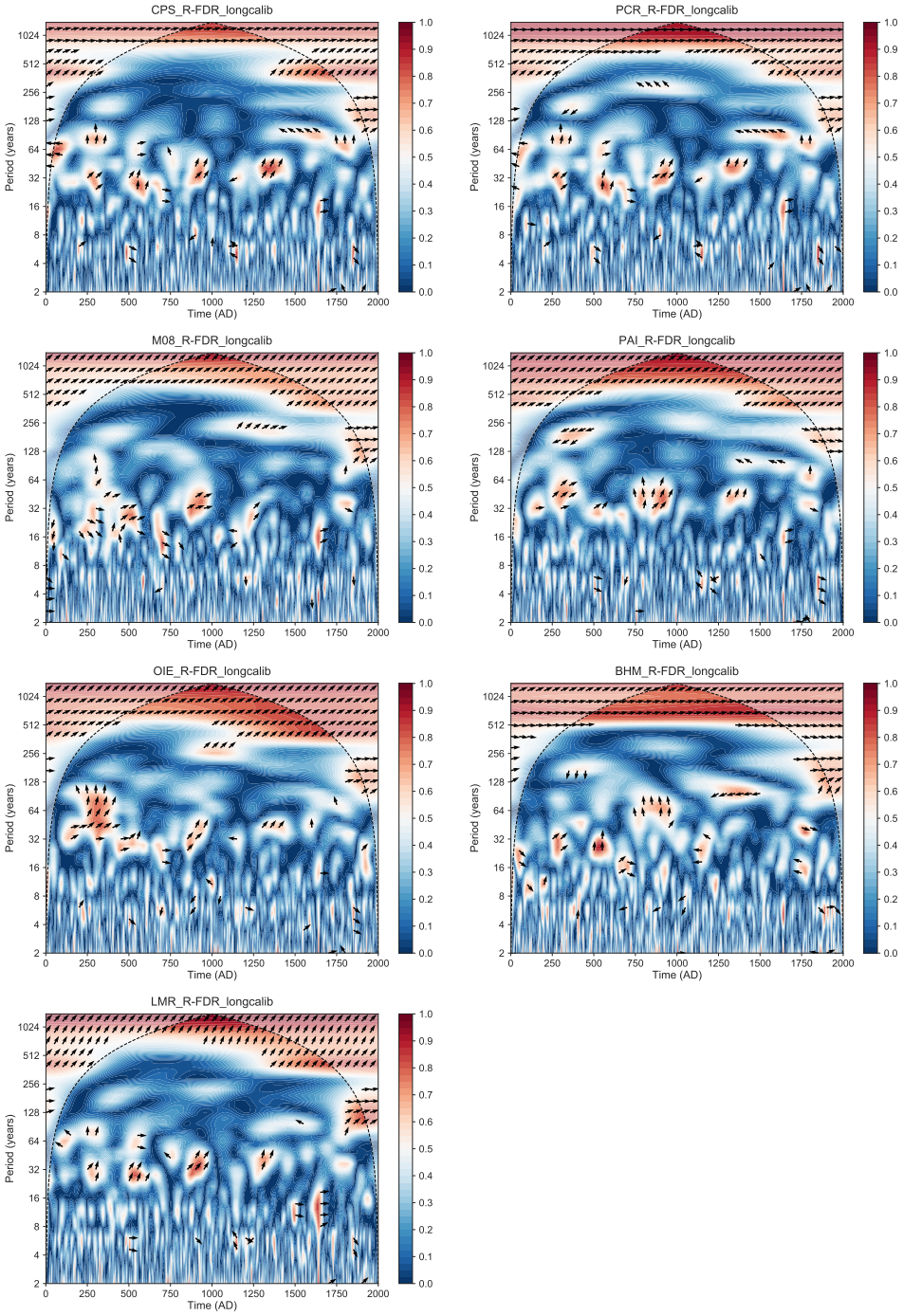


Supplementary Figure 15: Ratio of multi-decadal vs. total GMST variance as a function of the frequency of volcanic eruptions in the EBM. The ratio is calculated 1000 times for n volcanic eruptions per 2000 years for n between 2 and 133 (the value corresponding observations in Ref. 21). The insets show one realization of temperature history (black) and volcanic forcin (blue) in the EBM for $n = 133$ (upper inset) and $n = 14$ (lower inset).

This collection of evidence suggests that the multi-decadal response of GMST variability and trends presented in the main text is consistent with the physics of the climate system.

5.3 Solar forcing

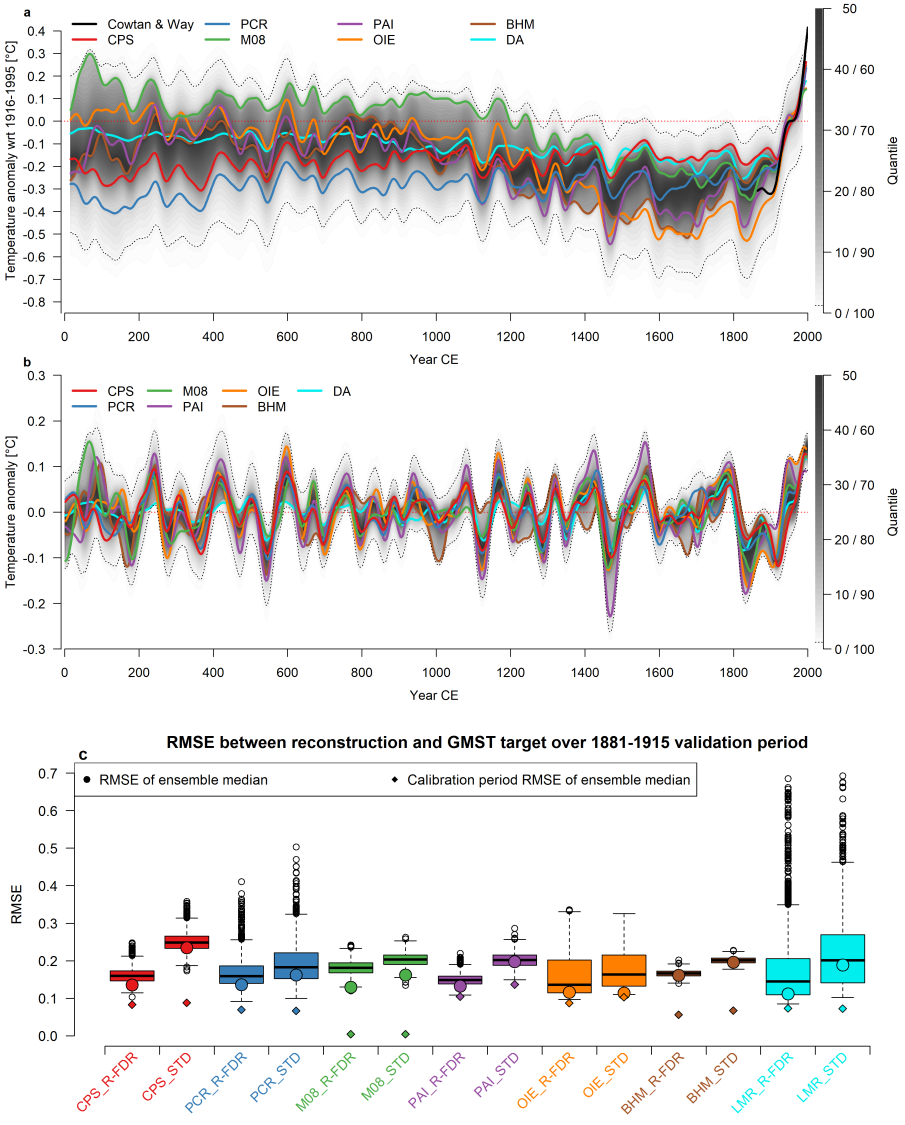
Supplementary Fig. **16** presents a cross wavelet analysis[22,23] of GMST and solar forcing[24,25]. The results indicate no robust relationship between GMST and solar forcing. Coherence is mostly found at the very lowest frequencies (below 500 years), where the degrees of freedom are too small to allow a robust significance analysis.



Supplementary Figure 16: Wavelet coherence between the solar forcing and the reconstructed temperature time series for the different methods. Red denotes coherence > 0.5; blue denotes < 0.5. The relative phase relationship is shown as arrows (plotted where coherence is > 0.5) with in-phase pointing right and anti-phase pointing left, and solar forcing leading temperature by 90 degrees pointing up. "LMR" in the header of the bottom panel refers to the DA method.

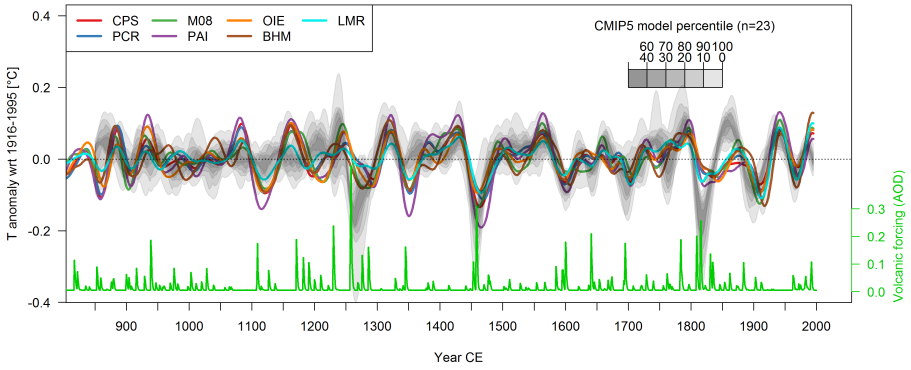
6 Sensitivity to proxy subsets and time-series filtering

Supplementary Fig. **17** shows the reconstructions and their skill based on the full unscreened PAGES 2k proxy matrix as opposed to the screened network used in the main text (see Methods section). The screened subset yields better agreement across methods and higher evaluation skill for all reconstruction techniques.



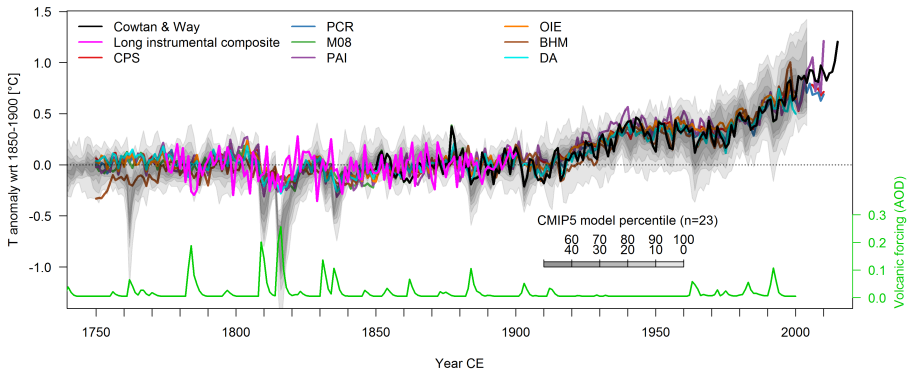
Supplementary Figure 17: **a**, **b** Same as Fig. 1 in the main text but using all proxies within the PAGES 2k database (not only the R-FDR screened records) and based on a 1916-1995 calibration period. **c** evaluation RMSE for the screened (R-FDR) and full (STD) proxy dataset for all reconstruction methods. Boxplots represent ensemble members, circles the skill of the ensemble median, and diamonds the calibration period RMSE of the ensemble median. Note that lower values indicate better performance.

Supplementary Fig. **18** provides an alternative window for the filtering of the time series to show multi-decadal GSMT variability. The figure indicates that the choice of the bandwidth does not influence our interpretations and conclusions.



Supplementary Figure 18: Same as Fig. 2 in the main text but using a window between 20 and 100 years (instead of 30-200 years) for the butterworth filtering of the time series.

There is an apparent data-model mismatch in the 19th century in Fig. 2 in the main text. Supplementary Fig. **19** compares the unfiltered data over 1750-2010 and indicates that the apparent difference is inflated by the filter applied in Fig. 2. The unfiltered data show that the difference is really only caused by the stronger cooling after Tambora in the models and a slightly faster recovery from the cooling caused by the 1830s eruptions in the models. Recent literature[26–28] and a composite of long instrumental stations show that the reconstructions are likely to better capture the temperature variability in the early 19th century compared to the models (with the caveat that the instrumental composite is strongly biased towards the European domain, where all stations are located).

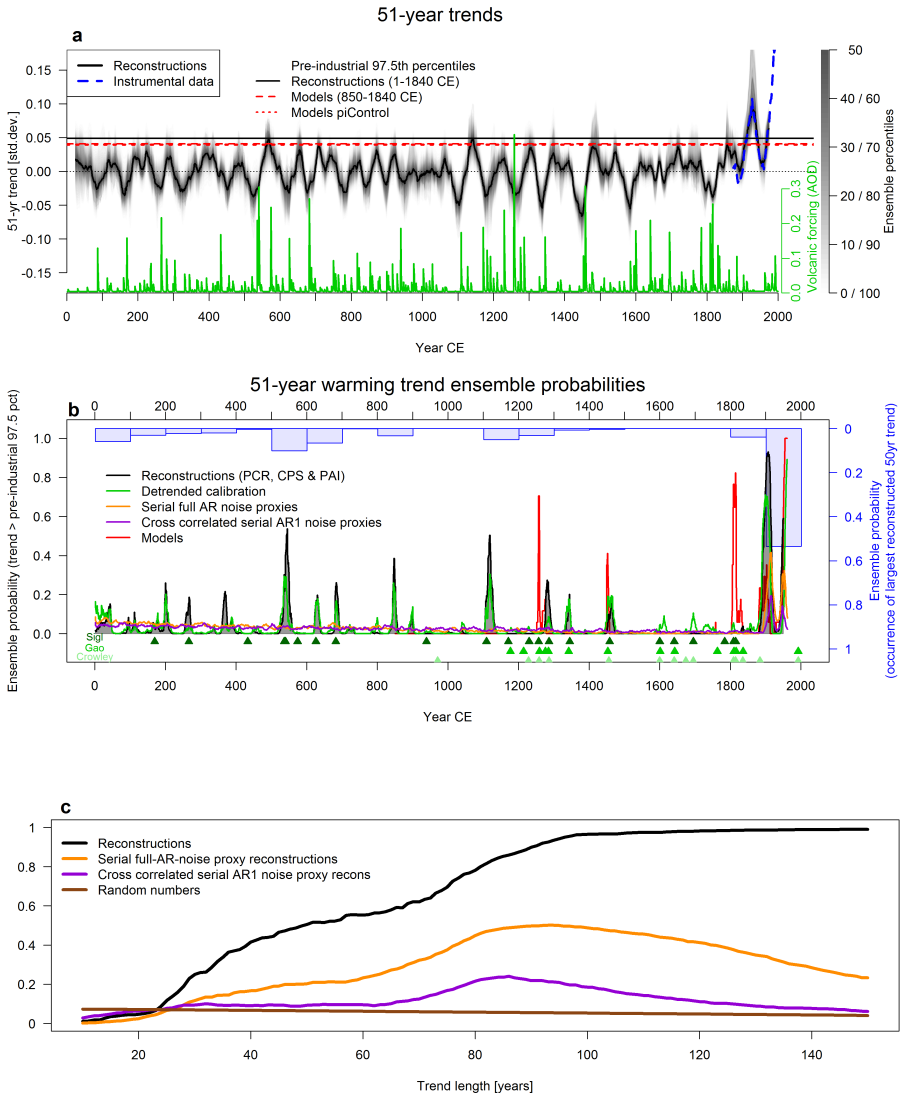


Supplementary Figure 19: Comparison of reconstructions, models and instrumental data over 1750-2010. Same as Fig. 2 but using unfiltered data and over the 1750-2010 time window. The instrumental target is also included (black line). Pink line shows a composite of long instrumental station data extending back to 1775. Station data are from GHCNm (v3, adjusted) and all stations covering 75% of years within 1775-1900 are included (at least nine months with non-missing data required per year). Included stations are Kremsmünster, Wien, Prag, Paris, Karlsruhe, Berlin/Dahlem, Berlin/Tempelhof, München, Hohenpeissenberg, Budapest, Milano, Torino, Vilnius, De Bilt, Trondheim, Warsaw, St. Petersburg, Stockholm, Basel, Genf, Edinburgh, Greenwich, New Haven. Because these are land station data, they have much larger variance than the GMST data. The composite was therefore scaled to the variance of the Cowtan & Way instrumental target over 1850-1900. All anomalies are with respect to the common period of overlap (1850-1900).

7 Trends: Sensitivity figures

Supplementary Fig. **20**a is similar to Fig. 4 in the main text but using relative trends. To correct for variance biases back in time due to proxy replication or other artifacts arising from the methodology[11], the trend of each 51-year period is divided by the temperature standard deviation over the respective 51-year period. Results are very similar to the uncorrected data, indicating that variance issues do not influence our conclusions about trends over different periods.

Supplementary Fig. **20**b shows an alternative illustration for the trends by displaying the ensemble probability for large trends at each time step (lines) and the ensemble probability for the largest 51-year trend within each century (blue bars). The figure also shows the results from a reconstruction using detrended data for calibration (green line). 20^{th} century trends remain large, even with detrended calibration, indicating that the large trends in this century are not a calibration artifact. Supplementary Fig. **20**b also includes results from pure-noise-proxy reconstructions based on two different methods to generate the noisse proxies: Serially correlated full AR noise proxies as described in the Methods section of the main text (orange). Cross-correlated noise proxies are generated such that the multivariate correlation among the noise-proxies is the same as in the real-proxy matrix (purple). These cross-correlated noise proxies have the same AR(1) coefficient as the real proxies (as opposed to the serially correlated noise proxies, which have the full AR-spectrum identical to the real proxies). Pure-noise-proxy reconstructions based on both methods also have the largest trends in the 20^{th} century. The magnitude is, however, much smaller and the ensemble probability for the largest 20^{th} century trends is clearly higher for the real data (Supplementary Fig. **20**c). Note that the noisy-proxy comparisons are only based on the CPS, PCR and PAI methods.



Supplementary Figure 20: **a** Same as Fig. 4a in the main text, but for relative trends. Trends are divided by the GMST standard deviation of each respective 51-year period. **b** Probability of occurrence of large trends. Fraction of reconstruction ensemble members showing a 51-year trend that exceeds the 97.5th percentile of all pre-industrial (1-1850 CE) trends. Gray: reconstructions. Red: model simulations. Green: reconstruction using detrended proxy and target data for calibration. Orange: pure AR-noise proxies (same data as in Fig. 4b). Purple: additional noise proxy experiment using cross-corelated (same covariance matrix as real proxy data) AR(1) noise proxies (see Methods in main text). These experiments were only conducted for the CPS, PCR and PAI methods. Green diamonds represent large volcanic eruptions in different data sets[21,29,30]. Blue bars represent the ensemble probability that the largest 51-year trend occurs in the given century (note the reversed y-axis). **c** Same as Fig. 4b in the main text, but also including cross-correlated noise proxies (purple). These experiments were only conducted for the CPS, PCR and PAI methods.

8 Additional data-model comparison tables

Supplementary Table 1: Same as the last two columns in Tab. 1 in the main text but using unfiltered data. Data present the variance ratios and correlations between unfilterred models and data, respectively, over the period 1300-2000. Ensemble medians are provided for each reconstruction method and 2.5th and 97.5th percentiles in square brackets. Percentage of ensemble members with correlations larger than expected from noise in parenthesis.

	model/data variance ratios unfiltered	model vs. data correlations unfiltered
Composite plus scaling (CPS)	1.4 [0.99, 3.07]	0.41 [0.3, 0.65] (100%)
Principal component regression (PCR)	1.38 [0.9, 2.98]	0.4 [0.29, 0.65] (100%)
Optimal information extraction (OIE)	1.05 [0.84, 2.29]	0.41 [0.32, 0.6] (100%)
Regularized errors in variables (M08)	1.17 [0.74, 2.53]	0.46 [0.35, 0.7] (100%)
Pairwise comparison (PaiCO)	1.62 [1.28, 3.59]	0.49 [0.37, 0.79] (100%)
Hierarchical Bayesian model (BHM)	0.7 [0.56, 1.54]	0.38 [0.29, 0.69] (100%)
Offline data assimilation (DA)	2.35 [1.36, 5.02]	0.51 [0.41, 0.78] (100%)

Supplementary Table 2: Same as the third column in Tab. 1 in the main text but showing the result for each individual model simulation. Data present the variance ratios between multi-decadal (30-200 year bandpass filtered) models and data over the period 1300-2000. Ensemble medians are provided for each reconstruction method and 2.5th and 97.5th percentiles in square brackets.

	CPS	PCR	OIE	M08	PAI	BHM	DA
CCSM4	2.07 [1.51, 2.7]	2.1 [1.43, 3.01]	2.19 [1.77, 2.71]	1.31 [0.83, 1.89]	2.58 [2.41, 2.74]	2.5 [2.11, 2.89]	2.32 [1.97, 3.29]
CESM 1	0.9 [0.66, 1.18]	0.91 [0.62, 1.31]	0.95 [0.77, 1.18]	0.57 [0.36, 0.82]	1.13 [1.05, 1.19]	1.09 [0.92, 1.26]	1.01 [0.86, 1.43]
CESM 2	0.74 [0.54, 0.97]	0.75 [0.51, 1.08]	0.78 [0.63, 0.97]	0.47 [0.3, 0.68]	0.93 [0.86, 0.98]	0.9 [0.75, 1.03]	0.83 [0.71, 1.18]
CESM 3	0.71 [0.52, 0.93]	0.72 [0.49, 1.03]	0.75 [0.61, 0.93]	0.45 [0.29, 0.65]	0.89 [0.83, 0.94]	0.86 [0.72, 0.99]	0.8 [0.67, 1.13]
CESM 4	0.83 [0.6, 1.08]	0.84 [0.57, 1.21]	0.88 [0.71, 1.09]	0.52 [0.33, 0.76]	1.03 [0.96, 1.1]	1 [0.84, 1.15]	0.93 [0.79, 1.31]
CESM 5	0.75 [0.55, 0.98]	0.76 [0.52, 1.09]	0.79 [0.64, 0.98]	0.47 [0.3, 0.68]	0.94 [0.87, 0.99]	0.91 [0.76, 1.05]	0.84 [0.71, 1.19]
CESM 6	0.83 [0.61, 1.09]	0.84 [0.58, 1.21]	0.88 [0.71, 1.09]	0.53 [0.33, 0.76]	1.04 [0.97, 1.1]	1.01 [0.85, 1.16]	0.93 [0.79, 1.32]
CESM 7	0.38 [0.28, 0.5]	0.39 [0.26, 0.55]	0.4 [0.33, 0.5]	0.24 [0.15, 0.35]	0.48 [0.44, 0.5]	0.46 [0.39, 0.53]	0.43 [0.36, 0.6]
CESM 8	0.66 [0.48, 0.87]	0.67 [0.46, 0.97]	0.7 [0.57, 0.87]	0.42 [0.27, 0.61]	0.83 [0.77, 0.88]	0.8 [0.68, 0.93]	0.74 [0.63, 1.05]
CESM 9	0.7 [0.51, 0.92]	0.71 [0.49, 1.02]	0.74 [0.6, 0.92]	0.44 [0.28, 0.64]	0.88 [0.82, 0.93]	0.85 [0.71, 0.98]	0.79 [0.67, 1.12]
CESM 10	0.72 [0.52, 0.94]	0.73 [0.5, 1.04]	0.76 [0.61, 0.94]	0.45 [0.29, 0.66]	0.9 [0.84, 0.95]	0.87 [0.73, 1]	0.8 [0.68, 1.14]
CESM 11	0.61 [0.44, 0.8]	0.62 [0.42, 0.89]	0.64 [0.52, 0.8]	0.39 [0.24, 0.56]	0.76 [0.71, 0.81]	0.74 [0.62, 0.85]	0.68 [0.58, 0.97]
CESM 12	0.54 [0.39, 0.7]	0.54 [0.37, 0.78]	0.57 [0.46, 0.7]	0.34 [0.22, 0.49]	0.67 [0.63, 0.71]	0.65 [0.55, 0.75]	0.6 [0.51, 0.85]
CESM 13	1.07 [0.78, 1.4]	1.09 [0.74, 1.56]	1.13 [0.92, 1.4]	0.68 [0.43, 0.98]	1.34 [1.25, 1.42]	1.29 [1.09, 1.5]	1.2 [1.02, 1.7]
BCC-CSM	1.96 [1.43, 2.56]	1.99 [1.36, 2.85]	2.07 [1.68, 2.57]	1.24 [0.79, 1.79]	2.45 [2.28, 2.6]	2.37 [1.99, 2.73]	2.2 [1.86, 3.11]
CSM1	1.09 [0.8, 1.43]	1.11 [0.76, 1.59]	1.16 [0.94, 1.43]	0.69 [0.44, 1]	1.36 [1.27, 1.45]	1.32 [1.11, 1.52]	1.23 [1.04, 1.74]
GISS 1	1.76 [1.29, 2.31]	1.79 [1.22, 2.57]	1.87 [1.51, 2.32]	1.12 [0.71, 1.62]	2.21 [2.06, 2.34]	2.14 [1.8, 2.47]	1.98 [1.68, 2.81]
GISS 2	1.7 [1.24, 2.23]	1.73 [1.18, 2.48]	1.8 [1.46, 2.24]	1.08 [0.69, 1.56]	2.13 [1.99, 2.26]	2.06 [1.74, 2.38]	1.91 [1.62, 2.71]
GISS 3	1.53 [1.12, 2.01]	1.56 [1.06, 2.24]	1.62 [1.31, 2.01]	0.97 [0.62, 1.4]	1.92 [1.79, 2.04]	1.86 [1.56, 2.14]	1.72 [1.46, 2.44]
HadCM3	1.85 [1.35, 2.42]	1.88 [1.28, 2.69]	1.96 [1.58, 2.42]	1.17 [0.74, 1.69]	2.31 [2.16, 2.45]	2.23 [1.88, 2.58]	2.07 [1.76, 2.94]
MPI 1	2.39 [1.75, 3.13]	2.43 [1.66, 3.49]	2.53 [2.05, 3.14]	1.52 [0.96, 2.19]	2.99 [2.79, 3.17]	2.9 [2.44, 3.34]	2.69 [2.28, 3.81]
MPI 2	1.57 [1.15, 2.06]	1.6 [1.09, 2.29]	1.67 [1.35, 2.06]	1 [0.63, 1.44]	1.97 [1.84, 2.09]	1.9 [1.6, 2.2]	1.77 [1.5, 2.5]
MPI 3	2.47 [1.8, 3.23]	2.51 [1.71, 3.6]	2.62 [2.12, 3.24]	1.56 [0.99, 2.26]	3.09 [2.88, 3.28]	2.99 [2.52, 3.45]	2.77 [2.35, 3.93]

Supplementary Table 3: Same as the last column in Tab. 1 in the main text but showing the result for each individual model simulation. Data present the correlations between multi-decadal (30-200 year bandpass filtered) models and data over the period 1300-2000. Ensemble medians are provided for each reconstruction method and 2.5th and 97.5th percentiles in square brackets. Percentage of ensemble members with correlations larger than expected from noise in parenthesis.

	CPS	PCR	OIE	M08
CCSM4	0.68 [0.54, 0.75] (99.9%)	0.6 [0.38, 0.73] (99.5%)	0.62 [0.52, 0.7] (100%)	0.63 [0.53, 0.7] (99.9%)
CESM 1	0.53 [0.37, 0.63] (98.8%)	0.51 [0.32, 0.63] (98.3%)	0.58 [0.48, 0.66] (99.3%)	0.47 [0.38, 0.54] (97.6%)
CESM 2	0.57 [0.4, 0.66] (99.9%)	0.58 [0.4, 0.69] (99.6%)	0.65 [0.55, 0.72] (100%)	0.56 [0.44, 0.63] (99.9%)
CESM 3	0.58 [0.45, 0.67] (99.9%)	0.56 [0.4, 0.68] (99.6%)	0.64 [0.55, 0.72] (99.9%)	0.55 [0.44, 0.64] (99.7%)
CESM 4	0.64 [0.51, 0.72] (100%)	0.59 [0.37, 0.71] (99.5%)	0.64 [0.55, 0.72] (100%)	0.67 [0.58, 0.74] (100%)
CESM 5	0.58 [0.43, 0.66] (99.9%)	0.59 [0.43, 0.71] (99.8%)	0.59 [0.49, 0.66] (100%)	0.56 [0.47, 0.63] (99.9%)
CESM 6	0.55 [0.4, 0.65] (99.1%)	0.5 [0.29, 0.63] (97.8%)	0.59 [0.49, 0.68] (99.5%)	0.61 [0.5, 0.68] (99.7%)
CESM 7	0.5 [0.37, 0.59] (98.7%)	0.49 [0.32, 0.6] (98.4%)	0.53 [0.43, 0.61] (99.5%)	0.51 [0.4, 0.59] (99.2%)
CESM 8	0.6 [0.47, 0.68] (99.6%)	0.57 [0.39, 0.69] (99.5%)	0.58 [0.48, 0.67] (99.8%)	0.61 [0.5, 0.68] (99.8%)
CESM 9	0.56 [0.45, 0.65] (99.5%)	0.54 [0.37, 0.65] (99.3%)	0.6 [0.5, 0.68] (99.8%)	0.6 [0.48, 0.68] (99.7%)
CESM 10	0.61 [0.49, 0.68] (99.9%)	0.61 [0.45, 0.71] (99.5%)	0.67 [0.59, 0.74] (100%)	0.64 [0.55, 0.7] (99.8%)
CESM 11	0.75 [0.65, 0.81] (100%)	0.72 [0.56, 0.81] (100%)	0.75 [0.67, 0.81] (100%)	0.77 [0.66, 0.83] (100%)
CESM 12	0.55 [0.43, 0.63] (99.7%)	0.55 [0.37, 0.67] (99.6%)	0.54 [0.44, 0.62] (99.8%)	0.56 [0.45, 0.63] (99.8%)
CESM 13	0.65 [0.51, 0.74] (99.8%)	0.59 [0.36, 0.73] (99.4%)	0.6 [0.48, 0.69] (99.8%)	0.64 [0.54, 0.72] (100%)
CSIRO	0.68 [0.55, 0.75] (100%)	0.61 [0.41, 0.74] (99.7%)	0.65 [0.54, 0.72] (99.9%)	0.67 [0.58, 0.74] (100%)
CSM1	0.59 [0.46, 0.67] (99.7%)	0.56 [0.36, 0.68] (99.2%)	0.63 [0.54, 0.7] (100%)	0.48 [0.35, 0.6] (97.7%)
GISS 1	0.75 [0.63, 0.82] (100%)	0.7 [0.53, 0.8] (100%)	0.78 [0.7, 0.83] (100%)	0.68 [0.59, 0.76] (99.8%)
GISS 2	0.8 [0.68, 0.86] (100%)	0.74 [0.52, 0.84] (100%)	0.78 [0.71, 0.84] (100%)	0.71 [0.62, 0.78] (100%)
GISS 3	0.73 [0.62, 0.8] (100%)	0.68 [0.5, 0.79] (99.8%)	0.72 [0.62, 0.79] (100%)	0.65 [0.54, 0.73] (99.9%)
HadCM3	0.75 [0.64, 0.81] (100%)	0.69 [0.47, 0.79] (99.8%)	0.73 [0.64, 0.8] (100%)	0.8 [0.72, 0.85] (100%)
MPI 1	0.71 [0.59, 0.78] (99.9%)	0.65 [0.42, 0.77] (99.7%)	0.67 [0.57, 0.75] (100%)	0.69 [0.59, 0.75] (99.9%)
MPI 2	0.74 [0.63, 0.81] (100%)	0.69 [0.52, 0.79] (99.8%)	0.75 [0.67, 0.81] (100%)	0.74 [0.66, 0.79] (100%)
MPI 3	0.74 [0.63, 0.8] (100%)	0.67 [0.46, 0.78] (99.8%)	0.71 [0.62, 0.78] (100%)	0.72 [0.64, 0.78] (100%)
	PAI	BHM	DA	
CCSM4	0.67 [0.64, 0.7] (99.9%)	0.47 [0.41, 0.52] (98.4%)	0.67 [0.63, 0.7] (100%)	
CESM 1	0.57 [0.54, 0.61] (99.8%)	0.46 [0.4, 0.52] (97.9%)	0.62 [0.57, 0.68] (100%)	
CESM 2	0.53 [0.49, 0.56] (99.8%)	0.63 [0.57, 0.68] (100%)	0.58 [0.53, 0.61] (99.9%)	
CESM 3	0.57 [0.52, 0.6] (99.8%)	0.55 [0.48, 0.6] (99.9%)	0.52 [0.48, 0.55] (99.4%)	
CESM 4	0.6 [0.56, 0.62] (99.7%)	0.5 [0.43, 0.55] (98.7%)	0.62 [0.57, 0.66] (99.8%)	
CESM 5	0.58 [0.54, 0.6] (100%)	0.56 [0.51, 0.62] (100%)	0.59 [0.54, 0.62] (100%)	
CESM 6	0.48 [0.44, 0.51] (97.6%)	0.44 [0.37, 0.5] (96.8%)	0.54 [0.48, 0.59] (99.2%)	
CESM 7	0.42 [0.37, 0.46] (96.8%)	0.52 [0.46, 0.57] (99.2%)	0.4 [0.37, 0.42] (95.8%)	
CESM 8	0.54 [0.5, 0.57] (99.9%)	0.4 [0.33, 0.46] (96.7%)	0.55 [0.52, 0.59] (100%)	
CESM 9	0.48 [0.44, 0.52] (98.3%)	0.52 [0.46, 0.57] (98.8%)	0.51 [0.45, 0.53] (98.8%)	
CESM 10	0.56 [0.51, 0.59] (99.7%)	0.6 [0.54, 0.65] (99.8%)	0.61 [0.56, 0.63] (99.9%)	
CESM 11	0.72 [0.69, 0.76] (100%)	0.62 [0.57, 0.66] (99.9%)	0.72 [0.68, 0.75] (100%)	
CESM 12	0.47 [0.43, 0.5] (99.4%)	0.38 [0.31, 0.44] (96.2%)	0.55 [0.52, 0.58] (99.7%)	
CESM 13	0.6 [0.56, 0.63] (99.9%)	0.54 [0.47, 0.59] (99.8%)	0.57 [0.51, 0.62] (99.9%)	
CSIRO	0.68 [0.64, 0.71] (100%)	0.45 [0.39, 0.51] (98.1%)	0.73 [0.66, 0.79] (100%)	
CSM1	0.67 [0.62, 0.71] (99.8%)	0.64 [0.59, 0.68] (99.9%)	0.59 [0.54, 0.63] (99.5%)	
GISS 1	0.81 [0.78, 0.83] (100%)	0.62 [0.57, 0.67] (99.9%)	0.85 [0.8, 0.87] (100%)	
GISS 2	0.87 [0.86, 0.89] (100%)	0.68 [0.62, 0.72] (99.9%)	0.89 [0.83, 0.9] (100%)	
GISS 3	0.81 [0.78, 0.84] (100%)	0.48 [0.42, 0.54] (98.6%)	0.84 [0.79, 0.86] (100%)	
HadCM3	0.71 [0.67, 0.75] (100%)	0.61 [0.54, 0.65] (99.9%)	0.79 [0.75, 0.84] (100%)	
MPI 1	0.72 [0.68, 0.75] (100%)	0.52 [0.45, 0.58] (99.4%)	0.76 [0.71, 0.81] (100%)	
MPI 2	0.73 [0.69, 0.76] (100%)	0.52 [0.45, 0.58] (99.2%)	0.81 [0.76, 0.86] (100%)	
MPI 3	0.76 [0.73, 0.8] (100%)	0.5 [0.43, 0.56] (98.5%)	0.79 [0.73, 0.84] (100%)	

References

1. PAGES2k Consortium. A global multiproxy database for temperature reconstructions of the Common Era. *Scientific Data* **4**, sdata201788 (2017).

2. Cook, E. R., Briffa, K. R. & Jones, P. D. Spatial regression methods in dendroclimatology: A review and comparison of two techniques. *International Journal of Climatology* **14**, 379–402 (1994).

3. Gneiting, T. & Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102**, 359–378 (2007).

4. Werner, J. P. & Tingley, M. P. Technical Note: Probabilistically constraining proxy ageâĂŞdepth models within a Bayesian hierarchical reconstruction model. *Clim. Past* **11**, 533–545 (2015).

5. Hakim, G. J. *et al.* The last millennium climate reanalysis project: Framework and first results. *Journal of Geophysical Research: Atmospheres* **121**, 6745–6764 (2016).

6. Osborn, T. J. & Jones, P. D. The CRUTEM4 land-surface air temperature data set: construction, previous versions and dissemination via Google Earth. *Earth System Science Data* **6**, 61–68 (2014).

7. Jungclaus, J. H. *et al.* Climate and carbon-cycle variability over the last millennium. *Climate of the Past* **6**, 723–737 (2010).

8. Cowtan, K. & Way, R. G. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society* **140**, 1935–1944 (2014).

9. Hofer, D., Raible, C. C. & Stocker, T. F. Variations of the Atlantic meridional overturning circulation in control and transient simulations of the last millennium. *Climate of the Past* **7**, 133–150 (2011).

10. Dee, S. G., Steiger, N. J., Emile-Geay, J. & Hakim, G. J. On the utility of proxy system models for estimating climate states over the common era. *Journal of Advances in Modeling Earth Systems* **8**, 1164–1179 (2016).

11. Christiansen, B. & Ljungqvist, F. C. Challenges and perspectives for large-scale temperature reconstructions of the past two millennia. *Reviews of Geophysics* **55**, 40–96 (2017).

12. McGregor, H. V. *et al.* Robust global ocean cooling trend for the pre-industrial Common Era. *Nature Geoscience* **8**, 671–677 (2015).

13. Otto-Bliesner, B. L. *et al.* Climate Variability and Change since 850 CE: An Ensemble Approach with the Community Earth System Model. *Bulletin of the American Meteorological Society* **97**, 735–754 (2016).

14. Miller, G. H. *et al.* Abrupt onset of the Little Ice Age triggered by volcanism and sustained by sea-ice/ocean feedbacks. *Geophysical Research Letters* **39**, L02708 (2012).

15. Lehner, F., Born, A., Raible, C. C. & Stocker, T. F. Amplified Inception of European Little Ice Age by Sea IceâĂŞOceanâĂŞAtmosphere Feedbacks. *Journal of Climate* **26**, 7586–7602 (2013).

16. Schurer, A. P., Hegerl, G. C., Mann, M. E., Tett, S. F. B. & Phipps, S. J. Separating Forced from Chaotic Climate Variability over the Past Millennium. *Journal of Climate* **26**, 6954–6973 (2013).

17. Goosse, H. *Climate System Dynamics and Modelling* (Cambridge University Press, New York, 2015).

18. Sigl, M. *et al.* No role for industrial black carbon in forcing 19th century glacier retreat in the Alps. *The Cryosphere Discuss.* **2018**, 1–34 (2018).

19. IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013).

20. Rehfeld, K., MÃijnch, T., Ho, S. L. & Laepple, T. Global patterns of declining temperature variability from the Last Glacial Maximum to the Holocene. *Nature* **554**, 356–359 (2018).

21. Sigl, M. *et al.* Timing and climate forcing of volcanic eruptions for the past 2,500 years. *Nature* **523**, 543–549 (2015).

22. Foster, G. Wavelets for period analysis of unevenly sampled time series. *The Astronomical Journal* **112**, 1709 (1996).

23. Grinsted, A., Moore, J. C. & Jevrejeva, S. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlin. Processes Geophys.* **11**, 561–566 (2004).

24. Vieira, L. E. A., Solanki, S. K., Krivova, N. A. & Usoskin, I. Evolution of the solar irradiance during the Holocene. *Astronomy & Astrophysics* **531**, A6 (2011).

25. Wu, C.-J. SATIRE-M reconstruction of spectral solar irradiance over the Holocene (2017).

26. Marotzke, J. & Forster, P. M. Forcing, feedback and internal variability in global temperature trends. *Nature* **517**, 565–570 (2015).

27. Zanchettin, D. *et al.* A coordinated modeling assessment of the climate response to volcanic forcing. *PAGES News* **23**, 54–55 (2015).

28. Zanchettin, D. *et al.* Clarifying the Relative Role of Forcing Uncertainties and Initial-Condition Unknowns in Spreading the Climate Response to Volcanic Eruptions. *Geophysical Research Letters* **46** (2019).

29. Gao, C., Robock, A. & Ammann, C. Volcanic forcing of climate over the past 1500 years: An improved ice core-based index for climate models. *Journal of Geophysical Research* **113**, D23111 (2008).

30. Crowley, T. J. & Unterman, M. B. Technical details concerning development of a 1200 yr proxy index for global volcanism. *Earth System Science Data* **5**, 187–197 (2013).