Supplementary information to ”Past millennial solar forcing magnitude: a statistical hemispheric-scale climate model versus proxy data comparison”, by:

Alistair Hind and Anders Moberg
Department of Physical Geography and Quaternary Geology, Stockholm University, Sweden
correspondence: aistair.hind@natgeo.su.se

September 17, 2012

1 Introduction

This document provides supplementary information to ”Past millennial solar forcing magnitude: a statistical hemispheric-scale climate model vs. proxy data comparison”, which compares the millennial simulations by Jungclaus et al. (2010) with a set of selected Northern Hemispheric-scale temperature reconstructions for the last millennium, using the statistical framework of Sundberg et al. (2012). The information provided here are: Details about the selected reconstructions (section 2), description of the selected large-scale instrumental temperature data (section 3), motivation for choice of calibration period (section 4), explanation of the estimation of instrumental measurement error variance (section 5), description of ensemble calibration experiments (section 6) and additional figures from the model vs. proxy data comparisons, not shown in the main paper (section 8). The mathematical notation used here is the same as in Sundberg et al. (2012).

2 Northern Hemispheric-scale temperature reconstructions

Six Northern Hemispheric-scale temperature reconstructions, reaching back at least to AD 1000, by different author teams were selected. Together they represent a variety of currently used reconstruction methods and proxy data collections. The six selected reconstructions are:

- Christiansen and Ljungqvist (2011); CHR
- Juckes et al. (2007); JUC
- Hegerl et al. (2007); HEG
- Jones and Mann (2004); JON
- Ljungqvist (2010); LJU
- D’Arrigo et al. (2006); DAR

The number and locations of local/regional proxies used by each author is illustrated in Figure 1 of the main article. It can be seen there that almost all proxy series are located on land areas between 30°N and 75°N. Essential details of the methods and data used by each author are provided below, together with URLs to the original data files. The following abbreviations are used: NH = entire Northern Hemisphere, N20N = areas north of 20°N, N30N = areas north of 30°N.

Christiansen and Ljungqvist (2011); CHR

Multi-proxy.
Proxies' latitude range: 32–81N.
Original calibration target: N30N.
Original seasonal target: annual mean temperature.
Method: 'classical' calibration of each proxy series against local temperatures, followed by arithmetic averaging of local temperature reconstructions ("LOC").
Temporal resolution: annual.
Calibration period: 1880–1960.
Reconstruction period: 1000–1975.

`ftp://ftp.ncdc.noaa.gov/pub/data/paleo/contributions_by_author/christiansen2011/christiansen2011.txt`

23 local/regional proxies (proxy # 1–4, 8–10, 12–14, 16–18, 22–23, 25, 30–31, 35, 37–40 in their Table 1) of which 11 have annual resolution (9 tree-ring).

Procedure: (i) preliminary selection of proxy records with "accurate" dating, "reasonably high sample resolution" and "good correlation" with local temperatures, (ii) interpolation to annual resolution (where necessary), (iii) screening to avoid proxies without significant correlation with local temperatures (annual mean temperatures in nearest corresponding grid-boxes in HadCRUT2v), (iv) linear regression of proxy on local temperatures (i.e. 'classical calibration'), (v) the local reconstructed temperatures are arithmetically averaged. Note: the instrumental N30N target temperature is never used in this procedure.

Juckes et al. (2007); JUC

Multi-proxy.
Proxies' latitude range: 14S–73N.
Original calibration target: NH.
Original seasonal target: annual mean temperature.
Method: unweighted composite with variance matching.
Temporal resolution: annual.
Calibration period: 1856–1980.
Reconstruction period: 1000–1980.

`http://www.clim-past-discuss.net/2/1001/2006/cpd-2-1001-2006-supplement.zip`

13 local/regional proxies (the proxies in their "Union" reconstruction, i.e. those marked with "*" in their Table 1), of which 7 have annual resolution (6 tree-ring).

Procedure: (i) selection of proxies based on physically grounded expectations that they have a temperature signal, not on statistical evaluation, (ii) (presumably) interpolation to annual resolution where necessary, (iii) normalization to unit variance, (iv) arithmetic average of the normalized proxies, (v) variance of composite matched to be the same as for the NH annual mean temperature (HadCRUT2v). Juckes et al. (2007) did not clearly state whether steps (iii)–(iv) were made using smoothed or unsmoothed data.

Hegerl et al. (2007); HEG

Multi-proxy.
Proxies' latitude range: 34–74N.
Original calibration target: N30N (land-only).
Original seasonal target: annual mean temperature.
Method: weighted composite calibrated with TLS regression.
Temporal resolution: annual, but decadally smoothed.

Calibration period: 1880–1960.
Reconstruction period: 946–1960.

`ftp://ftp.ncdc.noaa.gov/pub/data/paleo/contributions_by_author/hegerl2007/nh-temp-hegerl2007.txt`

7 local/regional proxies (their "CH-blend (long)"), of which 5 have annual resolution (5 tree-ring).

Procedure: (i) selection of proxies (apparently) based on physically grounded expectations that they have a temperature signal, not on statistical evaluation, (ii) (presumably) interpolation to annual resolution where necessary, (iii) decadal smoothing (where necessary), (iv) normalization (of decadal data) to unit variance, (v) composite formed by weighting the normalized proxies by their correlation with decadal N30N land-only annual-mean temperatures, (vi) TLS regression of composite against decadal N30N land-only annual-mean temperatures (using independent estimates of error variances in proxy composite and instrumental temperatures).

Jones and Mann (2004); JON

Multi-proxy.
Proxies' latitude range: 32–74N.
Original calibration target: NH.
Original seasonal target: annual mean temperature.
Method: weighted composite with variance matching.
Temporal resolution: annual, but is presented as 40-year smoothed.
Calibration period: 1856–1995 (1901–1980).
Reconstruction period: 200–1995.

`ftp://ftp.ncdc.noaa.gov/pub/data/paleo/contributions_by_author/jones2004/jonesmann-nhrecon-rescale.txt`

8 local/regional proxies, of which three have annual resolution (3 tree-ring; one of these is a composite from three distinct sites in northern Eurasia, wherefore it is relevant to say that 10 local/regional proxies (of which 5 are tree-rings), rather than eight, are used).

Procedure: (i) selection based on indicators with either (for non-tree-rings) no known limitations in resolving millennial-scale variability or (for tree-rings) with conservative standardization methods, (ii) (presumably) interpolation to annual resolution where necessary, (iii) decadal smoothing (low-pass filtering at $f = 0.1$ cycle/year), (iv) normalization to unit variance, (v) composite formed by weighting of proxies by both area (size of region represented by proxy) and correlation with decadal temperatures (1901-1980) for the corresponding region, (vi) composite re-scaled to the same decadal variance as the instrumental NH temperature record over the 1856-1995 period. Steps (i)–(v) define the 'standard' reconstruction of Mann and Jones (2003), while step (vi) was undertaken by Jones and Mann (2004).

Ljungqvist (2010); LJU

Multi-proxy.
Proxies' latitude range: 30–81N.
Original calibration target: N30N.
Original seasonal target: annual mean temperature.
Method: composite with variance matching.
Temporal resolution: decadal.
Calibration period: 1850–1989.
Reconstruction period: 9–1999.

`ftp://ftp.ncdc.noaa.gov/pub/data/paleo/contributions_by_author/`

ljungqvist2010/ljungqvist2010.txt

30 local/regional proxies, of which 16 have annual resolution (10 tree-ring).

Procedure: (i) selection made to include as many as possible "high-quality" temperature proxies with annual to multi-decadal resolution reaching back at least to AD 1000, while excluding tree-ring records from arid and semi-arid regions to avoid potential influence from drought stress, (ii) interpolation to annual resolution where necessary, (iii) normalization to unit variance, (iv) calculation of non-overlapping 10-year means, representing the years AD 0-9, 10- 19, ..., 1990-1999, (v) composite formed by taking arithmetic average of the proxies, (vi) composite re-scaled to have the same decadal variance as the instrumental N30N temperatures (HadCRUT3v) over 1859-1989. Note that normalization was made before calculating the 10-yr means, which implies that more weight is (presumably un-intentionally) given to proxies with lower resolution.

D'Arrigo et al. (2006); DAR

Tree-rings only.
Proxies' latitude range: 45–72N.
Original calibration target: N20N (land only).
Original seasonal target: annual mean temperature.
Method: composite calibrated with OLS regression ('inverse calibration').
Temporal resolution: annual.
Calibration period: 1856–1978.
Reconstruction period: 713–1995.

ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/n_hem_temp/nhtemp-darrigo2006.txt

19 local/regional proxies, all with annual resolution. The number of local/regional proxies drops backwards in time; all nineteen records reach back to AD 1686 and six reach back at least to AD 1000.

Procedure: (i) tree-ring width (and limited density) data were derived from coniferous tree species from 66 high-elevation and latitudinal treeline sites, and used to form 19 regional composite RCS-standardized chronologies which were assessed for tree-growth signal strength, (ii) the regional chronologies were screened by comparisons with local or large-scale temperature and precipitation data to ensure a strong temperature signal and avoid significant correlations with precipitation, (iii) regional chronologies were separated into a North American and a European subset, (iv) for each continental subset, a 'nested' approach was adopted to account for a decrease in number of chronologies back in time; for each nest the chronologies were normalized to unit variance over the common period followed by averaging of the normalized regional series to create a nest mean, (v) the mean and variance of the longer nests where adjusted to match that of the most replicated nest and the nests were spliced together to form a North American and a European series, (vi) the two continental series were averaged to form a composite, (vii) the composite was calibrated by regressing the N20N (land only) annual-mean temperature on the proxy (i.e. 'inverse calibration') over the period 1856-1978. (Apparently, parts of the screening in step (ii) was undertaken before step (i).)

3 Instrumental data

In order to obtain a single calibration target to be used with all six selected hemispheric-scale reconstructions, we select the temperature series averaged over all land areas between latitudes 30°N and 75°N in the CRUTEM3 dataset of Brohan et al. (2006). This choice of region is well motivated as nearly all individual temperature proxy data used are located within the chosen region (see Figure 1 in the main article). None of the six hemispheric-scale reconstructions, however, were originally calibrated against this regional target, as the different authors chose either the full Northern Hemisphere or areas north of either 20°N or 30°N. Originally, all six hemispheric-scale reconstructions were calibrated against annual mean temperatures, although many individual proxy series mainly reflect summer temperatures or temperatures in the summer half-year. We allow for two alternatives and use both the full annual mean temperature and the mean for the six-month season April to September (AMJJAS). 10-year moving average series for these two temperature records are shown in Figures 1a and 2a.

4 Calibration period

We use a common calibration period for (re-)calibration of all records against the common instrumental target data. To this end we select non-overlapping 10-year mean temperatures in the period AD 1880-1959. The choice of non-overlapping 10-year means is dictated by the fact that the LJU reconstruction is published at a 10-year resolution, but it is reasonable to use decadal mean temperatures also because two of the other reconstructions (HEG, JON) were originally calibrated using decadal temperatures and were only presented as smoothed data. The start of the calibration period is motivated by the circumstance that the uncertainty in the instrumental temperature series is particularly large in this period (see section 5). Moreover, 1880 is the latest start point of calibration periods originally used among the authors of the selected reconstructions. The end of the calibration period is dictated by the HEG reconstruction, which has the earliest end point (in 1960). The resulting common calibration period is thus 80 years long, which means that the calibration is undertaken using samples with only eight 10-year mean temperatures. This is a small sample size, which implies a large statistical uncertainty. The effect of this uncertainty, on the comparison of proxy-data vs. climate model simulation data, is investigated by undertaking a bootstrap approach to derive an ensemble of calibrated reconstructions (see section 6).

5 Estimation of the instrumental measurement error variance σ_θ^2

Brohan et al. (2006) discuss several sources of uncertainty in their global dataset of instrumental temperatures, which are illustrated for different temporal scales and global mean temperatures in their Figure 10. Here, we describe how these error estimates can be used within the framework of Sundberg et al. (2012), which allows an errors-in-variables model in the calibration of the proxy data.

Through personal communication with Philip Brohan (in June 2012), we have obtained data files with the same types of error estimates as shown in their Figure 10, but which are valid for the regional and seasonal definitons and temporal scale chosen for our analysis; i.e. 10-year mean temperatures averaged (areally-weighted) over all grid-boxes between latitudes 30°N and 75°N in the land-only dataset CRUTEM3, for both annual mean and AMJJAS temperatures. Figures 1a and 2a show the central temperature estimates together with two error bands representing 95% uncertainty ranges for two types of errors, using the same colour coding as in Brohan et al. (2006). The two error types are:

- uncertainty caused by station, sampling (within grid-boxes), and measurement errors (red band).
- additional uncertainty due to limited coverage (i.e. grid-boxes with missing data) (green band).

Brohan et al. (2006) also quantify a third error type, due to bias errors. However, we do not consider this error type in our analysis because it does not represent a simple variance term added to the 'true' temperatures.

As can be seen in Figures 1a and 2a, unique error estimates are provided for each time point. However, for (re-)calibration of the hemispheric-scale temperature reconstructions, we need a single estimate of the error variance in order to use eq. (6) in Sundberg et al. (2012); here labeled eq. (1):

$$\widehat{\beta}_0 = \frac{s_{yz_0}}{s_y^2 - \widehat{\sigma}_\theta^2}\,. \tag{1}$$

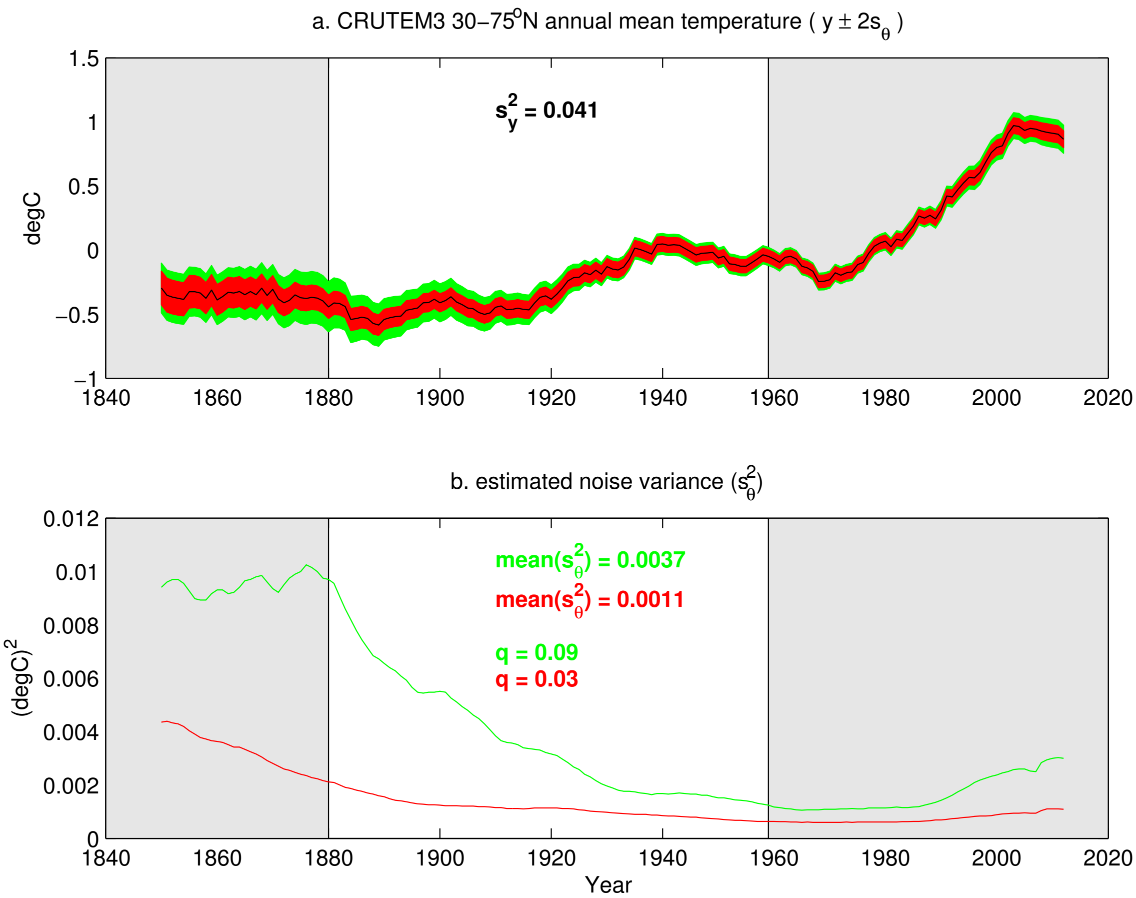

Figure 1: **a.** Running 10-year annual mean temperatures averaged over the land areas between 30°N and 75°N from the CRUTEM3 dataset of Brohan et al. (2006), expressed as anomalies from the mean over the 1961-1990 period. The temperature estimate (black line) is surrounded by two 95% uncertainty bands; due to station, sampling and measurement errors (red band) and the additional uncertainty due to limited coverage (green band). The sample variance of the temperatures (s_y^2) within the period 1880-1959 is indicated in the graph. **b.** Time sequences of the estimated error variances ($\widehat{\sigma}_\theta^2$, here denoted s_θ^2) deduced from the two types of errors shown in the upper panel (red and green lines). The mean value of these error variances in the period 1880-1959 are indicated, as well as the corresponding variance ratios $q = s_\theta^2/s_y^2$ (red and green text).

Here, the crucial quantity is $\widehat{\sigma}_\theta^2$, i.e. the estimate of the variance of the error that is added to the 'true' (but unobservable) temperatures (τ) for the chosen region and season. The other two components of eq. (1) are the sample covariance (s_{yz_0}) between the observed temperatures (y) and the un-calibrated proxy (z_0), and the sample variance (s_y^2) of the observed temperatures. Following an advice by Philip Brohan (pers. comm., January 2012), we regard the combined error due to station, sampling, measurement and coverage errors (i.e. the errors represented by the red and green bands together) as being most relevant to use when estimating σ_θ^2. Following the notation of Sundberg et al. (2012), the central estimates in Figures 1a and 2a can be denoted y, while the interval between the uppermost and lowermost limit of the green band is $y \pm 2\widehat{\sigma}_\theta$. Hence, the width of this interval can straightforwardly be used to quantify $\widehat{\sigma}_\theta^2$ at each time point. The resulting time sequence of the instrumental error variance estimate is presented with green lines in Figures 1b and 2b.

It is immediately seen that $\widehat{\sigma}_\theta^2$ varies considerably with time. For both annual and AMJJAS temperatures the error variance is largest before around AD 1880 and smallest in the period 1961-1990. The reasons for this behaviour are discussed in detail by Brohan et al. (2006). Most important here are likely the effect of

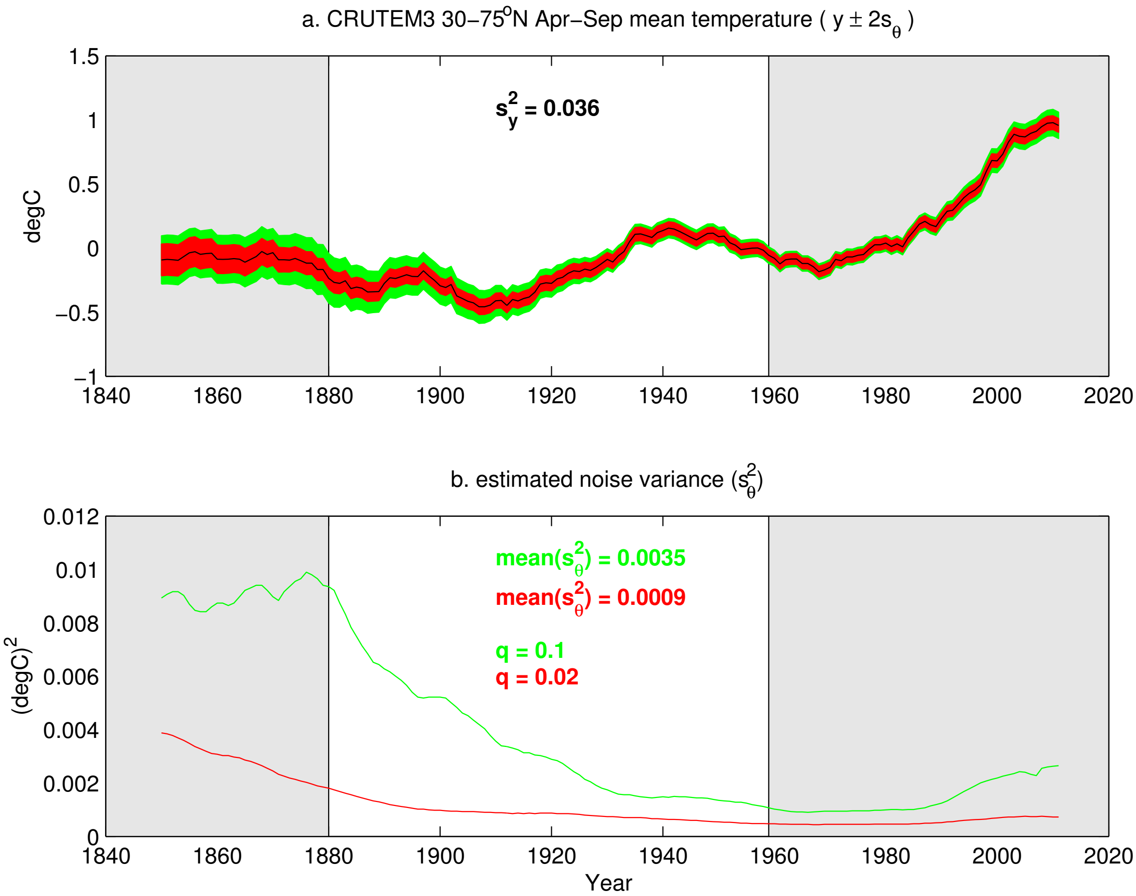

Figure 2: As Figure 1 of this supplementary material, but for AMJJAS temperatures

.

fewer observational sites and fewer grid-boxes with data before around AD 1880. The small errors in the 1961-1990 period is partly related to the fact that all temperatures are expressed as departures from the mean over this period.

Because $\widehat{\sigma}_\theta^2$ varies with time, it is necessary to find a value that is valid for the calibration period of interest. The mean value of $\widehat{\sigma}_\theta^2$ within the calibration period is theoretically the most reasonable value to use. Consequently, we use the mean value of $\widehat{\sigma}_\theta^2$ in the 1880-1959 period as our estimate. For annual mean temperatures, we obtain $\widehat{\sigma}_\theta^2 = 0.0037(°\mathrm{C})^2$ and for AMJJAS we get $\widehat{\sigma}_\theta^2 = 0.0035(°\mathrm{C})^2$.

Of particular interest is the ratio between the instrumental error variance and the variance of the observed temperatures, $q = \sigma_\theta^2/s_y^2$. This ratio is used in Sundberg et al. (2012) in the defintion of weights to be applied when using calibrated proxy data in the calulation of their distance and correlation measures. In the period 1880-1959, the sample variance of the 10-year running annual mean temperatures is $0.041(°\mathrm{C})^2$ and for the AMJJAS temperatures it is $0.036(°\mathrm{C})^2$. By using these numbers and the above obtained estimates of σ_θ^2, we find that $q = 0.09$ for the annual mean temperatures and $q = 0.10$ for AMJJAS temperatures. Thus, in both cases, the observational errors account for approximately 10% of the total variance in the observed 10-year mean temperatures.

If we, alternatively, exclude the errors due to incomplete coverage (the green bands in Figures 1a and 2a) and consider only the errors due to station, sampling and measurement errors (the red bands), then we instead obtain $q = 0.03$ for the annual mean temperatures and $q = 0.02$ for AMJJAS temperatures. Thus, these errors together account for only about two to three percent of the total variance in our calibration period.

In our experiments, we use both the larger (red plus green) and the smaller (red only) instrumental error estimates to study if the results are dependent on how $\widehat{\sigma}_\theta^2$ is chosen.

Finally, we should mention that eq. (1) can be re-written, using the above given definition of q, as:

$$\widehat{\beta}_0 = \frac{s_{yz_0}}{s_y^2(1-q)} \,. \tag{2}$$

In our bootstrap analysis of the calibration (see below), we use this alternative formulation where we let q be a fixed quantity (taken as one of the four values given above, depending on the actual experiment), and thus we only need to estimate s_{yz_0} and s_y^2 for each re-sampling of the calibration data set.

6 Ensemble calibration experiments

To investigate how the results of comparisons of the Northern Hemispheric-scale reconstructions and climate model simulations are affected by the choice of season (annual vs. AMJJAS) and choice of instrumental error variance (including or excluding the uncertainty due to limited coverage – referred to below as 'large q' and 'small q'), we undertake four experiments. For each experiment, we produce an ensemble of calibrated reconstructions using a bootstrap approach, where we resample (with replacement) the calibration dataset 100 times. This makes it possible to account for the statistical uncertainty in the calibration, in a fashion similar to the ensemble calibration used by Frank et al. (2010). Their procedure, however, was designed to systematically changing the length of the calibration period and shifting the calibration time window to obtain different calibration periods. Our calibration period, which is chosen to be longest feasible that is common for all six reconstructions, is too short to allow exactly the same procedure. Instead, we resample the same single calibration period, using sampling with replacement. The result is conceptually similar to theirs – an ensemble consisting of many versions of the calibrated reconstructions, which differ by the regression coefficients used but where all are within the limits determined by the available calibration dataset.

The ensemble calibration approach provides, for each reconstruction, a sample of 100 values of every quantity of interest. This makes it possible to study the distribution of, e.g., the sample correlation between the proxy and the instrumental data ($\widehat{\rho}$), the scaling factor that converts each reconstruction from its original numeric values to represent the new common calibration target ($\widehat{\beta}_0^{-1}$), or the statistical measures in the main analysis of the proxy data and the climate model simulations (U_R and U_T).

The very small size of the calibration data set (eight pairs of 10-year mean temperature and proxy values) turned out to cause some problems, due to the large associated statistical uncertainty. In particular, some random re-samplings were found to lead to very weak, or even occasionally negative, values of $\widehat{\rho}$. Although this is certainly allowable from a pure statistical point of view, very low (insignificant) or even negative correlations would imply that the proxy record is not a useful proxy for the temperature series in question. This, however, is in conflict with prior knowlege about the proxy data; the original authors have demonstrated and concluded that their reconstructions are valid proxies for large-scale temperature averages within the Northern Hemisphere. We do not see any reason to question these conclusions and judgements. Thus, to avoid physically meaningless results from the re-sampling of the very small calibration dataset, we inserted a routine to check if $\widehat{\rho} < 0.2$; if this happened, we dismissed that re-sampling and generated a new one. This happened only a few times, as in most cases the sample correlations were considerably larger (see Figures 3–6 below).

While undertaking the ensemble calibration, we also noted a source of conflict between the estimated values of ρ^2 and q. Under the assumptions made in the statistical models in Sundberg et al. (2012) – in particular the assumed independence of the various error terms – it is not possible for *true* values that ρ^2 is larger than $(1-q)$. However, we found that this could happen for the corresponding *estimated* values. We did not investigate the reasons for this (although the small size of calibration data sample may be one of the contributing factors), but to avoid violations to the assumptions we included a routine to check (for each re-sampling of the calibration dataset) if this happens. If it happened, we kept the estimate $\widehat{\rho^2}$ unchanged and forced q to equal $(1 - \widehat{\rho^2})$. Obviously, this problem occurred more often for the experiments with 'large q' than for those with 'small q'.

We provide below some information about the results of each of the four experiments. In all cases, the instrumental data are the temperatures averaged over land areas between 30-75°N from the CRUTEM3 dataset and the calibration period is 1880-1959, where the calibration data sample is obtained by resampling the pairs of mean values for the eight non-overlapping 10-year periods.

The four experiments are:

- exp. 1: Annual temperature, small q
- exp. 2: Annual temperature, large q
- exp. 3: AMJJAS temperature, small q
- exp. 4: AMJJAS temperature, large q

Tables 1–3 give the mean values of $\widehat{\rho}$, $\widehat{\rho^2}$ and $\widehat{\beta_0}^{-1}$ for each 100-member ensemble in each experiment. Frequency distributions (histograms) for the corresponding elements in the full ensembles are given in Figures 3–6.

Some observations of interest can be made. To begin, the ensemble-mean correlations between the reconstructions and the instrumental temperatures are always stronger for annual temperatures than for AMJJAS. For annual temperatures, the correlation is at least 0.96 for all reconstructions except JON, where it is 0.92 and 0.93 in the two experiments respectively. The correlations with AMJJAS temperatures are not much weaker; they are always at least 0.90, except for JON (0.86 and 0.87). By all means, all ensemble-mean correlations are strong, and the histograms of $\widehat{\rho}$ show that correlations below 0.7 are rare for individual re-samplings. The warming trend throughout the calibration period certainly helps to explain these strong correlations.

As regards the conflict between too high q and $\widehat{\rho^2}$, this happens for annual temperatures when $\widehat{\rho^2} > 0.91$ in the 'large q' case and when $\widehat{\rho^2} > 0.97$ in the 'small q' case. Thus, because the correlations are mostly very strong, these conflicts occur often for the 'large q' case. Because we reduced the q-value whenever the conflict occurred, the possibility to study the difference between the experiments with 'large q' and 'small q' becomes somewhat reduced. For AMJJAS temperatures the conflicts are less common because of the weaker correlations in this season, as clearly illustrated when inspecting the histograms of $\widehat{\rho^2}$.

Due to the re-calibration of all reconstructions against a common target and the use of a single method, the amplification or diminuation of the amplitudes in the reconstructions compared to their original values (i.e. the scaling factor $\widehat{\beta_0}^{-1}$) are sometimes substantial. Only in the case of HEG, the re-calibration cause just a very small change; with an ensemble-mean scaling factor of between 1.03 and 1.07 depending on the experiment considered. The amplitude change was rather small also for DAR; a scaling factor between 0.83 and 0.91. However, the amplitude of CHR was drastically reduced with a mean factor between 0.35 and 0.37, while the amplitude of JON was roughly doubled. JUC and LJU were also amplified relative to their original values, with factors between 1.25 and 1.46 for LJU and between 1.53 to 1.61 for JUC. These findings can be understood by considering how the reconstructions were originally calibrated (to different targets and with different methods) and how they were calibrated here (against a single target and according to the theory of Sundberg et al. (2012)). The important consequence of our re-calibration is that the results of our main model vs. proxy data comparisons are not dependent on the choice of calibration method and target area and season, as all reconstructions are calibrated in a consistent way to represent the same aspect of large-scale temperatures and to provide an unbiased ranking of the simulations according the the theory of Sundberg et al. (2012). The final results are instead mainly affected by the choice of individual proxy series and how they were combined into a single composite series by the original authors.

exp.	CHR	JUC	HEG	JON	LJU	DAR
1	0.96	0.96	0.97	0.93	0.96	0.97
2	0.96	0.97	0.98	0.92	0.96	0.97
3	0.93	0.92	0.91	0.87	0.94	0.93
4	0.93	0.90	0.88	0.86	0.94	0.92

Table 1: Mean $\widehat{\rho}$ for experiment 1–4

exp.	CHR	JUC	HEG	JON	LJU	DAR
1	0.93	0.93	0.95	0.87	0.93	0.95
2	0.92	0.94	0.95	0.86	0.92	0.94
3	0.87	0.84	0.83	0.76	0.89	0.86
4	0.86	0.82	0.79	0.75	0.89	0.84

Table 2: Mean $\widehat{\rho^2}$ for experiment 1–4

exp.	CHR	JUC	HEG	JON	LJU	DAR
1	0.37	1.61	1.06	2.05	1.46	0.91
2	0.37	1.53	1.03	1.96	1.39	0.89
3	0.35	1.60	1.07	2.09	1.39	0.89
4	0.35	1.57	1.04	1.96	1.25	0.83

Table 3: Mean $\widehat{\beta}_0^{-1}$ for experiment 1–4

6.1 Conclusions for the choice of experiment to highlight

Based on the discussion above, together with the observations from sections 5 and 6, and considering that many of the individual proxy series are representative of the summer season or an extended summer season, we decide to choose results for the AMJJAS season 'large q' case for being highlighted in our main text. Although the stronger correlations between the reconstructions and instrumental data for annual data – and also the fact that all six reconstructions were originally calibrated against annual mean temperatures – would speak for calibration against annual temperatures, we find AMJJAS to be a preferable choice as this leads to a less often occurring conflict between q and $\widehat{\rho^2}$. This, in turn, means that we most often can use the estimated instrumental errors ('large q') as devised by Brohan et al. (2006) without artificial modifications. We show the results for the annual mean and large q experiment in the supplementary Section 8 for the sake of comparison with the main text. We found that for a given season, there is almost no difference between results obtained with 'large q' and 'small q'. There are some differences between the choice of annual (supplementary) vs. AMJJAS (main article) temperatures, as discussed further in Section 8. We make efforts in the main text to point out in what way the choice of season affect the result of the study.

7 Autocorrelation test for the COSMOS CTRL simulation

In order to check whether or not the assumption in Sundberg et al. (2012) of white noise behaviour in the unforced (CTRL) simulated surface temperature δ holds, we calculated the lag-1 autocorrelation for land-only 30-75°N averages with April-September (Figure 7) and annual (Figure 8) means. It was found that in both cases, δ can be considered white noise at the 20 year time-resolution in keeping with the statistical assumptions of this investigation.

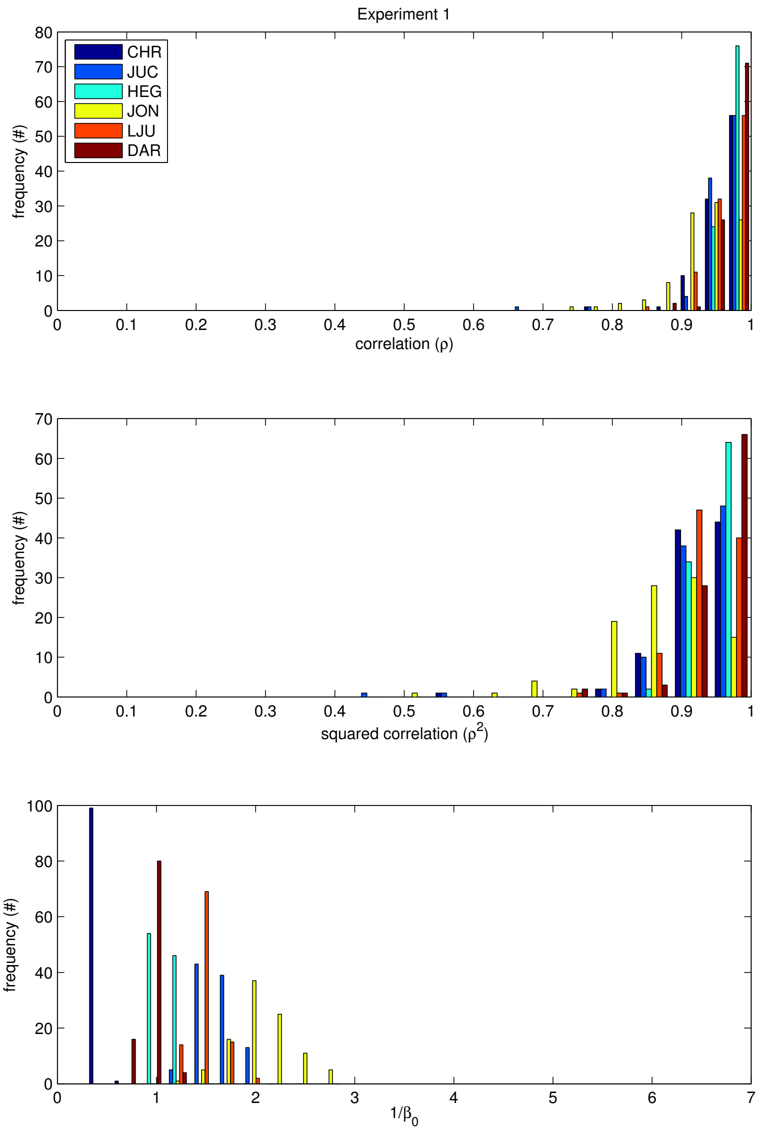

Figure 3: Frequency distributions of $\widehat{\rho}$ (top), $\widehat{\rho^2}$ (middle) and $\widehat{\beta}_0^{-1}$ (bottom) for experiment 1 (Annual temperature, small q).

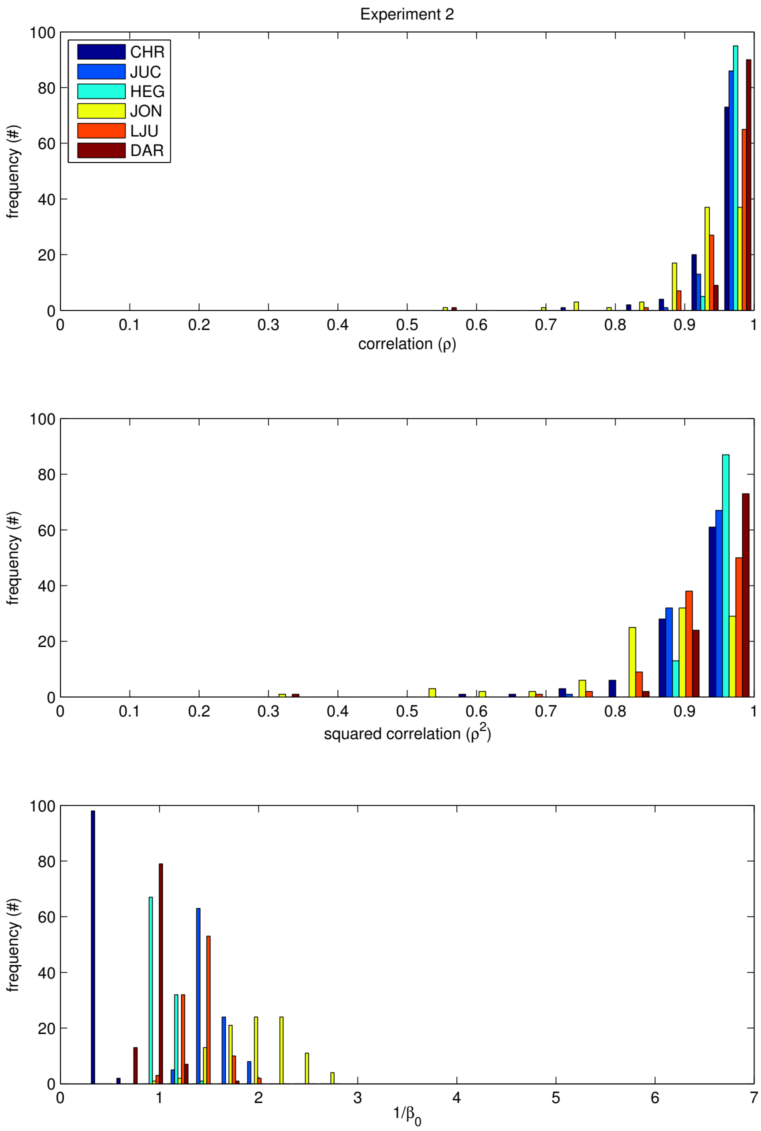

Figure 4: Frequency distributions of $\widehat{\rho}$ (top), $\widehat{\rho^2}$ (middle) and $\widehat{\beta}_0^{-1}$ (bottom) for experiment 2 (Annual temperature, large q).

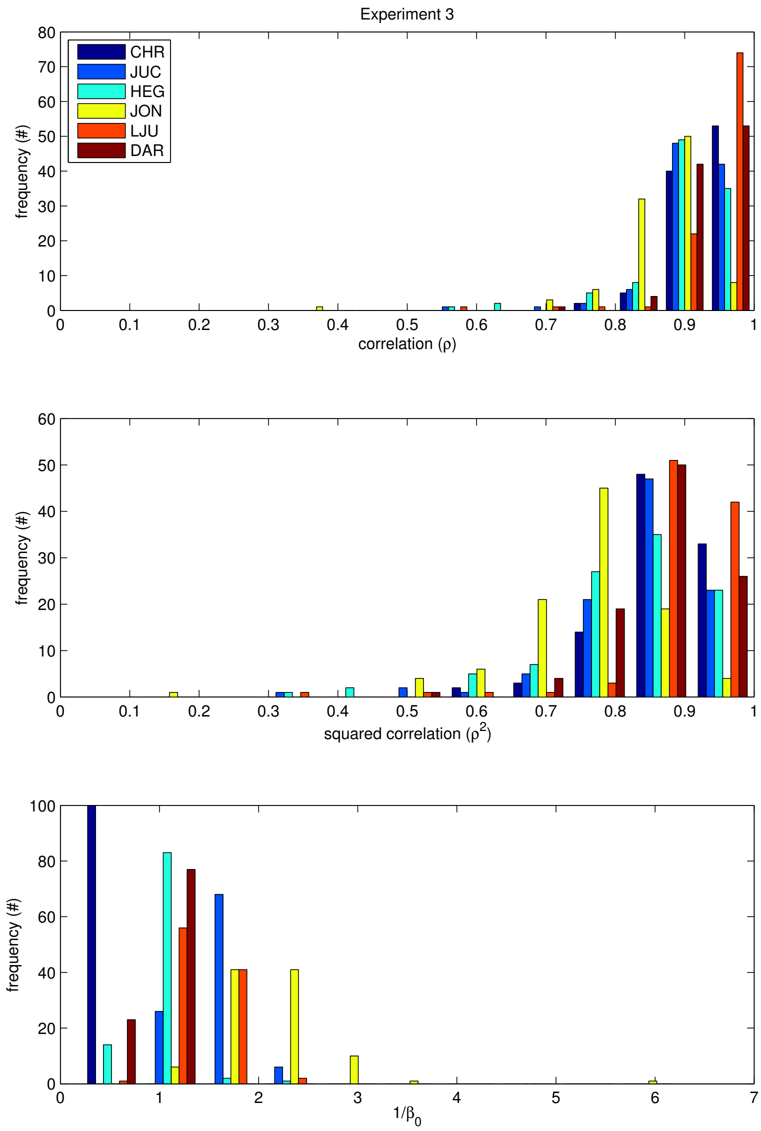

Figure 5: Frequency distributions of $\widehat{\rho}$ (top), $\widehat{\rho^2}$ (middle) and $\widehat{\beta}_0^{-1}$ (bottom) for experiment 3 (AMJJAS temperature, small q).

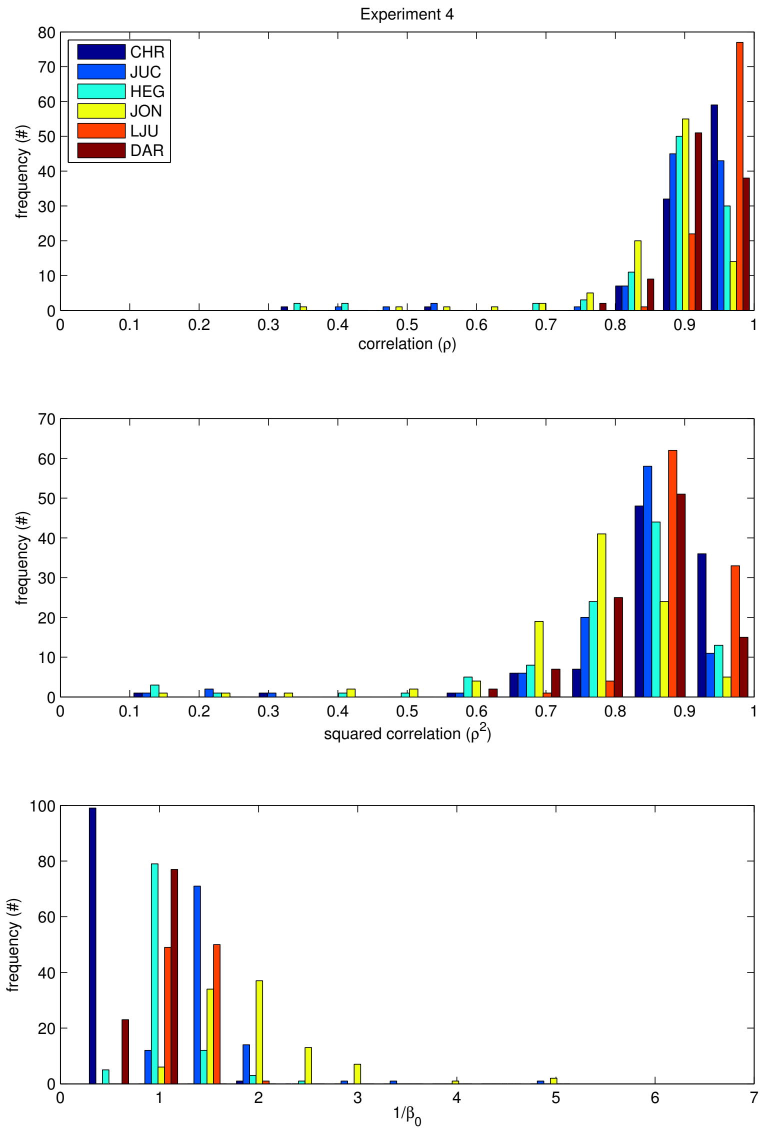

Figure 6: Frequency distributions of $\widehat{\rho}$ (top), $\widehat{\rho^2}$ (middle) and $\widehat{\beta}_0^{-1}$ (bottom) for experiment 4 (AMJJAS temperature, large q).

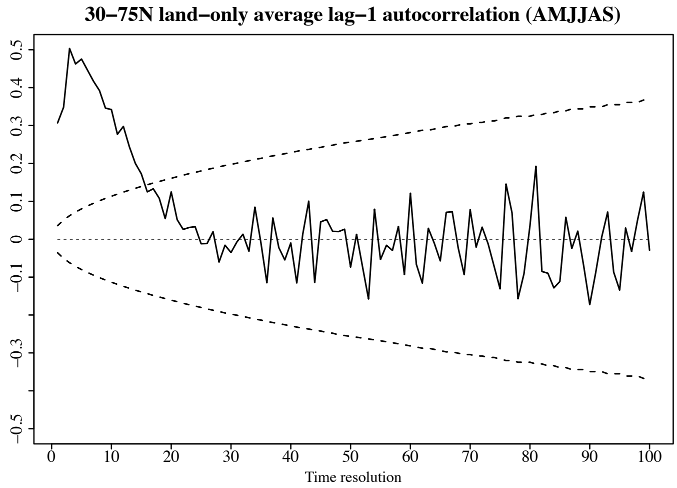

Figure 7: Lag-1 autocorrelation of the April-September (AMJJAS) mean land-only 30-75°N average 3000-yr CTRL simulation against changing time resolution. The curved dashed lines represents the 0.05 (two-sided significance level).

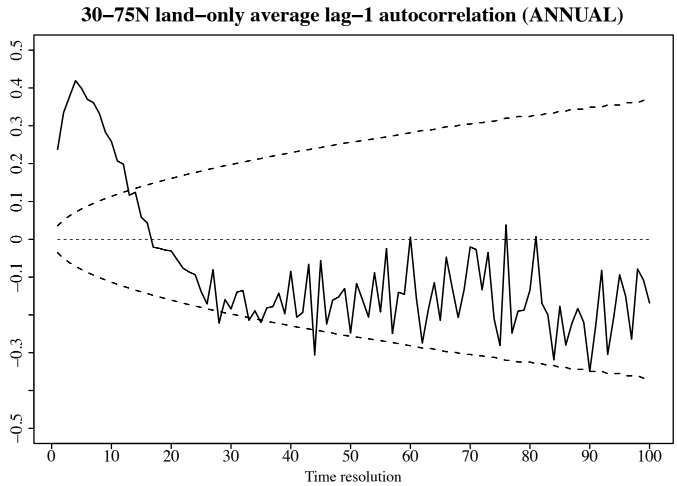

Figure 8: As Figure 7 of this supplementary material, but for annual averages.

8 Additional figures from the model vs. data comparisons, not shown in the main paper

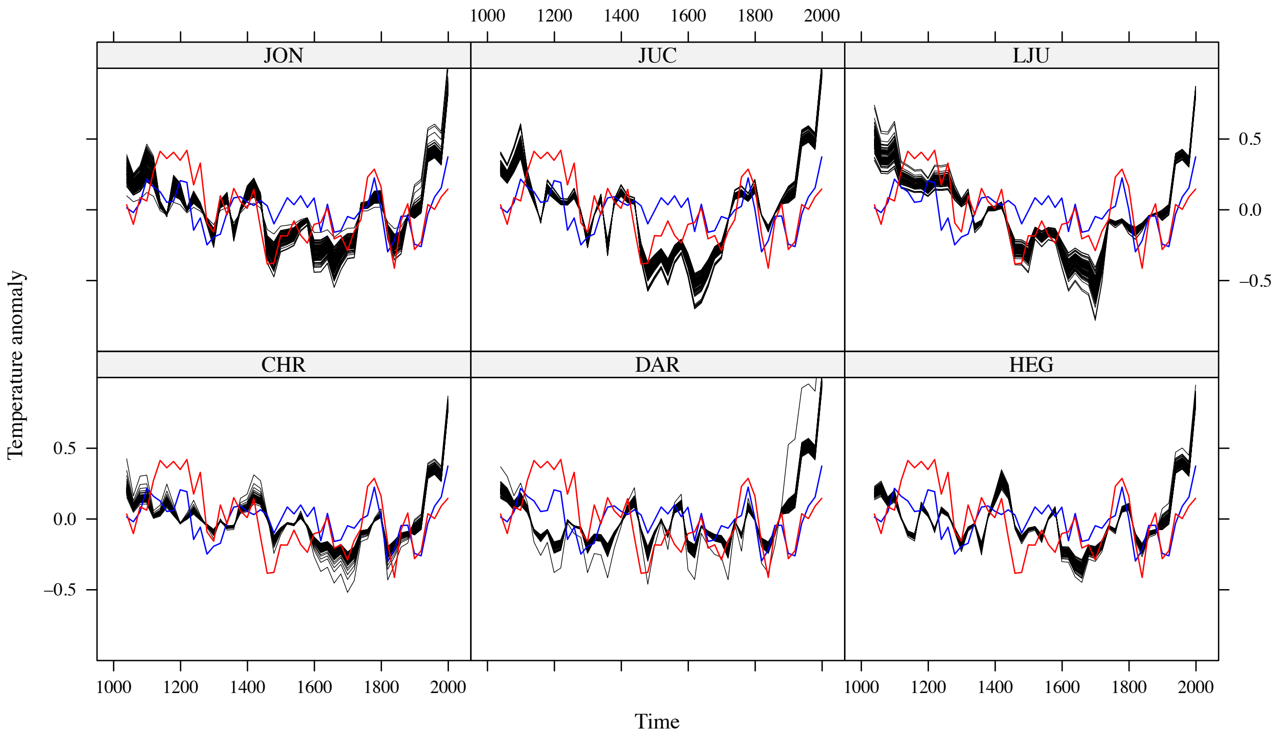

Figure 9: As Figure 2 in the article but for annual averages.

8.1 Simulation and reconstruction time series

Figure 9 represents the annually averaged, 20-year non-overlapping mean 30-75°N time series for the E1 (blue) and E2 (red) COSMOS simulations against 100 alternative bootstrap (with replacement) calibrated hemispheric proxy reconstructions for Jones and Mann (2004) (JON), Juckes et al. (2007) (JUC), Ljungqvist (2010) (LJU), Christiansen and Ljungqvist (2011) (CHR), D'Arrigo et al. (2006) (DAR) and Hegerl et al. (2007) (HEG). The simulations are very similar to the AMJJAS mean equivalents seen in Figure 2 of the main article, though there is clearly less calibration uncertainty in the reconstructions using annual mean instrumental data as a target, associated with the generally higher correlation scores (Section 6).

8.2 U_R and U_T graphs

Figure 10 shows the annual-mean U_R correlation test statistics comparing E1 (blue) and E2 (red) ensemble simulation mean time-series with the (re-)calibration ensembles of the six hemispheric reconstructions (one hundred replicants) over the periods 1000-2000, 1000-1850, 1400-2000 and 1400-1850. The capitalized reconstruction abbreviated names are the same as used previously. The dashed lines indicate statistical significance at the 0.05 two-sided levels. It can be concluded from Figure 10 that there is a similarly significant positive correlation between the E1/E2 simulations and the target hemispheric-scale reconstructions, as compared with the AMJJAS analysis of the main article. Figure 11 shows the U_T test statistic distributions for the annual means (though is otherwise an identical analysis to that of Figure 4 in the main article).

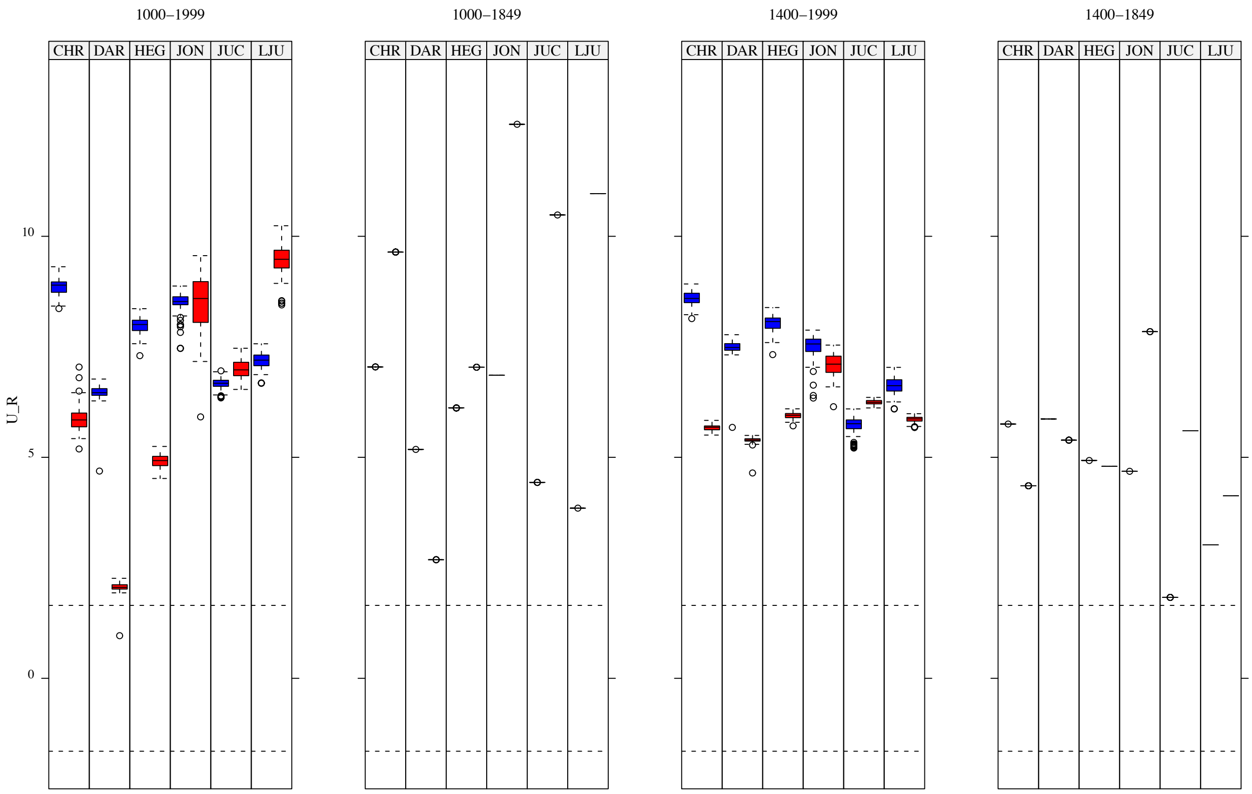

Figure 10: As Figure 3 in the article using a large $q = 0.09$ but for annual averages (U_R). Namely exp. 2 as described in Section 6 of this supplementary material.

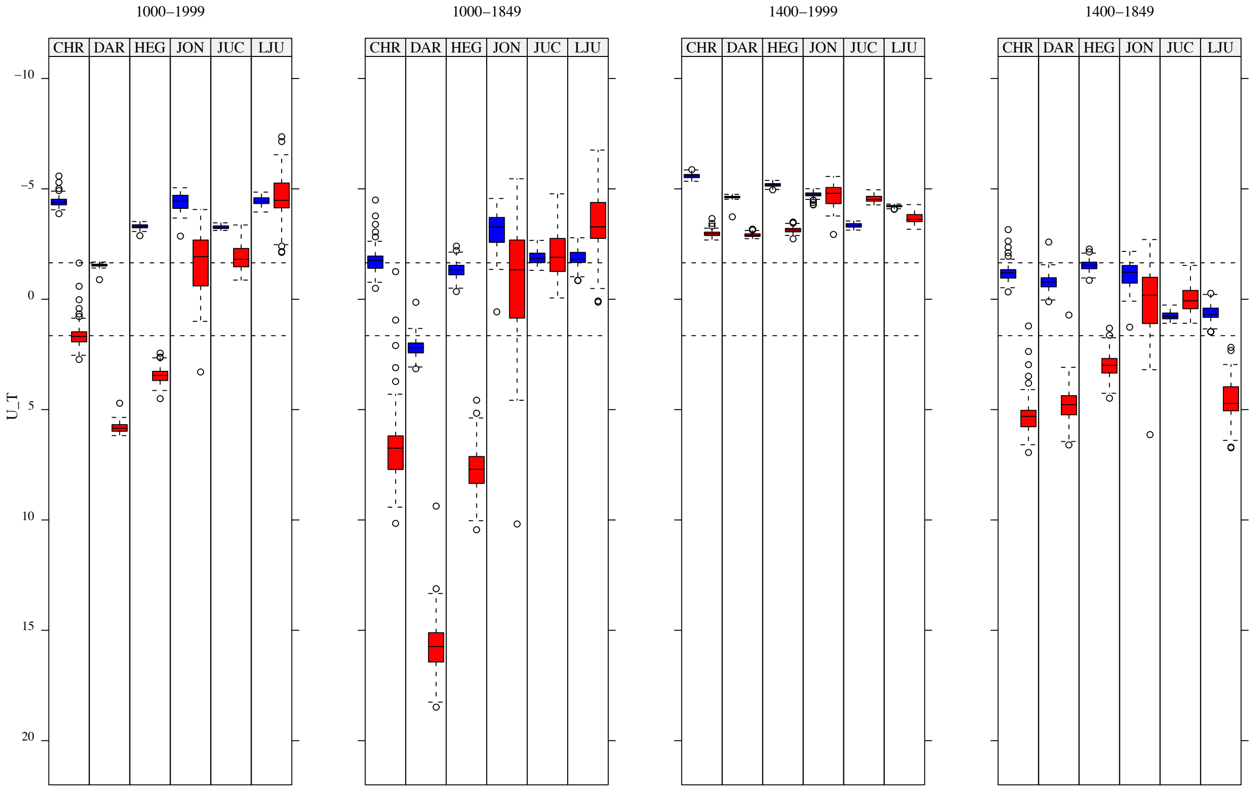

Figure 11: As Figure 4 in the article using a large $q = 0.09$ but for annual averages (U_T). Namely exp. 2 as described in Section 6 of this supplementary material.

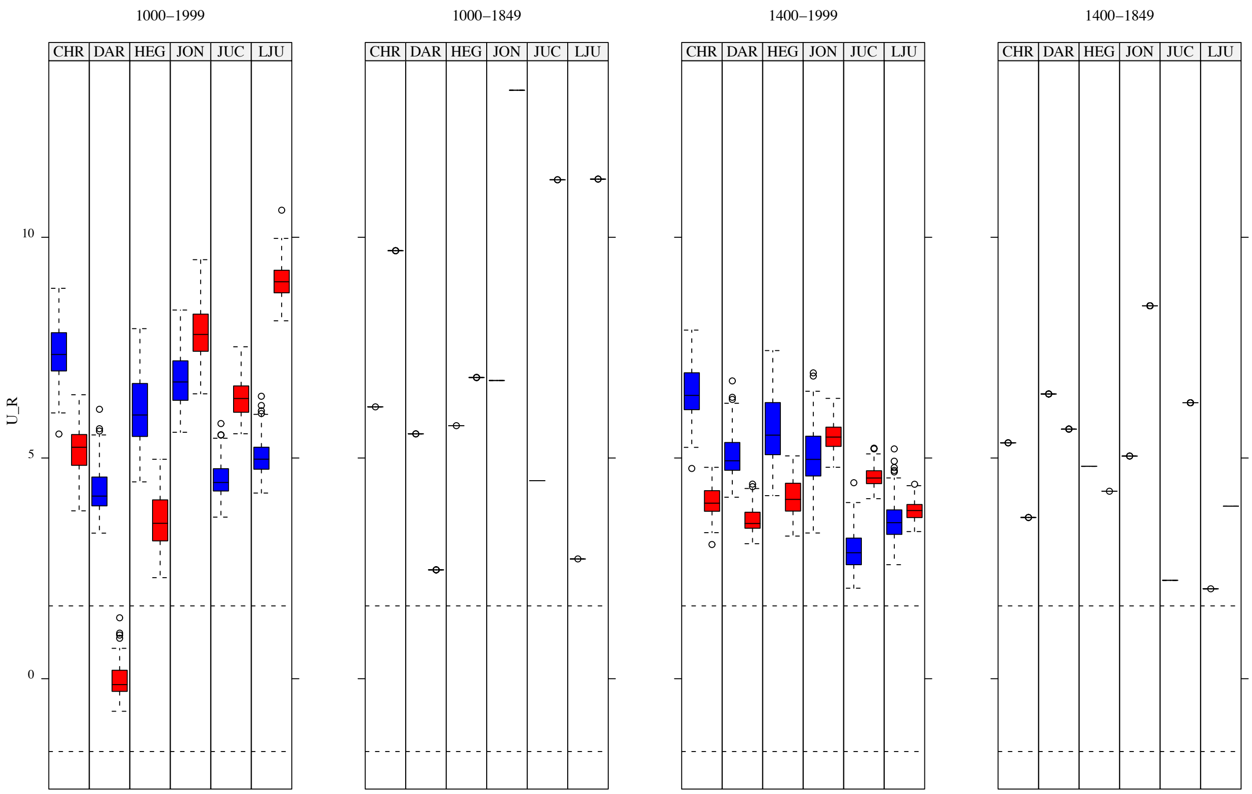

Figure 12: As Figure 3 in the article (U_R), but using a smaller $q = 0.02$ estimated from station & grid-box sampling errors (exp. 3 in Section 6).

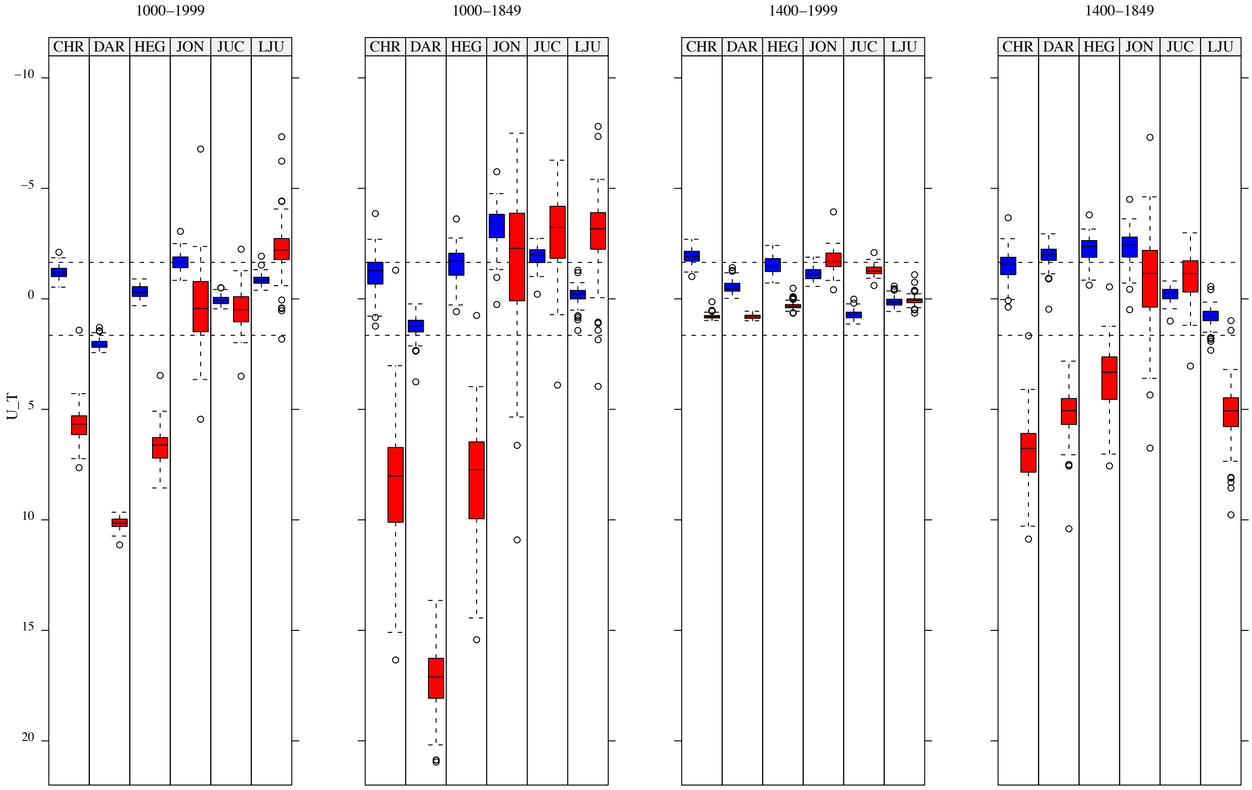

Figure 13: As Figure 4 in the article (U_T), but using a smaller $q = 0.02$ estimated from station & grid-box sampling errors (exp. 3 in Section 6).

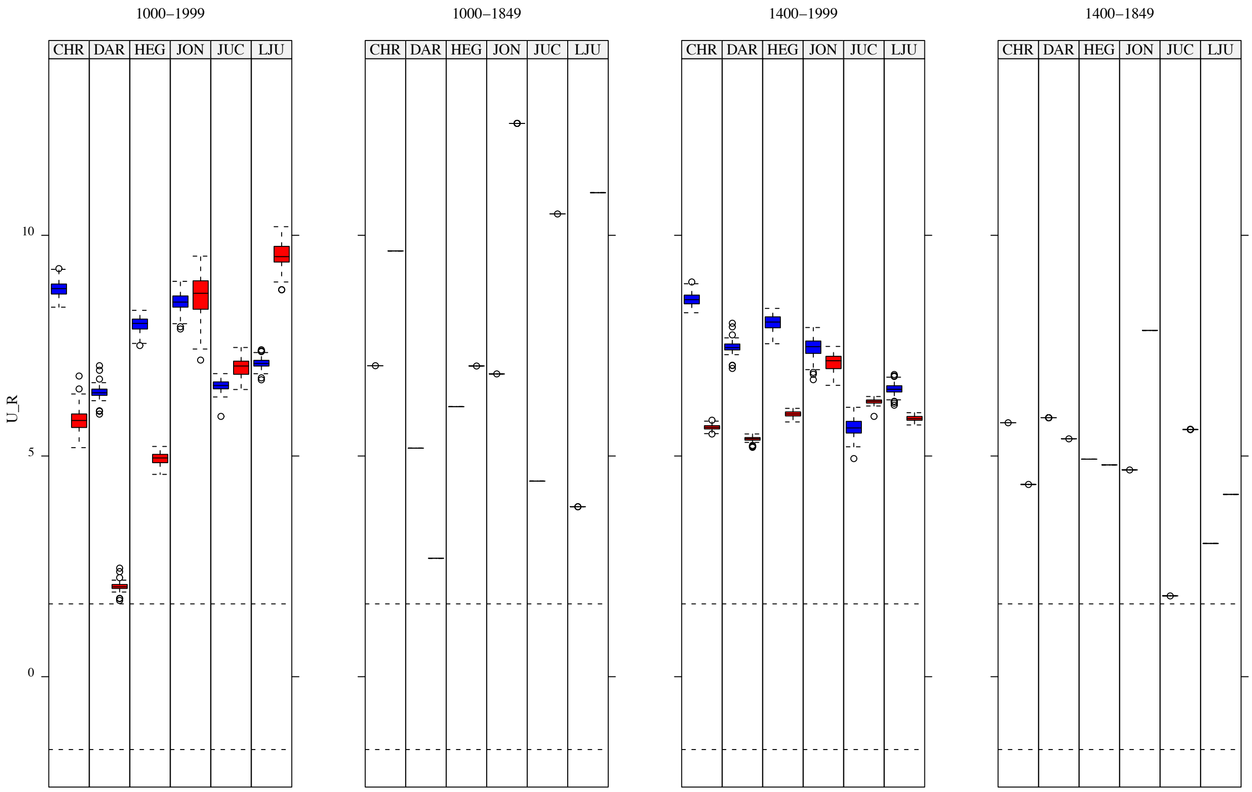

Figure 14: As Figure 10 in this supplement (U_R) based on annual temperatures using a smaller $q = 0.03$ (exp. 1 in Section 6).

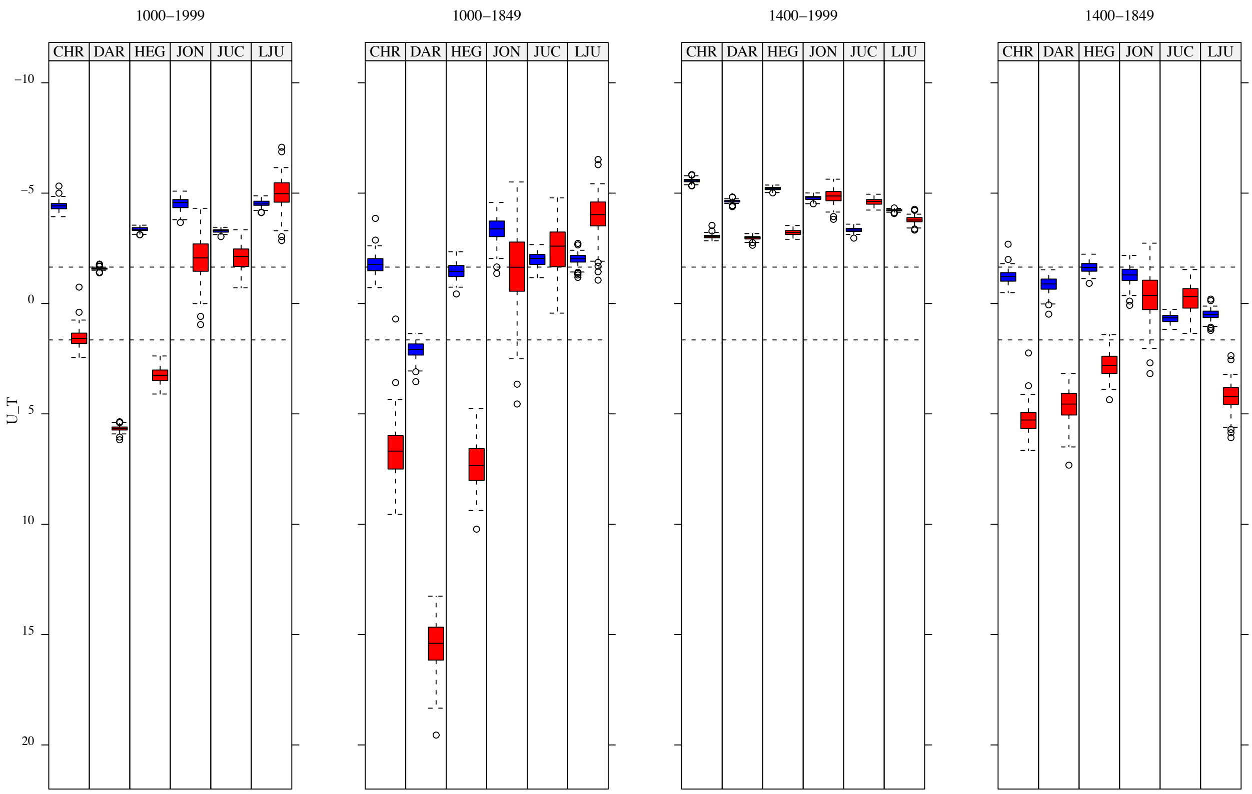

Figure 15: As Figure 11 in this supplement (U_T) based on annual temperatures using a smaller $q = 0.03$ (exp. 1 in Section 6).

Here, the ranking of E1 "low" solar and E2 "high" solar simulations is largely unchanged from the main analysis results. Namely, that E1 simulations are often closer to the target hemispheric-scale series than the E2 simulations. However, these results contrast with the AMJJAS mean results of the main text, in that the E1 and E2 simulations more frequently outperform the CTRL simulations when the instrumental period is included in the analysis. We also include the experiments featuring smaller q values (i.e. excluding error associated with coverage uncertainty) for both the summer AMJJAS (Figure 12 and 13) and annual averages (Figure 14 and 15). The lower q values result in a reduced spread associated with calibration uncertainty for all comparisons, but otherwise there is no notable difference when compared with the results using a larger q.

References

Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., Jones, P. D., 2006. Uncertainty estimates in regional and global observed temperature changes: A new data seet from 1850. Journal of Geophysical Research 111 (D12106).

Christiansen, B., Ljungqvist, F. C., 2011. Reconstruction of the extra-tropical NH mean temperature over the last millennium with a method that preserves low-frequency variability. Journal of Climate 24, 6013–6034.

D'Arrigo, R. D., Wilson, R., Jacoby, G., 2006. On the long-term context for late twentieth century warming. Journal of Geophysical Research 111 (D3).

Frank, D. C., Esper, J., Raible, C. C., Büntgen, U., Trouet, V., Stocker, B., Joos, F., 2010. Ensemble reconstruction constraints on the global carbon cycle sensitivity. Nature 463, 527–530.

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

Jones, P. D., Mann, M. E., 2004. Climate over past millennia. Reviews of Geophysics 42, RG2002.

Juckes, M. N., Allen, M. R., Briffa, K. R., Esper, J., Hegerl, G. C., Moberg, A., Osborn, T. J., Weber, S. L., 2007. Millennial temperature reconstruction intercomparison and evaluation. Climate of the Past 3 (4), 591–609.

Jungclaus, J. H., Lorenz, S. J., Timmreck, C., Reick, C. H., Brovkin, V., Six, K., Segschneider, J., Giorgetta, M. A., Crowley, T. J., Pongratz, J., Krivova, N. A., Vieira, L. E., Solanki, S. K., Klocke, D., Botzet, M., Esch, M., Gayler, V., Haak, H., Raddatz, T. J., Roeckner, E., Schnur, R., Widmann, H., Claussen, M., Stevens, B., Marotzke, J., 2010. Climate and carbon-cycle variability over the last millennium. Climate of the Past 6 (5), 723–737.

Ljungqvist, F. C., 2010. A new reconstruction of temperature variability in the extra-tropical northern hemisphere during the last two millennia. Geografiska Annaler 92 A (3), 339–351.

Mann, M. E., Jones, P., 2003. Global surface temperatures over the past two millennia. Geophysical Research Letters 30 (1820).

Sundberg, R., Moberg, A., Hind, A., 2012. Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium - part 1: Theory. Climate of the Past 8, 1339–1353.