Ranking of tree-ring based temperature reconstructions of the past millennium

Jan Esper [a,*], Paul J. Krusic [b, c], Fredrik C. Ljungqvist [d, e], Jürg Luterbacher [f], Marco Carrer [g], Ed Cook [h], Nicole K. Davi [h, i], Claudia Hartl-Meier [a], Alexander Kirdyanov [j], Oliver Konter [a], Vladimir Myglan [k], Mauri Timonen [l], Kerstin Treydte [m], Valerie Trouet [n], Ricardo Villalba [o], Bao Yang [p], Ulf Büntgen [m]

[a] *Department of Geography, Johannes Gutenberg University, 55099 Mainz, Germany*
[b] *Department of Physical Geography, Stockholm University, 10691 Stockholm, Sweden*
[c] *Navarino Environmental Observatory, Messinia, Greece*
[d] *Department of History, Stockholm University, 10691 Stockholm, Sweden*
[e] *Bolin Centre for Climate Research, Stockholm University, 10691 Stockholm, Sweden*
[f] *Department of Geography, Justus-Liebig University, 35390 Giessen, Germany*
[g] *Università degli Studi di Padova, Dipartimento Territorio e Sistemi AgroForestali, 35020 Legnaro, Italia*
[h] *Tree Ring Laboratory, Lamont-Doherty Earth Observatory, Palisades, NY 10964, USA*
[i] *Department of Environmental Science, William Paterson University, Wayne, NJ 07470, USA*
[j] *Institute of Forest SB RAS, Akademgorodok, Krasnoyarsk, 660036, Russia*
[k] *Institute for the Humanities, Siberian Federal University, Krasnoyarsk, 660041, Russia*
[l] *Natural Resources Institute Finland (Luke), Rovaniemi Unit, Rovaniemi, Finland*
[m] *Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland*
[n] *Laboratory of Tree-Ring Research, University of Arizona, Tucson, AZ 85721, USA*
[o] *Instituto Argentino de Nivología, Glaciología y Ciencias Ambientales, CCT-Mendoza, 5500 Mendoza, Argentina*
[p] *Key Laboratory of Desert and Desertification, Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou 730000, China*

Submitted as a **Review Paper** to *Quaternary Science Reviews*

* Corresponding author: esper@uni-mainz.de, Tel. +49 6131 3922296

Abstract

Tree-ring chronologies are widely used to reconstruct high- to low-frequency variations in growing season temperatures over centuries to millennia. The relevance of these timeseries in large-scale climate reconstructions is often determined by the strength of their correlation against instrumental temperature data. However, this single criterion ignores several important quantitative and qualitative characteristics of tree-ring chronologies. Those characteristics are *(i) data homogeneity, (ii) sample replication, (iii) growth coherence, (iv) chronology development*, and *(v) climate signal* including the correlation with instrumental data. Based on these 5 characteristics, a reconstruction-scoring scheme is proposed and applied to 39 published, millennial-length temperature reconstructions from Asia, Europe, North America, and the Southern Hemisphere. Results reveal no reconstruction scores highest in every category and each has their own strengths and weaknesses. Reconstructions that perform better overall include N-Scan and Finland from Europe, E-Canada from North America, Yamal and Dzehlo from Asia. Reconstructions performing less well include W-Himalaya and Karakorum from Asia, Tatra and S-Finland from Europe, and Great Basin from North America. By providing a comprehensive set of criteria to evaluate tree-ring chronologies we hope to improve the development of large-scale temperature reconstructions spanning the past millennium. All reconstructions and their corresponding scores are provided at www.blogs.uni-mainz.de/fb09climatology.

Keywords

Paleoclimate, Climate change, Proxy data, Dendrochronology, Dendroclimatology

1. Introduction

Tree-ring chronologies (TRCs) are an important source of information in large-scale temperature reconstructions (IPCC 2013, St. George 2014). The latter are used to estimate temperature variability at continental (Euro-Med 2k consortium 2016, Pages 2k consortium 2013, Trouet et al. 2013), hemispheric (Christiansen and Ljungqvist 2012, D'Arrigo et al. 2006, Esper et al. 2002a, Ljungqvist 2010, Ljungqvist et al. 2012, 2016, Mann et al. 2008, Schneider et al. 2015, Shi et al. 2013, Stoffel et al. 2015, Wilson et al. 2016, Xing et al. 2016) and global scales (Mann and Jones 2003, Neukom et al. 2014) over the past 1000 years, enabling comparisons between climate variations during pre-industrial and industrial periods. The importance of TRCs in these reconstructions arises from the precise annual dating inherent to this proxy (Douglass 1941) and a well-defined mechanistic understanding of the influence of temperature on tree growth (Fritts 1976). The relative significance of tree-ring chronologies, compared to other proxies in large-scale reconstructions, increases back in time, as the number of annually resolved proxies rapidly declines towards the early centuries of the past millennium (Esper et al. 2004).

1.1 Basic tree-ring chronology characteristics

TRCs are typically composed of tree-ring width (TRW) or maximum latewood density (MXD) measurement series from many trees (Fritts 1976). A TRC might extend over the entire past millennium if one or more individual trees are 1000 years or more in age. Such longevity, however, is restricted to only a few known locations (OldList at: www.rmtrr.org/oldlist.htm). Most millennial-length TRCs are therefore produced by combining samples from living trees with older material from archeological and historical structures (hereafter: historical samples), dead wood on the ground (remnant samples), or wood preserved under ground and in lakes (sub-fossil samples). The successful combination of living trees with historical/remnant/sub-fossil material improves when the provenance of all samples is ecologically consistent. If not, older sections of a millennial-length chronology can have different growth rates and climate signals than those sections dominated by samples from living trees (Boswijk et al. 2014, Linderholm et al. 2014, Tegel et al. 2010). For example, remnant samples from a sub-alpine site in the Alps are ideally combined with samples from living trees growing on the same slope, at the same elevation and aspect (Neuwirth et al. 2004); sub-fossil trees from a shallow lake in Fennoscandia are ideally combined with trees growing around the lake, as opposed to drier inland locations (Düthorn et al. 2013, 2015).

Combining living trees with historical/remnant/sub-fossil samples is not always straightforward. Habitat homogeneity in a TRC derived from living trees and in-situ remnant or sub-fossil wood from the same location may be high, but their combination with historical material can be more complicated. If, for example, the historical samples were obtained from an old building in a mountain valley, it often remains unclear in which position in the surrounding forests the harvested tree grew

(Büntgen et al. 2006b). It is not uncommon for historical structures, particularly in alpine environments, to contain recycled material of unknown origin as a consequence of repairs and additions (Bellwald 2000, Kalbermatten and Kalbermatten 1997). Without detailed construction histories the researcher's ability to trace the origin of samples is limited (Büntgen et al. 2005, Wilson et al. 2004). The situation is further complicated if the samples in a TRC are from multiple locations spread over a large region, and if this region extends over several hundreds of kilometers. These problems, affecting the *Homogeneity* of a tree-ring dataset, are seemingly reduced in TRCs from only living trees sampled at a single site.

Another important characteristic of millennial-length TRCs includes the number and temporal distribution of TRW (or MXD) measurement series averaged in the mean chronology. Varying sample replication is often reported when describing a new TRC, but is usually disregarded in large-scale temperature reconstructions. Typically, the number of measurement series included in a TRC declines back in time and might change from more than 100 living-tree samples in the 20th century to only a handful of samples (perhaps from a single historical structure) at the beginning of the last millennium. Acknowledging the effects of changing sample size by calculating temporally varying uncertainty estimates is not usually considered outside the tree-ring community (IPCC 2013). However, this characteristic is important as the relevance of an individual TRC in large-scale proxy networks is commonly based on the strength of instrumental calibration of only the well-replicated 20th century data, thereby overlooking any pre-instrumental replication changes.

Similarly, the coherence among the TRW (MXD) series combined in a TRC, and temporal change thereof, is not considered in the non-dendrochronological literature (Frank et al. 2007). The inter-series correlation among sample measurements is an important characteristic of a mean chronology and is commonly computed to evaluate temporal changes of the chronology's signal strength (Fritts 1976). The inter-series correlation is rarely stable and can change at *(i)* the transition from living trees to series from historical/remnant/sub-fossil material, or *(ii)* from a cluster of measurement series of a certain building to another building, or *(iii)* by the proportion of juvenile, mature, and adult growth rings (Cook and Kairiukstis 1990). Gradual trends in the inter-series correlation, as well as step changes, are common in long TRCs and bare important information on the reliability of dendroclimatic reconstructions during pre-instrumental periods. Measures that assess the affect of changing sample size and inter-series correlation include the *Expressed Population Signal* and *Subsample Signal Strength* (Wigley et al. 1984). However, these metrics are not widely recognized beyond the tree-ring community and their combination with other uncertainties, e.g. from the unexplained variance of the calibration model or the choice of the detrending model, remains challenging (Esper et al. 2007).

Another important TRC characteristic is the degree to which a chronology retains the full spectrum of pre-instrumental temperature variance, which is affected by the method used for chronology development and the age-structure of the underlying data (Cook et al. 1995). Recent assessments of large datasets showed that instrumental meteorological measurements and tree-ring timeseries contain different frequency spectra (Ault et al. 2014, Bunde et al. 2013, Büntgen et al. 2015, Franke et al. 2013, Zhang et al. 2015), and that TRCs are limited in capturing millennial scale temperature trends (Esper et al. 2012b). To minimize the loss of long-term information, dendrochronologists apply detrending techniques that are specifically designed to preserve low frequency variance. The preferred approach is the Regional Curve Standardization (RCS) method, introduced to dendroclimatology by Briffa et al. (1992). However, RCS demands a large number of TRW (MXD) measurement series and requires the underlying data to represent a combination of short segments (trees) distributed more or less evenly throughout the entire chronology (Esper et al. 2003a). For example, if a TRC is composed of only very old living trees, the chronology's biological age will steadily increase towards the present. This causes the biologically younger rings to be concentrated at the beginning of the past millennium and the older rings in the modern period. This age structure limits the comparison of tree-rings of the same age over time, which is the backbone of RCS and related tree-ring detrending techniques (Melvin and Briffa 2008).

1.2 Objectives and structure

These basic characteristics of *Data Homogeneity*, *Sample Replication*, *Growth Coherence*, and *Chronology Development* are well known to dendroclimatologists. However, they are not usually recognized in the multi-proxy paleoclimate community and rarely, if ever, considered in large-scale temperature reconstructions derived from these data (IPCC 2013). The most widely considered criterion for data screening and TRC selection is the calibration against instrumental climate data used to compose the predictor networks for large-scale temperature reconstruction (Mann et al. 2008, Neukom et al. 2014, Stoffel et al. 2015, Xing et al. 2016). While we acknowledge the importance of calibration statistics, as well as the methods used to transfer proxy records into estimates of climate variability (Bürger et al. 2006, Christiansen 2011, Christiansen et al. 2009, Esper et al. 2005, Juckes et al. 2007, Lee et al. 2008, Smerdon et al. 2011, 2015, von Storch et al. 2004, Zorita et al. 2003), we believe additional consideration of the aforementioned TRC characteristics will improve the development of large-scale reconstructions.

In this article, we first describe *Data Homogeneity*, *Sample Replication*, *Growth Coherence*, *Chronology Development* and *Climate Signal*, and detail how these characteristics are combined in an ordinal scoring scheme. We apply this scheme to 39 tree-ring based temperature reconstructions reaching back to AD 1000 and use the results to rank the timeseries. This is done for each of the five characteristics separately and by combining their results in a final scoring scheme considering all characteristics. Potential implications of this ranking are discussed towards the end of the article and a list of recommendation that might help to improve the development of large-scale reconstructions is provided. The main objective of this paper is to promote basic dendrochronological characteristics on *Data Homogeneity*, *Sample Replication*, *Growth Coherence*, and *Chronology Development*, and to encourage their consideration when selecting records for future research, in addition to the classical calibration against instrumental climate data.

2. Data and methods

2.1 Temperature reconstructions

A survey of tree-ring based temperature reconstructions, with a minimum replication of three TRW (or MXD) measurement series reaching back to AD 1000, returned 39 records (Table 1; „reconstructions“ are derived from “TRCs” typically by applying a linear transfer function or simple scaling; Esper et al. 2005). Fourteen records are from Asia, 13 from Europe, 8 from North America, and 4 from the Southern Hemisphere (SH). The reconstructions are not evenly distributed over the hemispheres, but are clustered in Fennoscandia, the European Alps, northern Siberia, high Asia, the Rocky Mountains, southwestern US, southern South America, Tasmania and New Zealand. The TRCs are located in regions characterized by different summer warming trends over the past 100 years (see the colored areas in Fig. 1). Compared to the Northern Hemisphere (NH), the SH is clearly underrepresented with only four records.

There are precedents of long TRCs with an inferred climate signal (e.g. LaMarche 1973, 1974), but the first instrumentally calibrated, millennial-length record is the summer (previous-year December to current-year February; pDec-Feb) temperature reconstruction from very old (living) *Fitzroya cupressoides* growing in the Rio Alerce valley in southern Argentina (Villalba 1990; Table 1). Other records developed at that time were later updated by including new measurement series and/or reprocessed using new methods. A good example is the Torneträsk MXD chronology that was originally developed in the 1980s (Schweingruber et al. 1988), calibrated and reprocessed in the early 1990s (Briffa et al. 1990, 1992), updated in the early 2000s (Grudd 2008), and recently again updated and reprocessed (Melvin et al. 2013). In those instances where there are multiple versions of a reconstruction, we cite the most recently published account as it contains references to all previous work.

The millennial-length temperature reconstructions are derived from various conifer species representing nine genera, with *Pinus* (n=14 records) and *Larix* (n=12) being most common. Seven reconstructions, including the early *Fitzroya cupressoides* record from Argentina (Villalba 1990), are produced from only living trees, whereas the majority of chronologies (n=32) are composed of tree-ring series from living trees combined with series from historical samples (e.g. Lötschental TRC from Switzerland; Büntgen et al. 2006a), remnant samples (e.g. Polar Ural TRC from Russia; Briffa et al. 2013), and sub-fossil samples (e.g. Oroko Swamp TRC in New Zealand, Cook et al. 2002). Some of these chronologies are composed of samples collected in well-constrained, ecologically homogeneous (Schweingruber 1996) sites (e.g. Dzehlo in Russia, Myglan et al. 2012b), whereas others combine data from different sites (e.g. Yamal in Russia, Briffa et al. 2013), and even from several valleys within a larger region (e.g. Karakorum in Pakistan; Esper et al. 2002b).

All TRCs included in this survey have either been calibrated against regional instrumental climate data and transferred into temperature units, or interpreted by the original authors as a temperature proxy. Interestingly, the different methods used to transfer TRW and MXD data into temperature units (Briffa et al. 1983, Cook et al. 1994, Esper et al. 2005) produces vastly different reconstructed temperature ranges, varying by only a few tenths to several degrees Celsius over the past millennium (thin black curves in Figure 2; see Esper et al. 2012a for a regional example). Also, the season of maximum response to temperature (e.g. June-August, May-September, etc.) and the reconstructed climate target (e.g. mean, maximum, and minimum temperature) differ among the records (last column in Table 1).

Surprisingly, despite differences in *(i)* location and regional 20th century temperature trends (Fig. 1), *(ii)* species composition and sample sources (historical/remnant/sub-fossil), *(iii)* seasonality of the temperature signal, and *(iv)* transfer technique and reconstructed variance, the simple arithmetic mean of each "continent" (acknowledging that the records do not spatially represent NH continents) coheres astonishingly well over the past 1000 years (Fig. 2e). Correlations range from r = 0.42 between Asia and North America to r = 0.48 between Europe and Asia, and increase at decadal resolution to 0.66 (Asia/N-America) and 0.82 (Europe/Asia). This large-scale coherence indicates that some common external forcing affects this dendrochronological network (Fernández-Donado et al. 2013, Pages 2K PMIP3 group 2015) and confirms the paleoclimatic significance of tree-ring data over the past millennium.

2.2 TRC characteristics and metrics

In this section, we describe the five basic TRC characteristics *Data Homogeneity* (2.2.1), *Sample Replication* (2.2.2), *Growth Coherence* (2.2.3), *Chronology Development* (2.2.4), and *Climate Signal* (2.2.5), commonly used by dendrochronologists to evaluate a chronology for climate reconstruction, and explain how statistical measures of these characteristics are used in an ordinal scoring scheme that is understandable to non-specialists. In those instances where the raw TRW and MXD data are publically available or contributed by the authors (raw data at: www.blogs.uni-mainz.de/fb09climatology), we have re-calculated the metrics of interest. Where the original cross-dated measurements are not available (see last column in Table 7), we have estimated chronology scores based on information provided in the original articles. Such estimates are highlighted in red in the tables that follow. The calibration scores, resulting from the TRC's correlation against temperature data (2.2.5 *Climate signal*), are taken from the original articles. In the instances where no measure of calibration is detailed in the original article, we used nearby gridded data to provide an estimate of climate calibration.

For each characteristic (2.2.1 to 2.2.5) we used an ordinal scoring scheme to rank the reconstructions. To aid reconstruction comparison, results of the TRC scores are stratified into four classes: class-A (highlighted in green in Tables 2-7), class-B (light green), class-C (light blue), and class-D (blue). Except for the first characteristic (2.2.1 *Data Homogeneity*), we highlight the ten top-ranked TRCs in green (ranks 1-10), the TRCs ranking 11-20 in light green, the TRCs ranking 21-30 in light blue, and the TRCs ranking 31-39 in blue. This hierarchal color scale is expanded in the *Data Homogeneity* category (5 green, 9 light green, 16 light blue, 9 blue) to account for the larger number of intermediate TRCs. For all reconstructions their individual ranks for each characteristic (2.2.1 to 2.2.5) are finally summed into an overall score.

2.2.1 Data homogeneity

Of the five characteristics introduced here, *Data Homogeneity* is the most descriptive as it is based on a combination of qualitative traits rather than quantitative measures. *Homogeneity* integrates information on the *(i)* source of wood samples, *(ii)* type of chronology, *(iii)* number of species, *(iv)* temporal clustering, and *(v)* a remark (results shown in Table 2). "Source" includes information on the origin of wood samples and the number of sampling sites. We use "Sub-fossil" for samples from lakes, bogs, etc., "Remnant" for dead wood on the ground, and "Historic" for samples from old buildings and archaeological structures. The *Homogeneity* score also considers whether the samples originate from one, several, or multiple sites, as far as this information could be obtained from the original publication or via personal communication with the authors. "Chronology type" differentiates between "C"; records composed of living plus relict (sub-fossil/remnant/historical) material, and "L"; records composed of samples from only living trees. The "Number of Species" in a TRC is typically one, but occasionally may be two. "Temporal clustering" refers to cases where the contribution of data from distinct homogeneous sites dominates specific periods of the past 1000 years (a condition that might require the application of multiple RCS runs, Melvin et al. 2013). Finally, we included a "Remark" section summarizing specific features that are relevant to the *Homogeneity* score in support of the reconstruction's ranking.

2.2.2 Sample replication

The temporal distribution of TRW (or MXD) measurement series in the reconstructions differs dramatically over the past millennium (Fig. 3). These changes are considered in the second metric by combining information on *(i)* mean replication, *(ii)* maximum replication, *(iii)* minimum replication, and *(iv)* the 11th/20th century ratio of measurement numbers. "Mean replication" is the average number of measurement series (core samples or radii from disks) over the last millennium, considering all years from AD 1000 to the most recent year of a reconstruction. "Maximum replication" and "Minimum replication" refer to the maximum and minimum numbers of measurement series, which are typically reached in the modern and the early periods of a reconstruction, respectively (see the black curves in Fig. 3). The "11th/20th century ratio" acknowledges this exemplar replication curve shape, as well as its significance in the reconstruction: all TRCs are calibrated over the well-replicated 20th century, but the reconstruction period extends back to the, often weakly replicated, 11th century. The metric equals the mean 11th century replication divided by the mean 20th century replication, multiplied by 100. To produce the final *Replication* score, the first three values are summed (*i*+*ii*+*iii*) and the resulting sum multiplied by (*iv*). If the reconstruction is produced using MXD data the *Replication* score ((*i*+*ii*+*iii*)**iv)* is multiplied by 2 to account for MXD's increased signal strength and higher production costs. Note that these choices, as well as those described below for the other TRC characteristics, are not statistically validated but made with the intention of combining descriptive measures commonly used in dendrochronology into an ordinal scoring system that can be used to compare and rank reconstructions.

2.2.3 Growth coherence

Another important characteristic influencing the temporally changing skill of tree-ring based climate reconstructions is the correlation between the TRW (MXD) measurement series (Frank et al. 2007, Osborn et al. 1997, Wigley et al. 1984). For those reconstructions where the raw data are available, we calculated the inter-series correlation (abbreviated "Rbar" in the dendrochronological literature; Cook and Kairiukstis 1990) for 100-year segments, sliding in 10-year steps along the chronology (Fig. 4). The resulting timeseries reveal substantial differences among the TRCs (the black curves in Fig. 4), as well as a minor tendency towards reduced values back in time, particularly in some records from Europe and Asia. These characteristics are considered here in the *Growth Coherence* score by summing the *(i)* average inter-series correlation over the past millennium (mean Rbar), *(ii)* maximum inter-series correlation in a single 100-year period (max. Rbar), and *(iii)* minimum inter-series correlation in a single 100-year segment (min. Rbar). The sum (*i*+*ii*+*iii*) is multiplied by *(iv)* the 11th/20th century Rbar ratio (in %).

2.2.4 Chronology development

A key component in the process of building a TRC is the detrending method used to remove tree-age related growth trends from the raw measurement series (Bräker 1981, Cook and Kairiukstis 1990, Cook et al. 1995). As mentioned above, RCS (Esper et al. 2003a) is currently accepted as the preferred method to preserve low frequency variance in TRCs. We acknowledge this view by *(i)* assigning

TRCs produced using RCS a “1”, and TRCs produced using individual-series detrending methods (e.g. ratios from negative exponential curves or smoothing splines) a “2” (Cook and Peters 1997). However, RCS only works well if the underlying measurement series are derived from a composite of (many) living and relict trees, ideally including young and old tree-rings evenly distributed throughout the past millennium (Esper et al. 2014). TRCs composed this way are characterized by age curves that are nearly horizontal over the past 1000 years (Fig. 5). In practice this is rarely the case. The age curves of some TRCs composed of very old living trees in, for example, North America and Asia are particularly steep (increasing-age towards present). In contrast, in Europe, where the majority of reconstructions are derived from composite chronologies of historical and living-tree samples, the mean age curves are relatively flat (the blue curve in Fig. 5e).

We score these attributes by considering *(ii)* the maximum difference between the highest and lowest value in the age curve over the past millennium, and *(iii)* the slope of a linear regression fit to the age curve. We further consider *(iv)* the maximum retained low frequency information, ranging from multi-centennial = 1, to centennial = 2, to decadal = 3. For the final *Chronology Development* score we multiply *(i)* the method score (1 for RCS, 2 for individual-series detrending), with *(ii)* the (square root of the) max.-min. age difference, *(iii)* the (absolute) slope of the linear regression (times 100), and *(iv)* the maximum retained low frequency score (1 to 3, for multi-centennial, centennial, and decadal).

2.2.5 Climate signal

This final score considers some of the classic metrics used in paleoclimatic research, such as the correlation against monthly instrumental temperature data, averaged over the season of maximum response (see the last column in Table 1). However, as the period of overlap between instrumental and proxy data varies considerably among the reconstructions – largely due to the lengths of observational data available to researchers – we score *Climate Signal* by *(i)* the square root of the number of years of overlap between the TRC and instrumental record, multiplied by the residual between, *(ii)* the correlation against climate data and *(iii)* a split calibration/verification difference. The latter metric is a standard criterion in dendroclimatology used to benchmark the temporal robustness of the relationship between proxy and instrumental data (Cook and Kairiukstis 1990). However, the split calibration/verification differences are not always reported. In those instances we estimated the split calibration/verification difference based on our calculations using gridded temperature data. Finally, we include an additional variable *(iv)* to account for a calibration period that was intentionally shortened to avoid potential divergence issues (for details see Büntgen et al. 2008, D’Arrigo et al. 2008, Esper and Frank 2009, Esper et al. 2010, Wilson et al. 2007). If such problems are reported in the original article, and the calibration period was truncated, we used 0.5 as a multiplier (1 if no such problem was detected). The final *Climate Signal* score was derived by: square root *i* * *(ii - iii)* * *iv*.

3. Results and discussion

3.1 Overall TRC ranking

Our assessment of 39 millennial-long TRCs' *Data Homogeneity*, *Sample Replication*, *Growth Coherence*, *Chronology Development*, and *Climate Signal* is presented in Tables 2-6. The final ranking (Table 7), derived from the sum of all scores, reveals that no reconstruction consistently dominates in the top group (class-A, dark green dots in the tables) in all five categories. Four records (N-Scan, E-Canada, Finland, Dzhelo) score high (class-A or class-B) in four out of the five categories, and one record (Yamal) scores high in three. However, each of these, overall best-ranked reconstructions, scores less well (class-C: light blue dot) on at least one criterion, mostly *Data Homogeneity* (four records).

There are ten records (W-Himalaya, Tatra, Karakorum, Great Basin, S-Finland, Tien Shan, Jämtland, Wulan, Gulf of Alaska, French Alps) with weak scores (class-C or class-D) in four categories. W-Himalaya is the only reconstruction scoring in class-D in four: *Data Homogeneity*, *Sample replication*, *Growth Coherence*, and *Chronology Development*. This low scoring, combined with a class-B rank in *Climate Signal*, places the W-Himalaya reconstruction at the bottom of table 7. However, the raw data are not available for this reconstruction (see the last column) and several scores had to be estimated (highlighted in red in tables 3-6). The same is true for several other reconstructions, and it seems advisable to emphasize the coarse categorization into four classes (A to D) rather than the precise ordering in our tables.

3.2 Detailed TRC rankings

Four reconstructions (Dzehlo, Tasmania, Rio Alerce, Qamdo, Mongolia) rank at the top in *Data Homogeneity* (Table 2). The data used in these TRCs include samples from living trees, as well as remnant and sub-fossil material from a single site or valley (with one exception; Rio Alerce comprised of only living trees). These top-ranked records are followed by a group of nine reconstructions that were sampled from slightly less homogeneous conditions, e.g. including data at moist and dry micro-sites, from different elevations, and measured using different techniques. Despite their less than ideal *Data Homogeneity* score, these reconstructions are still more homogenous compared to a number of TRCs (n=16) that integrate data from multiple sites in different valleys, regions, and/or elevations (light blue dots in Table 2). Such large differences in habitat can introduce substantial growth rate variations that are difficult to differentiate from long-term temperature variations. In addition, the climate signal might change between samples from different elevations and micro-sites. These potential biases are likely most severe in the nine TRCs ranking last (dark blue dots in Table 2). The two end members (Karakorum and W-Himalaya) are produced using living trees only, sampled from multiple sites, located in different valleys and at different elevations, with distances up to 100 km between sites, and including two tree species (Karakorum). Clearly these TRCs contain a less homogeneous sample composition compared to the top-ranked records that include samples from one, well-constrained site.

The reconstructions scoring well in *Data Homogeneity* are not necessarily top-ranked in *Sample Replication* (Table 3). To appear in the top group in *Sample Replication*, it is necessary not only to include a large number of TRW or MXD measurement series, but have these samples evenly distributed throughout the past millennium. Bumps from very high to very low replications in certain periods, as well as large differences between 20th and 11th century replications, result in a lower score. Among the records performing well in *Replication* are two TRCs from Central Asia (Mongun and Dzhelo) and one from New Zealand (Oroko Swamp). These records score particularly well in the 11th/20th century ratio, reaching values >100%. Other reconstructions, such as the Alps (larch) and Swiss/Austrian Alps TRCs include many samples (530 and 253 respectively over the past millennium), but contain a dramatic replication decline from the 20th to the 11th centuries (down to 2% and 25%, respectively), limiting these reconstructions' skill in the early period of the past millennium. The TRCs scoring weakest in *Replication* (Tatra, Boreal Plateau, Rio Alerce, Tien Shan) are characterized by low minimum replications (n≤5 series) and small 11th/20th century ratios (≤15%). These records might perform well when calibrated against 20th century instrumental temperature data, but there is considerable risk that this 20th century skill does not persist over the past millennium simply because the number of samples changes drastically back in time.

Since more than the sheer number of measurement series is important, we also considered the reconstructions' inter-series correlations (Table 4). The three TRCs scoring best in this category (Indigirka, Yamal, Taimyr) are all located in northern Siberia, where growth variations among trees are synchronized by harsh climatic conditions during a rather short growing season. These top-ranked records are characterized by inter-series correlations that do not fall below Rbar=0.20 at any time over the past millennium (minimum correlation in Table 4) and reach values >100% in their 11th/20th century ratio. Other mid-ranked TRCs, such as Polar Ural (class-B) and Jämtland (class-C), display either very low minimum Rbar values (-0.20 in Polar Ural) or substantially decreasing Rbar values from the 20th century back to the 11th century (42% in Jämtland). Another interesting example of a class-C TRC is Oroko Swamp, which is characterized by only minor Rbar changes back in time (92%), but an overall low mean inter-series correlation (Rbar=0.18). Finally, the TRCs scoring weakest (Tatra, S-Finland, Central Alps) are characterized by severe correlation declines, down to ≤20% back in the 11th century, and either a low mean Rbar (0.20 estimated for S-Finland and Central Alps) or negative minimum Rbar values (-0.07 in Tatra). In these cases it seems advisable to anticipate substantial changes in the chronologies' signal strength over the past millennium as the coherence among their constituent measurement series is extremely variable. If the inter-series correlation drops significantly, reductions in TRC variance, and a tendency towards the long-term mean are to be expected.

The three top-ranked reconstructions in the *Chronology Development* category are all from Northern Europe (N-Scan, Finland, Torneträsk (TRW)), followed by records from the Alps (Lötschental) and Canada (E-Canada) (Table 5). These reconstructions, as well as the other class-A and class-B TRCs (green and light green in Table 5, total n=20), are all composed of a mixture of living trees and historical/remnant/sub-fossil samples, facilitating the application of RCS for optimal conservation of low frequency variance (Autin et al. 2015, Briffa et al. 1992, Esper et al. 2003a). The top-scoring Northern European records are, however, additionally characterized by small age ranges (<110 years) and only minor (positive and negative) linear trends in the mean age curves. The top-ranked N-Scan record is reported to contain millennial scale temperature variance (Esper et al. 2012b), a feature also seen in the Taimyr reconstruction from Northern Siberia. The subsequent mid-ranked TRCs are characterized by age ranges from ~150-300 years, as well as linear trend angles ranging from ~5-30 degrees. Some class-C records were standardized using individual detrending methods, including the Swiss/Austrian Alps, Lauenen and Mongolia reconstructions, an approach more commonly found in the TRCs towards the bottom of Table 5. The application of individual detrending methods has been shown to systematically limit the low frequency variance retained in TRCs (Cook et al. 1995). This limitation is reflected in the maximum frequency metric included here, indicating that six reconstructions (Rio Alerce, Wulan, Gulf of Alaska, Mongun, S-Chile, Lauenen) maximally retain decadal scale temperature variance. These records, as well as some of the individually detrended TRCs, should not be used with the objective of reconstructing the full spectrum of temperature variance over the past millennium (e.g. Mann et al. 2008).

By comparison to *Data Homogeneity*, *Sample Replication*, *Growth Coherence*, and *Chronology Development*, measures of climate signal strength are widely recognized in the paleoclimatic community. However, a good correlation between tree-ring proxy and instrumental temperature data alone is a fairly incomplete description of reconstruction skill. For example, if a TRC includes many more samples during the 20th century (*Sample Replication* metric), or the samples originate from different valleys (*Data Homogeneity*), or the mean age curve declines severely back in time (*Chronology Development*), the 20th century calibration statistics provide little information about the signal strength over past centuries. That being said, we here assess climate signal strength based on the length of the calibration period, the correlation strength with instrumental data, the calibration/verification difference and any, seemingly arbitrary, truncation of the calibration period.

The reconstructions scoring best for *Climate Signal* are all from regions where instrumental records of 100 years and longer are available for calibration (Table 6). The three top-ranked records (Torneträsk

(MXD), N-Scan, Alps (larch)) all correlate at r≥0.70 against instrumental temperature data, with only minor differences (<0.10) between calibration and verification periods. Other reconstructions, with calibration period correlations r≥0.70, albeit over shorter periods (53 years in Qamdo, 57 years in Taimyr), contain larger calibration/verification differences (0.18 in Taimyr) and appear in class-B. These reconstructions certainly meet the criteria for a successful TRC calibration, but they may contain a marginally verifiable climate signal. This is either because the calibration/verification differences are large (e.g. 0.63 in Qilian), the calibration period was truncated due to some inconsistency (e.g. Tatra, see the fourth column in Table 6), or the overall correlation is low (e.g. 0.17 in Upper Wright Lakes). However, a weak calibration result does not necessarily mean that a TRC contains no climate signal, but might indicate that the instrumental station record is too short (Esper et al. 2010), of poor quality (Böhm et al. 2001, 2010, Parker et al. 1994), or too remote (Cook et al. 2013).

Perhaps a good example, highlighting the importance of using several categories to evaluate a TRC, is the case of the Alps (larch) record. The Alps TRC correlates well (r=0.70) over 140 years of regional instrumental temperatures, and thus ranks #3 in the *Climate Signal* metric (calibration/verification difference is 0.07, calibration period not truncated). However, these calibration statistics were obtained over the period 1864-2003 during which the TRC's mean replication is 1379 series. Concurrently, the average number of TRW series in the 11th century reaches only 22, which produces an 11th/20th century ratio of 2% (see Table 3). Though certainly an extreme example, it nicely demonstrates how a large-scale reconstruction produced focusing on 20th century climate signals, can result in an overestimation of statistical skill over the past millennium.

3.3 Ranking implications

Over recent decades a number of statistically valid methods have been developed to describe a TRC's signal strength. Examples include the *Expressed Population Signal* (Wigley et al. 1984), bootstrap confidence intervals (Briffa et al. 1992), ensemble calibration technique (Frank et al. 2010), and reduced sample calibration trials (Esper et al. 2012b). All of these dendro-specific statistics help estimate the temporally varying skill of tree-ring based climate reconstructions, but the methods are largely inapplicable to other proxy archives, and are not used in large-scale, multi-proxy reconstructions (Pages 2k consortium 2013).

By providing an assessment and ranking of TRCs, we attempt to bridge the gap between the tree-ring, modeling, and multi-proxy communities. While some of the scores and metrics used here have not been rigorously validated, we believe that the development of an intuitive ranking system that can be universally applied to all TRCs will foster the judicious use of tree-ring data in large-scale reconstructions. For example, if NH temperature variability during medieval times is of interest, it is not meaningful to include TRCs with only a few samples during the 11th century, i.e. researchers might want to avoid reconstructions with low *Sample Replication* scores (Table 3). Similarly, if the full spectrum of past temperature variability is of interest, one might want to include only those TRCs retaining centennial to millennial scale variance, i.e. exclude records with low *Chronology Development* scores (Table 5).

These arguments lead to a list of recommendations:

R1 Avoid integrating tree-ring chronologies that emphasize decadal scale variance when intending to reconstruct centennial to millennial scale temperature variance.

R2 Avoid overrating tree-ring chronologies that average many measurement series in the 20th century, but only few series at the beginning of the last millennium.

R3 Pay attention to the tree-ring chronology sample composition and potentially changing data sources over the past millennium (different sites, buildings, valleys).

R4 Consider replication and inter-series correlation changes when interpreting tree-ring based climate reconstructions.

R5 Differentiate between composite tree-ring chronologies that integrate data from varying sources (living/remnant/historical/sub-fossil) and tree-ring chronologies that integrate data from only old living trees, and acknowledge potential biases due to changing tree ages over the past millennium.

R6 Do not only focus on the calibration statistics from comparisons with instrumental climate data, as this perspective can give the false impression that reconstruction skill persists throughout the past millennium.

We acknowledge that some of the metrics presented here contain partly redundant information, e.g. lower replication or reduced Rbar values typically result in weaker correlations with instrumental climate. There are also other TRC characteristics that could be used to assess tree-ring based temperature reconstructions, though these appeared difficult to quantify with simple measures. Examples include the TRC serial correlation (Meko 1981) and climate signal after trend removal (von Storch et al. 2004). For instance, an assessment of serial correlation in both tree-ring and instrumental temperature data might reveal a larger lag-1 autocorrelation in a TRC (likely due to biological memory effects; Esper et al. 2015), suggesting a coherence deficiency and reduced skill of a long-term climate reconstruction. Similarly, an assessment of the climate signal after removing low frequency variance (e.g. increasing 20th century temperature trend), from the instrumental and proxy data, increases the degrees of freedom of the calibration statistics and supports the estimation of signal strength in the high frequency domain. However, correctly evaluating these properties in a large network of millennial-length TRCs, including several records for which the underlying measurement data are not available, is not feasible.

Our review clearly indicates that solely focusing on the calibration statistics overlooks a number of additional, important characteristics inherent to tree-ring based climate reconstructions. When evaluating large TRC networks it is important to keep in mind that the 20th century instrumental data *(i)* contain gaps, breakpoints, and biases (Hinkel et al. 2003, Landsberg 1981, Oke 2007), *(ii)* are of substantially varying length depending on the study region (e.g. in Europe versus central Asia; Cook et al. 2013), and *(iii)* are recorded at greatly differing distances from the tree-ring sampling sites. The suitability of a station record is additionally influenced by the topography (flat or mountainous), the elevation difference between tree and station sites, and regional synoptic weather patterns. The use of gridded climate data does not necessarily overcome these shortcomings as they rely on the same (Jones et al. 1999) or even fewer (Krusic et al. 2015) station data.

5. Conclusions

Thirty-nine millennial-length temperature reconstructions are ranked based on a rating scheme that considers basic TRC characteristics commonly considered by dendrochronologists. The TRC characteristics were grouped into five composite scores: *Data Homogeneity*, *Sample Replication*, *Growth Coherence*, *Chronology Development*, and *Climate Signal*. It is argued that consideration of these characteristics, beyond the tree-ring community, will improve the development of large-scale temperature reconstructions that utilize tree-ring data from different regions and continents. Similarly, the rankings produced for each score supports this objective, as they facilitate the selection process of TRCs when addressing paleoclimatic objectives. For example, researchers might not want to include TRCs resting on only a few trees during the 11th century, in a study addressing the magnitude and spatial extent of warmth during medieval times. This, and other recommendations are expressed towards the end of this review paper.

A systematic comparison of the TRC characteristics permitted ranking of the 39 millennial-length temperature reconstructions into four groups (class-A to class-D) for each of the five categories. No reconstruction scores top in all five categories, but each record has its particular strengths and weaknesses. Nevertheless, there are some reconstructions that overall perform better than others. These include N-Scan and Finland from Europe; E-Canada from North America; Yamal and Dzehlo from Asia. Reconstructions performing less well include W-Himalaya and Karakorum from Asia; Tatra and S-Finland from Europe; and Great Basin from North America. The rankings presented here can be used to select and exclude particular records for producing hemispheric scale reconstructions. The fact that some of the records appear more often towards the bottom of a ranking table does not mean they cannot be used for climate reconstruction purposes, but implies users of these data need to be aware of potential weaknesses that may inadvertently affect their experiment. This review of millennial-long TRCs will be updated as new reconstructions are produced. Updates will be published online at: www.blogs.uni-mainz.de/fb09climatology

Acknowledgements

We thank all the tree-ring data producers for sharing their chronologies and measurement series. Supported by the German Science Foundation, Grant 161/9-1. JL acknowledges the German Science Foundation project "Attribution of forced and internal Chinese climate variability in the Common Era", and the National Natural Science Foundation of China (Grant 41325008).

References

Ault, T.R., Cole, J.E., Overpeck, J.T., Pederson, G.T., Meko, D.M., 2014. Assessing the risk of persistent drought using climate model simulations and paleoclimate data. J. Clim. 27, 7529–7549.

Autin, J., Gennaretti, F., Arseneault, D., Bégin, Y., 2015. Biases in RCS tree ring chronologies due to sampling heights of trees. Dendrochronologia 36, 13–22.

Bellwald, I., 2000. Der Rote Segensonntag 1900. Der Dorfbrand von Wiler. Ein Rückblick aus dem Jahre 2000. Gem. Wiler, Kippel.

Böhm, R., Auer, I., Brunetti, M., Maugeri, M., Nanni, T., Schöner, W., 2001. Regional temperature variability in the European Alps: 1760-1998 from homogenized instrumental time series. Int. J. Climatol. 21, 1779–1801.

Böhm, R., Jones, P.D., Hiebl, J., Frank, D., Brunetti, M., Maugeri, M., 2010. The early instrumental warm-bias: a solution for long central European temperature series 1760–2007. Clim. Change 101, 41–67.

Boswijk, G., Fowler, A.M., Palmer, J.G., Fenwick, P., Hogg, A., Lorrey, A., Wunder, J., 2014 The late Holocene kauri chronology: assessing the potential of a 4500-year record for palaeoclimate reconstruction. Quat. Sci. Rev. 90, 128–142.

Bräker, O.U., 1981. Der Alterstrend bei Jahrringdichten und Jahrringbreiten von Nadelhölzern und sein Ausgleich. Mitteil. Forstl. Bundesversuchsanst. Wien 142, 75–102.

Briffa, K.R., Jones, P.D., Wigley, T.M.L., Pilcher, J.R., Baillie, M.G.L., 1983. Climate reconstruction from tree rings: Part 1, basic methodology and preliminary results for England. J. Climatol. 3, 233–242.

Briffa, K.R., Bartholin, T.S., Eckstein, D., Jones, P.D., Karlén, W., Schweingruber F.H., Zetterberg, P., 1990. A 1,400-year tree-ring record of summer temperatures in Fennoscandia. Nature 346, 434–439.

Briffa, K.R., Jones, P.D., Bartholin, T.S., Eckstein, D., Schweingruber, F.H., Karlen, W., Zetterberg, P., Eronen, M., 1992. Fennoscandian summers from AD 500: temperature changes on short and long timescales. Clim. Dyn. 7, 111–119.

Briffa, K.R., Shishov, V.V., Melvin, T.M., Vaganov, E.A., Grudd, H., Hantemirov, R.M., Eronen, M., Naurzbaev, M.M., 2008. Trends in recent temperature and radial tree growth spanning 2000 years across northwest Eurasia. Philosoph. Trans. Royal Soc. B 363, 2269–2282.

Briffa, K.R., Melvin, T.M., Osborn, T.J., Hantemirov, R.M., Kirdyanov, A.V., Mazepa, V.S., Shiyatov, S.G., Esper, J., 2013. Reassessing the evidence for tree-growth and inferred temperature change during the Common Era in Yamalia, Northwest Siberia. Quat. Sci. Rev. 72, 83–107.

Büntgen, U., Esper, J., Frank, D.C., Nicolussi, K., Schmidhalter, M., 2005. A 1052-year tree-ring proxy for Alpine summer temperatures. Clim. Dyn. 25, 141–153.

Büntgen, U., Frank, D.C., Nievergelt, D., Esper, J., 2006a. Summer temperature variations in the European Alps, A.D. 755-2004. J. Clim. 19, 5606–5623.

Büntgen, U., Bellwald, I., Kalbermatten, H., Schmidhalter, M., Frank, D.C., Freund, H., Bellwald, W., Neuwirth, B., Nüsser, M., Esper, J., 2006b. 700 years of settlement and building history in the Lötschental/Switzerland. Erdkunde 60, 96–112.

Büntgen, U., Frank, D.C., Wilson, R.J.S., Carrer, M., Urbinati, C., Esper, J., 2008. Testing for tree-ring divergence in the European Alps. Glob. Change Biol. 14, 2243–2453.

Büntgen, U., Frank, D., Carrer, M., Urbinati, C., Esper, J., 2009. Improving Alpine summer temperature reconstructions by increasing sample size. Trace 7, 36–43.

Büntgen, U., Tegel, W., Nicolussi, K., McCormick, M., Frank, D., Trouet, V., Kaplan, J.O., Herzig, F., Heussner, K.U., Wanner, H., Luterbacher, J., Esper, J., 2011. 2500 years Of European climate variability and human susceptibility. Science 331, 578–582.

Büntgen, U., Neuschwander, T., Frank, D., Esper, J., 2012. Fading temperature sensitivity of Alpine tree growth at its Mediterranean margin and associated effects on large-scale climate reconstructions. Clim. Change 114, 651–666.

Büntgen, U., Kyncl, T., Ginzler, C., Jacks, D.S., Esper, J., Tegel, W., Heussner, K.U., Kyncl, J., 2013. Filling the Eastern European gap in millennium-length temperature reconstructions. Proc. Nat. Acad. Sci. 5, 1773–1778.

Büntgen, U., Trnka, M., Krusic, P.J., Kyncl, T., Kyncl, J., Luterbacher, J., Zorita, E., Ljungqvist, F.C., Auer, I., Konter, O., Schneider, L., Tegel, W., Štěpánek, P., Brönnimann, S., Hellmann, L., Nievergelt, D., Esper, J., 2015. Tree-ring amplification of the early nineteenth-century summer cooling in central Europe. J. Clim. 28, 5272–5288.

Bürger, G., Fast, I., Cubasch, U., 2006. Climate reconstruction by regression—32 variations on a theme. Tellus 58, 227–235.

Bunde, A., Büntgen, U., Ludescher, J., Luterbacher, J., von Storch, H., 2013. Is there memory in precipitation? Nat. Clim. Change 3, 174–175.

Christiansen, B., 2011. Reconstructing the NH mean temperature: Can underestimation of trends and variability be avoided? J. Clim. 24, 674–692.

Christiansen, B., Schmith, T., Thejll P., 2009. A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness. J. Clim. 22, 951–976.

Christiansen, B., Ljungqvist, F.C., 2012. The extra-tropical Northern Hemisphere temperature in the last two millennia: reconstructions of low-frequency variability. Clim. Past 8, 765–786.

Cook, E.R., Kairiukstis, L.A., 1990. Methods of Dendrochronology – Applications in the Environmental Science. Kluwer, Dordrecht.

Cook, E.R., Briffa, K.R., Jones, P.D., 1994. Spatial regression methods in dendroclimatology: a review and comparison of two techniques. Int. J. Climatol 14, 379–402.

Cook, E.R., Briffa, K.R., Meko, D.M., Graybill, D.A., Funkhouser, G., 1995. The 'segment-length curse' in long tree-ring chronology development for palaeoclimatic studies. Holocene 5, 229–237.

Cook, E.R., Peters, K., 1997. Calculating unbiased tree-ring indices for the study of climatic and environmental change. Holocene 7, 361–370.

Cook, E.R., Buckley, B.M., D'Arrigo, R.D., Peterson, M.J., 2000. Warm-season temperatures since 1600BC reconstructed from Tasmanian tree rings and their relationship to large-scale sea surface temperature anomalies. Clim. Dyn. 16, 79–91.

Cook E.R., Palmer, J.G., Cook, B.I., Hogg, A., D'Arrigo, R.D., 2002. A multi-millennial palaeoclimatic resource from *Lagarostrobos colensoi* tree-rings at Oroko Swamp, New Zealand. Glob. Plan. Change 33, 209–220.

Cook, E.R., Krusic, P.J., Anchukaitis, K.J., Buckley, B.M., Nakatsuka, T., Sano, M., Pages Asia2k Members, 2013. Tree-ring reconstructed summer temperature anomalies for temperate East Asia since 800 CE. Clim. Dyn. 41, 2957–2972.

D'Arrigo, R.D., Jacoby, G., Frank, D., Pederson, N., Cook, E.R., Buckley, B.M., Nachin, B., Mijiddorj, R., Dugarjav, C., 2001. 1738 years of Mongolian temperature variability inferred from a tree-ring width chronology of Siberian pine. Geophys. Res. Lett. 28, 543–546.

D'Arrigo, R., Wilson, R., Jacoby, G. 2006. On the long-term context for late 20th century warming. J. Geophys. Res. 111, D03103, doi: 10.1029/2005JD006352.

D'Arrigo, R.D., Wilson, R., Liepert, B., Cherubini, P., 2008. On the 'divergence problem' in northern forests: a review of the tree-ring evidence and possible causes. Global Planet. Change 60, 289–305.

Douglass, A.E., 1941. Crossdating in dendrochronology. J. Forestry 39, 825–832.

Düthorn, E., Holzkämper, S., Timonen, M., Esper, J., 2013. Influence of micro-site conditions on tree-ring climate signals and trends in Central and Northern Sweden. Trees 27, 1395–1404.

Düthorn, E., Schneider, L., Konter, O., Schön, P., Timonen, M., Esper, J., 2015. On the hidden significance of differing micro-sites in dendroclimatology. Silva Fennica 49, doi: org/10.14214/sf.1220.

Esper, J., Cook, E.R., Schweingruber, F.H., 2002a. Low-frequency signals in long tree-ring chronologies for reconstructing of past temperature variability. Science 295, 2250–2253.

Esper, J., Schweingruber, F.H., Winiger, M. 2002b. 1,300 years of climate history for Western Central Asia inferred from tree-rings. Holocene 12, 267–277.

Esper, J., Cook, E.R., Krusic, P.J., Peters, K., Schweingruber, F.H., 2003a. Tests of the RCS method for preserving low-frequency variability in long tree-ring chronologies. Tree-Ring Res. 59, 81–98.

Esper, J., Shiyatov, S.G., Mazepa, V.S., Wilson, R.J.S., Graybill, D.A., Funkhouser, G., 2003b. Temperature-sensitive Tien Shan tree-ring chronologies show multi-centennial growth trends. Clim. Dyn. 8, 699–706.

Esper, J., Frank, D.C., Wilson, R.J.S., 2004. Climate reconstructions: low frequency ambition and high frequency ratification. EOS 85, 113–130.

Esper, J., Frank, D.C., Wilson, R.J.S., Briffa, K.R., 2005. Effect of scaling and regression on reconstructed temperature amplitude for the past millennium. Geophys. Res. Lett. 32, doi: 10.1029/2004GL021236.

Esper, J., Frank, D.C., Büntgen, U., Verstege, A., Luterbacher, J., Xoplaki, E., 2007. Long-term drought severity variations in Morocco. Geophys. Res. Lett. 34, doi: 10.1029/2007GL030844.

Esper, J., Frank, D., 2009. Divergence pitfalls in tree-ring research. Clim. Change 94, 261–266.

Esper, J., Frank, D., Büntgen, U., Verstege, A., Hantemirov, R.M., Kirdyanov, A.V., 2010. Trends and uncertainties in Siberian indicators of 20th century warming. Glob. Change Biol. 16, 386–398.

Esper, J., Büntgen, U., Timonen, M., Frank, D.C., 2012a. Variability and extremes of Northern Scandinavian summer temperatures over the past two millennia. Glob. Plan. Change 88-89, 1-9.

Esper, J., Frank, D.C., Timonen, M., Zorita, E., Wilson, R.J.S., Luterbacher, J., Holzkämper, S., Fischer, N., Wagner, S., Nievergelt, D., Verstege, A., Büntgen U., 2012b. Orbital forcing of tree-ring data. Nat. Clim. Change 2, 862–866.

Esper, J., Düthorn, E., Krusic, P., Timonen, M., Büntgen, U., 2014. Northern European summer temperature variations over the Common Era from integrated tree-ring density records. J. Quat. Sci. 29, 487–494.

Esper, J., Schneider, L., Smerdon, J., Schöne, B., Büntgen, U., 2015. Signals and memory in tree-ring width and density data. Dendrochronologia 35, 62–70.

Euro-Med 2k consortium, 2016. European summer temperatures since Roman times. Environ. Res. Lett., in press.

Fernández-Donado, L., González-Rouco, J.F., Raible, C.C., Ammann, C.M., Barriopedro, D., García-Bustamante, E., Jungclaus, J.H., Lorenz, S.J., Luterbacher, J., Phipps, S.J., Servonnat, J., Swingedouw, D., Tett, S.F.B., Wagner, S., Yiou, P., Zorita, E., 2013. Large-scale temperature response to external forcing in simulations and reconstructions of the last millennium. Clim. Past 9, 393–421.

Franke, J., Frank, D., Raible, C., Esper, J., Brönnimann, S., 2013. Spectral biases in tree-ring climate proxies. Nat. Clim. Change 3, 1–5.

Frank, D., Esper, J., Cook E.R., 2007. Adjustment for proxy number and coherence in a large-scale temperature reconstruction. Geophys. Res. Lett. 34, doi: 10.1029/2007GL030571.

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

Fritts, H.C., 1976. Tree Rings and Climate. Academic press, London.

Gennaretti, F., Arseneault, D., Nicault, A., Perreault, L., Bégin, Y., 2014. Volcano-induced regime shifts in millennial tree-ring chronologies from northeastern North America. Proceed. Nat. Acad. Sci. 111, 10077–10082.

Graumlich, L.J., 1993. A 1000-year record of temperature and precipitation in the Sierra Nevada. Quat. Res., 39, 249–255.

Grudd, H., 2008. Torneträsk tree-ring width and density AD 500-2005: a test of climatic sensitivity and a new 1500-year reconstruction of north Fennoscandian summers. Clim. Dyn. 31, 843–857.

Helama, S., Fauria, M.M., Mielikäinen, K., Timonen, M., Eronen, M., 2010. Sub-Milankovitch solar forcing of past climates: mid and late Holocene perspectives. GSA Bulletin; 122, 1981–1988.

Helama, S., Vartiainen, M., Holopainen, J., Mäkelä, H.M., Kolström, T., Meriläinen, J. 2014. A palaeotemperature record for the Finnish Lakeland based on microdensitometric variations in tree rings. Geochronometria 41, 265–277.

Hinkel, K.M., Nelson, F.E., Klene, A.E., Bell, J.H., 2003. The urban heat island in winter at Barrow, Alaska. Int. J. Climatol. 23, 1889–1905.

IPCC, 2013. 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.

Jones, P.D., New, M., Parker, D.E., Martin, S., Rigor, I.G., 1999. Surface air temperature and its variations over the last 150 years. Rev. Geophys. 37, 173–199.

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. Clim. Past 3, 591–609.

Kalbermatten, H., Kalbermatten L., 1997. Blatten. Was alte Menschen, alte Häuser und Schriften erzählen. Druckerei Bloch, Arlesheim.

Krusic, P.J., Cook, E.R, Dukpa, D., Putnam, A.E., Rupper, S., Schaefer, J., 2015. 638 years of summer temperature variability over the Bhutanese Himalaya. Geophys. Res. Lett. 42, doi: 10.1002/2015GL063566.

LaMarche, V.C. Jr., 1973. Holocene climatic variations inferred from treeline fluctuations in the White Mountains, California. Quat Res 3, 632–660.

LaMarche, V.C. Jr., 1974. Paleoclimatic inferences from long tree-ring records. Science 183, 1043–1048.

Landsberg, H.E., 1981. The urban climate. Academic press, London.

Lara, A., Villalba, R., 1993. A 3620-year temperature record from *Fitzroya cupressoides* tree rings in southern South America. Science 260, 1104–1106.

Lee, T.C.K., Zwiers, F.W., Tsao, M., 2008. Evaluation of proxy-based millennial reconstruction methods. Clim. Dyn. 31, 263–281.

Linderholm, H.W., Gunnarson, B.E., 2005. Summer temperature variability in central Scandinavia in the last 3600 years. Geogr. Ann. 87, 231–241.

Linderholm, H.W., Zhang, P., Gunnarson, B.E., Björklund, J., Farahat, E., Fuentes, M., Rocha, E., Salo, R., Seftigen, K., Stridbeck, P. and Liu, Y., 2014. Growth dynamics of tree-line and lake-shore Scots pine (*Pinus sylvestris L.*) in the central Scandinavian Mountains during the Medieval Climate Anomaly and the early Little Ice Age. Fron. Ecol. Evol. 2, doi: 10.3389/fevo.2014.00020.

Liu, Y., An, Z., Linderholm, H.W., Chen, D., Song, H., Cai, Q., Sun, J., Tian, H., 2009. Annual temperatures during the last 2485 years in the mid-eastern Tibetan Plateau inferred from tree rings. Sci. China Series D: Earth Sci. 52, 348–359.

Ljungqvist, F.C., 2010. A new reconstruction of temperature variability in the extra-tropical Northern Hemisphere during the last two millennia. Geogr. Ann. 92, 339–351.

Ljungqvist, F.C., Krusic, P.J, Brattström, G., Sundqvist, H.S., 2012. Northern Hemisphere temperature patterns in the last 12 centuries. Clim. Past 8, 227–249.

Ljungqvist, F.C., Krusic, P.J., Sundqvist, H., Zorita, E., Brattstrom, G., Frank, D., 2016. Northern Hemisphere hydroclimatic variability over the past 12 centuries. Nature 532, 94–98.

Lloyd, A.H., Graumlich, L.J., 1997. Holocene dynamics of treeline forests in the Sierra Nevada. Ecology 78, 1199–1210.

Luckman, B.H., Wilson, R.J.S., 2005. Summer temperatures in the Canadian Rockies during the last millennium: a revised record. Clim. Dyn. 24, 131–144.

Mann, M.E., Jones, P.D., 2003. Global surface temperatures over the past two millennia. Geophys. Res. Lett. 30, doi: 10.1029/2003GL017814.

Mann, M.E., Zhang, Z., Hughes, M.K., Bradley, R.S., Miller, S.K., Rutherford, S., Ni, F., 2008. Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. Proc. Nat. Acad. Sci. 105, 13252–13257.

Meko, D.M., 1981. Applications of Box-Jenkins methods of time series analysis to the reconstruction of drought from tree rings. Ph.D. Dissertation. University of Arizona, Tucson.

Melvin, T.M., Briffa, K.R., 2008. A “signal-free” approach to dendroclimatic standardisation. Dendrochronologia 26, 71–86.

Melvin, T.M., Grudd, H., Briffa, K.R., 2013. Potential bias in ‘updating’ tree-ring chronologies using Regional Curve Standardization: re-processing the Torneträsk maximum-latewood-density data. Holocene 23, 364–373.

Myglan, V.S., Oidupaa, O.C., Vaganov, E.A., 2012a. A 2367-year tree-ring chronology for the Altai-Sayan region (Mongun-Taiga Mountain Massif). Archaeol. Ethnol. Anthropol. Eurasia 40, 76–83.

Myglan, V.S., Zharnikova, O.A., Malysheva, N.V., Gerasimova, O.V., Vaganov, E.A., Sidorov, O.V., 2012b. Constructing the tree-ring chronology and reconstructing summertime air temperatures in southern Altai for the last 1500 years. Geogr. Nat. Resour. 33, 200–207.

Neukom, R., et al., 2014. Inter-hemispheric temperature variability over the past millennium. Nature Clim. Change 4, 362–367.

Neuwirth, B., Esper, J., Schweingruber, F.H., Winiger, M., 2004. Site ecological differences to the climatic forcing of spruce pointer years from the Lötschental, Switzerland. Dendrochronologia 21, 69–78.

Oke, T.R., 2007. Siting and exposure of meteorological instruments at urban sites. In: Air Pollution Modeling and Its Application XVII. Springer US, 615-631.

Osborn, T.J., Briffa, K.R., Jones, P.D., 1997. Adjusting variance for sample-size in tree-ring chronologies and other regional-mean time-series. Dendrochronologia 15, 89–99.

Pages 2k Consortium, 2013. Continental-scale temperature variability over the Common Era. Nature Geosc. 6, 339–346.

Pages 2k PMIP3 group, 2015. Continental-scale temperature variability in PMIP3 simulations and Pages 2k regional temperature reconstructions over the past millennium. Clim. Past Discuss. 11, 2483–2555.

Parker, D.E., 1994. Effects of changing exposure of thermometers at land stations." Int. J. Climatol. 14, 1–31.

Salzer, M.W., Kipfmueller, K.F., 2005. Reconstructed temperature and precipitation on a millennial timescale from tree-rings in the southern Colorado Plateau, USA. Clim. Change 70, 465–487.

Salzer, M.W., Bunn, A.G., Graham, N.E., Hughes, M.K., 2014. Five millennia of paleotemperature from tree-rings in the Great Basin, USA. Clim. Dyn. 42, 1517–1526.

Schneider, L., Smerdon, J.E., Büntgen, U., Wilson, R.J.S., Myglan, V.S., Kirdyanov, A.V., Esper, J., 2015. Revising midlatitude summer temperatures back to A.D. 600 based on a wood density network. Geophys. Res. Lett. 42, doi: 10.1002/2015GL063956.

Schweingruber, F.H.,1996. Tree Rings and Environment: Dendroecology. Haupt Verlag, Bern.

Schweingruber, F.H., Bartholin, T., Schär, E., Briffa, K.R., 1988. Radiodensitometric-dendroclimatological conifer chronologies from Lapland (Scandinavia) and the Alps (Switzerland). Boreas 17, 559–566.

Shi, F., Yang, B., Mairesse, A., von Gunten, L., Li, J., Bräuning, A., Yang, F., Xiao, X., 2013. Northern Hemisphere temperature reconstruction during the last millennium using multiple annual proxies. Clim. Res. 56, 231–244.

Sidorova, O.V., Naurzbaev, M.M., Vaganov, E.A., 2006. An integral estimation of tree-ring chronologies from subarctic regions of Eurasia. Trace 4, 84–91.

Smerdon, J.E., Kaplan, A., Zorita, E., González-Rouco, J.F., Evans, M.N., 2011. Spatial performance of four climate field reconstruction methods targeting the Common Era. Geophyss Res. Lett. 38, doi: 10.1029/2011GL047372.

Smerdon, J.E., Coats, S., Ault, T.R., 2015. Model-dependent spatial skill in pseudoproxy experiments testing climate field reconstruction methods for the Common Era. Clim. Dyn., doi: 10.1007/s00382-015-2684-0.

St. George, S., 2014. An overview of tree-ring width records across the Northern Hemisphere. Quat. Sci. Rev. 95, 132–150.

Stoffel, M, Khodri, M., Corona, C., Guillet, S., Poulain, V., Bekki S., Guiot, J., Luckman, B.H., Oppenheimer, C., Lebas N., Beniston M., Masson-Delmotte, V., 2015. Estimates of volcanic-induced cooling in the Northern Hemisphere over the past 1,500 years. Nature Geosc. 8, 784–788.

Støve, B., Ljungqvist, F.C., Thejll, P., 2012. A test for non-linearity in temperature proxy records. J. Clim. 25, 7173–7186.

Tegel, W., Vanmoerkerke, J., Büntgen, U., 2010. Updating historical tree-ring records for climate reconstruction. Quat. Sci. Rev. 29, 1957–1959.

Trouet, V., Diaz, H.F., Wahl, E.R., Viau, A.E., Cook, E.R., 2013. A 1500-year reconstruction of annual mean temperature for temperate North America on decadal-to-multidecadal time-scales. Environ. Res. Lett. 8, doi: 10.1088/1748-9326/8/2/024008.

Villalba, R., 1990. Climatic fluctuations in northern Patagonia during the last 1000 years as inferred from tree-ring records. Quat. Res. 34, 346–360.

von Storch, H., Zorita, E., Jones, J., Dimitriev, Y., González-Rouco, J.F, Tett, S., 2004. Reconstructing past climate from noisy data. Science 306, 679–682.

Wang, J., Yang, B., Qin, C., Kang, S., He, M., Wang, Z., 2014. Tree-ring inferred annual mean temperature variations on the southeastern Tibetan Plateau during the last millennium and their relationships with the Atlantic Multidecadal Oscillation. Clim. Dyn. 43, 627–640.

Wigley, T.M.L., Briffa K.R., Jones, P.D., 1984. On the average of correlated time series, with applications in dendroclimatology and hydrometeorology. J. Clim. Appl. Meteorol. 23, 201–213.

Wiles, G.C., D'Arrigo, R.D., Barclay, D., Wilson, R.S., Jarvis, S.K., Vargo, L., Frank, D., 2014. Surface air temperature variability reconstructed with tree rings for the Gulf of Alaska over the past 1200 years. Holocene 24, 198–208 .

Wilson, R.J.S., Esper, J., Luckman, B.H., 2004. Utilizing historical tree-ring data for dendroclimatology: a case study from the Bavarian Forest, Germany. Dendrochronologia 21, 53–68.

Wilson, R., D'Arrigo, R.D., Buckley, B., Büntgen, U., Esper, J., Frank, D., Luckman, B., Payette, S., Vose, R., Youngblut, D., 2007. A matter of divergence: tracking recent warming at hemispheric scale using tree ring data. J. Geophys. Res. 112, D17103.

Wilson, R.J.S., et al., 2016. Last millennium Northern Hemisphere summer temperatures from tree rings. Part I: the long term context. Quat. Sci. Rev. 134, 1–18.

Xing, P., Chen, X., Luo, Y., Nie. S., Zhao, Z., Huang, J., Wang, S., 2016. The extratropical Northern Hemisphere temperature reconstruction during the last millennium based on a novel method. Plos One, doi: 10.1371/journal.pone.0146776.

Yadav, R.R., Braeuning, A., Singh, J., 2011. Tree ring inferred summer temperature variations over the last millennium in western Himalaya, India. Clim. Dyn. 36, 1545–1554.

Yang, B., Qin, C., Wang, J.L., He, M.H., Melvin, T.M., Osborn, T.J., Briffa, K.R., 2014. A 3,500-year tree-ring record of annual precipitation on the northeastern Tibetan Plateau. Proc. Nat. Acad. Sci. 111, 2903–2908.

Zhang, Y., Shao, X.M., Yin, Z.Y., Wang, Y., 2014. Millennial minimum temperature variations in the Qilian Mountains, China: evidence from tree rings. Clim. Past 10, 1763–1778.

Zhang, H., Yuan, N., Esper, J., Werner, J.P., Xoplaki, E., Büntgen, U., Treydte, K., Luterbacher, J., 2015. Modified climate with long term memory in tree ring proxies. Environ. Res. Lett. 10, doi: 10.1088/1748-9326/10/8/084020.

Zhu, H., Zheng, Y., Shao, X., Liu, X., Xu, Y., Liang, E., 2008. Millennial temperature reconstruction based on tree-ring widths of Qilian juniper from Wulan, Qinghai Province, China. Chin. Sci. Bull. 53, 3914–3920.

Zorita, E., González-Rouco, J.F., Legutke. S., 2003. Testing the Mann et al. (1998) approach to paleoclimate reconstructions in the context of a 1000-yr control simulation with the ECHO-G coupled climate model. J. Clim. 16, 1378–1390.

Table and Figure Captions

Table 1 Millennial-length tree-ring based temperature reconstructions. Superscript * indicates reconstructions developed using MXD (instead of TRW). The Icefield reconstruction contains both MXD and TRW data. *Signal* specifies the seasonality of reconstructed temperatures, with *p* indicating previous-year months. *T* is temperature, *Tmax* is maximum temperature, *Tmin* is minimum temperature.

Table 2 *Data Homogeneity* scores. Chronology type *C* refers to reconstructions derived from a composite of material from living trees, remnant, historical and/or sub-fossil wood. Type *L* refers to reconstructions derived from only living trees. Temporal clustering (*Yes*) indicates records composed of data from distinct sites or species concentrated in discrete periods over the past 1000 years.

Table 3 *Sample Replication* scores. The number of TRW (or MXD) measurement series included in the reconstructions. *11th/20th* is the ratio of the mean replication during the 11th century relative to the mean replication during the 20th century. Values in red are estimates.

Table 4 *Growth Coherence* scores. Mean, maximum, and minimum correlations among the TRW (or MXD) series included in the reconstructions. *11th/20th* is the ratio of the correlation during the 11th century relative to the 20th century correlation. Values in red are estimates.

Table 5 *Chronology Development* scores. *Detrending method* 1 = RCS (and Signal Free), and 2 = individual detrending. *Age range* is the difference between highest and lowest point on the mean age curve over the past millennium. *Age trend* is the slope of a linear regression fit to the mean age curve over the past millennium (times 100). *Maximum frequency* indicates the wavelength of lowest frequency information retained in a reconstruction, with 1 = multi-centennial, 2 = centennial, and 3 = decadal. Values in red are estimates.

Table 6 *Climate Signal* scores. *Length* is the period of overlap with instrumental temperature data in years. *Correlation* is the Pearson correlation coefficient between the tree-ring chronology and the instrumental data over the calibration period. *Calibration/verification difference* indicates the correlation range between different periods of overlap with instrumental data. *Truncation* = 0.5 if the calibration period was shortened (e.g. due to divergence), *truncation* = 1 if this is not the case. Values in red are estimates.

Table 7 Ranking of 39 tree-ring based temperature reconstructions based on their *Data Homogeneity, Sample replication, Growth Coherence, Chronology Development, and Climate Signal* scores. Last column indicates which datasets are publicly available.

Fig. 1 Location of millennial-length tree-ring based temperature reconstructions (circles). Colors indicate the June-August temperature change between the mean of the period 1964-2013 minus the mean of the period 1914-1963 using GISS 1200 km gridded data.

Fig. 2 Tree-ring based temperature reconstructions. Black curves are the 13 reconstructions from Europe (**a**), 14 from Asia (**b**), 8 from North America (**c**), and 4 from the Southern Hemisphere (**d**) shown as anomalies from their 20th century means. Note that the reconstructed temperature variance differs substantially among records, largely as a result of the differing calibration schemes used in the original publications. Colored curves are the arithmetic means calculated over the common period of all reconstructions in each region. **e**, Comparison of the mean timeseries from Europe, Asia, and North America.

Fig. 3 TRC replication curves. Black curves show the changing numbers of TRW (or MXD) measurement series within the temperature reconstructions from Europe (**a**), Asia (**b**), North America (**c**), and the Southern Hemisphere (**d**). The replication curve of the Alps (larch) reconstruction in (**a**) refers to the right axis. Colored curves are the arithmetic means calculated over the common period covered by all reconstructions in each region. **e**, Comparison of the mean curves.

Fig. 4 TRC inter-series correlations. Black curves show the correlation coefficients among the TRW (or MXD) measurement series integrated in the local temperature reconstructions from Europe (**a**), Asia (**b**), North America (**c**), and the Southern Hemisphere (**d**). Correlations are calculated over 100-year periods shifted in 10-year steps throughout the past millennium. The earliest value is centered on 1050, the most recent value on 1950. Colored curves are the arithmetic means calculated for each region, and dashed lines indicate the mean values over the millennium. **e**, Comparison of the mean inter-series correlation curves.

Fig. 5 TRC age curves. Black curves show the mean tree age of the TRW (or MXD) data integrated in the temperature reconstructions from Europe (**a**), Asia (**b**), North America (**c**), and the Southern Hemisphere (**d**). Colored curves are the arithmetic means calculated over the common period covered by all reconstructions in each region. **e**, Comparison of mean replication curves.

Table1

Record	Reference	Continent	Lat./Lon.	Species	Signal
Alps (Larch)	Büntgen et al. 2009	Europe	45-47N 6-14E	Larix decidua	Jun-Jul T
Boreal Plateau	Lloyd & Graumlich 1997	N-America	36.3N 118.5W	Pinus balfouriana	Annual T
Central Alps	Büntgen et al. 2011	Europe	46-47N 10-12E	Larix decidua, Pinus cem.	Jun-Aug T
Crabtree	Graumlich 1993	N-America	36.5N 118.3W	Pinus balfouriana	Jun-Aug T
Dulan	Liu et al. 2009	Asia	36N 98-99E	Sabina przewalskii	pJan-pDec T
Dzhelo	Myglan et al. 2012b	Asia	50N 87.9E	Larix sibirica	Jun-Jul T
E-Canada	Gennaretti et al. 2014	N-America	54-55N 70-72W	Picea mariana	Jul-Aug T
Finland	Helama et al. 2010	Europe	67-69N 20-28E	Pinus sylvestris	Jun-Aug T
French Alps	Büntgen et al. 2012	Europe	44N 7.3E	Larix decidua	Jun-Aug T
Great Basin	Salzer et al. 2014	N-America	37-40N 114-118W	Pinus longaeva	Jul-Sep T
Gulf of Alaska	Wiles et al. 2014	N-America	58-61N 134-149W	Tsuga mertensiana	Feb-Aug T
Icefield*	Luckman & Wilson 2005	N-America	51-53N 117-119W	Picea engel., Abies lasio.	May-Aug Tmax
Indigirka	Sidorova et al. 2006	Asia	70N 148E	Larix kajanderi	Jun-Jul T
Jämtland	Linderholm & Gunnarson 2005	Europe	63.2N 12–13E	Pinus sylvestris	Jun-Aug T
Karakorum	Esper et al. 2002b	Asia	35-36N 74-75E	Juniperus spec.	Annual T
Lauenen*	Schweingruber et al. 1988	Europe	46.4N 7.3E	Picea abies, Abies alba	Jun-Aug T
Lötschental*	Büntgen et al. 2006a	Europe	46.3N 7.8E	Larix decidua	Jun-Sep T
Mongolia	D'Arrigo et al. 2001	Asia	48.3N 98.9E	Pinus sibirica	pAug-Jul T
Mongun	Myglan et al. 2012a	Asia	50.3N 90E	Larix sibirica	Jun-Jul T
N-Scan*	Esper et al. 2012	Europe	67-69N 20-28E	Pinus sylvestris	Jun-Aug T
Oroko Swamp	Cook et al. 2002	Australia	43.2S 170.3E	Lagarostrobos colensoi	Jan-Mar T
Polar Ural*	Briffa et al. 2013	Asia	66.8N 65.6E	Larix sibirica	Jun-Aug T
Qamdo	Wang et al. 2014	Asia	31.1N 97.2E	Sabina tibetica	pJan-pDec T
Qilian	Zhang et al. 2014	Asia	38.7N 99.7E	Sabina przewalskii	Jan-Aug Tmin
Rio Alerce	Villalba 1990	S-America	41S 73W	Fitzroya cupressoides	Dec-Feb T
S-Chile	Lara & Villalba 1993	S-America	41.5S 72.6W	Fitzroya cupressoides	pDec-pMar T
S-Finland*	Helama et al. 2014	Europe	61-62N 28-29E	Pinus sylvestris	May-Sep T
Southern Colorado	Salzer & Kipfmueller 2005	N-America	35.3N 111.7W	Pinus aristata	Annual Tmax
Swiss/Austrian Alps	Büntgen et al. 2005	Europe	46-47N 7-11E	Larix decidua, Pinus cem.	Jun-Aug T
Taimyr	Briffa et al. 2008	Asia	70-72N 95-105E	Larix gmelinii	Jun-Jul T
Tasmania	Cook et al. 2000	Australia	41.8S 145.5E	Lagarostrobos franklinii	Nov-Apr T
Tatra	Büntgen et al. 2013	Europe	48-49N 19-21E	Larix decidua	May-Jun T
Tien Shan	Esper et al. 2003b	Asia	40N 71-72E	Juniperus spec.	Jun-Sep T
Torneträsk (MXD)*	Melvin et al. 2013	Europe	68.2N 19.5E	Pinus sylvestris	May-Aug T
Torneträsk (TRW)	Melvin et al. 2013	Europe	68.2N 19.5E	Pinus sylvestris	May-Aug T
Upper Wright	Lloyd & Graumlich 1997	N-America	36.3N 118.3W	Pinus balfouriana	Annual T
W-Himalaya	Yadav et al. 2011	Asia	32-33N 76-77E	Juniperus polycarpos	May-Aug T
Wulan	Zhu et al. 2008	Asia	37N 98.7E	Sabina przewalskii	pSep-Apr T
Yamal	Briffa et al. 2013	Asia	67-68N 69-71E	Larix sibirica	Jun-Jul T

Table2

	1. Source	2. Chronology type	3. Species number	4. Temporal clustering	5. Remark	6. Homogeneity
Dzhelo	Fossil. One site.	C	1	No	Single treeline site.	🟢
Tasmania	Sub-fossil. One site.	C	1	No		🟢
Rio Alerce	One valley.	L	1	No	Living trees from one valley.	🟢
Qamdo	Fossil. One site.	C	1	No	Single site in high elevation (4350–4500 m asl.).	🟢
Mongolia	Fossil. One site.	C	1	No	Single treeline site.	🟢
Torneträsk (MXD)	Fossil. Around one lake.	C	1	No	Varying measurement techniques.	🟩
S-Chile	One valley.	L	1	No	Logged and living trees from one slope.	🟩
Oroko Swamp	Sub-fossil. One site.	C	1	No	Including samples from moist and dry microsites.	🟩
Southern Colorado	Fossil. Several sites.	C	1	No	Two summit areas of San Francisco Peaks.	🟩
Polar Ural	Fossil. Several sites.	C	1	No	Elevational transects.	🟩
Crabtree	Two sites.	L	1	No		🟩
Qilian	Fossil. Four sites.	C	1	No	152 out of 250 samples used for reconstruction.	🟩
Upper Wright	Fossil. One site.	C	1	Yes	Snag material from above current treeline.	🟩
Boreal Plateau	Fossil. One site.	C	1	Yes	Snag material from above current treeline.	🟩
N-Scan	Sub-fossil. Several lakes & sites.	C	1	No	Living trees from lakeshores.	🟦
E-Canada	Sub-fossil. Several lakes & sites.	C	1	No	Living trees from lakeshores.	🟦
Finland	Sub-fossil. Several lakes & sites.	C	1	Yes	Living trees from dry sites, sub-fossil from lakes.	🟦
Yamal	Sub-fossil. Multiple sites.	C	1	No	Two-curve RCS. Normal distribution transformation.	🟦
Lötschental	Historical. Two valleys.	C	1	Yes	Pre-1200 data from Simplon valley.	🟦
Torneträsk (TRW)	Fossil & sub-fossil. Several lakes & sites.	C	1	No		🟦
Indigirka	Fossil & sub-fossil. Several sites.	C	1	Yes	Trees from upper timberline & flood plain trerrace.	🟦
Swiss/Austrian Alps	Historical & sub-fossil. Several sites.	C	2	Yes	Multiple RCS.	🟦
Mongun	Fossil. Multiple sites.	C	1	No	Multiple sites within 35 km^2.	🟦
Central Alps	Fossil, sub-fossil & historic. Several sites.	C	2	Yes	Several sites in Austria and Switzerland.	🟦
Dulan	Historical. Several sites.	C	1	Yes	Historical material from lower elevations.	🟦
French Alps	Fossil. Several sites.	C	1	No	Temprally varying climate signal reported.	🟦
Lauenen	Historical. Several sites.	C	2	No	Various buildings in lower elevation.	🟦
Wulan	One side.	L	1	Yes	50% weakly correlating samples removed.	🟦
Jämtland	Sub-fossil. Several lakes & sites.	C	1	No		🟦
Tatra	Historical. Several sites.	C	1	Yes	Historical data from lower elevation sites.	🟦
Taimyr	Sub-fossil. Multiple sites.	C	1	Yes	Elevational & latitudinal ecotones over larger region.	🔵
Icefield	Fossil & sub-fossil. Several sites.	C	2	Yes	Multiple RCS runs.	🔵
Alps (Larch)	Fossil/sub-fossil/historic. Multiple sites.	C	1	No	Multiple sites from larger region.	🔵
Gulf of Alaska	Fossil & sub-fossil. Multiple sites.	C	1	Yes	Multiple sites from larger region.	🔵
Tien Shan	Multiple sites.	L	2	Yes	Increasing correlation back in time.	🔵
S-Finland	Sub-fossil. Several lakes & sites.	C	1	Yes	Varying measurement techniques and micro-sites.	🔵
Great Basin	Fossil. Three sites.	C	1	No		🔵
Karakorum	Multiple sites.	L	2	Yes	Ten sites in five valleys.	🔵
W-Himalaya	Multiple sites.	L	1	No	Sites from 3200-3600 m asl. 100 km distance.	🔵

🟢 Class-A 🟩 Class-B 🟦 Class-C 🔵 Class-D

Table3

	1. Mean	2. Maximum	3. Minimum	4. 11th/20th [%]	5. Replication
Mongun	79	124	14	250	🟢
Oroko Swamp	103	186	15	113	🟢
Dzhelo	36	57	11	251	🟢
E-Canada	185	282	84	47	🟢
Tasmania	61	76	11	170	🟢
N-Scan	53	197	25	38	🟢
Gulf of Alaska	176	352	43	33	🟢
Central Alps	208	405	74	25	🟢
Swiss/Austrian Alps	253	407	10	25	🟢
Dulan	159	218	42	34	🟢
Southern Colorado	55	69	17	85	○
Qilian	68	95	36	53	○
Torneträsk (TRW)	74	178	5	39	○
Lauenen	20	54	2	64	○
Indigirka	43	65	13	75	○
Qamdo	67	104	19	45	○
Icefield	123	274	29	20	○
Great Basin	73	116	24	34	○
Finland	91	436	24	13	○
Karakorum	115	203	23	18	○
Lötschental	43	80	8	23	🔵
S-Chile	21	25	16	97	🔵
Yamal	57	155	27	24	🔵
Upper Wright	34	43	20	58	🔵
Alps (Larch)	530	1490	19	2	🔵
Mongolia	30	51	13	39	🔵
Torneträsk (MXD)	21	76	5	17	🔵
Wulan	60	101	14	18	🔵
Polar Ural	22	56	7	18	🔵
Jämtland	35	105	3	20	🔵
French Alps	104	292	10	6	●
Crabtree	25	31	7	34	●
W-Himalaya	70	150	13	9	●
Taimyr	61	131	4	10	●
S-Finland	15	60	3	11	●
Tien Shan	83	203	5	5	●
Rio Alerce	35	49	5	15	●
Boreal Plateau	28	50	3	14	●
Tatra	73	271	3	3	●

🟢 Class-A ○ Class-B 🔵 Class-C ● Class-D

Table4

	1. Mean	2. Maximum	3. Minimum	4. 11th/20th [%]	5. Growth coherence
Indigirka	0.45	0.59	0.31	153	🟢
Yamal	0.49	0.66	0.38	125	🟢
Taimyr	0.39	0.54	0.20	151	🟢
Finland	0.37	0.53	0.13	158	🟢
Boreal Plateau	0.44	0.65	0.29	116	🟢
Rio Alerce	0.26	0.49	0.30	149	🟢
S-Chile	0.27	0.45	0.18	152	🟢
Tien Shan	0.22	0.39	0.11	178	🟢
Crabtree	0.39	0.53	0.18	110	🟢
French Alps	0.41	0.58	0.31	92	🟢
Dulan	0.42	0.50	0.37	87	○
Upper Wright	0.42	0.57	0.30	80	○
E-Canada	0.23	0.48	0.11	120	○
N-Scan	0.41	0.74	0.24	69	○
Wulan	0.41	0.74	0.24	69	○
Qilian	0.31	0.40	0.20	105	○
Lötschental	0.33	0.58	0.02	100	○
Mongun	0.44	0.60	0.30	69	○
Dzhelo	0.44	0.63	0.29	66	○
Torneträsk (MXD)	0.44	0.61	0.24	66	○
Southern Colorado	0.35	0.49	0.23	79	🔵
Mongolia	0.39	0.72	0.16	61	🔵
Polar Ural	0.47	0.70	-0.20	75	🔵
Gulf of Alaska	0.26	0.39	0.14	92	🔵
Torneträsk (TRW)	0.38	0.57	0.18	61	🔵
Great Basin	0.33	0.46	0.24	62	🔵
Lauenen	0.42	0.74	-0.31	68	🔵
Jämtland	0.41	0.78	0.16	42	🔵
Oroko Swamp	0.18	0.26	0.13	92	🔵
Tasmania	0.27	0.46	0.18	55	🔵
Qamdo	0.20	0.39	0.10	65	●
Icefield	0.15	0.31	0.08	81	●
Karakorum	0.22	0.39	0.15	53	●
W-Himalaya	0.23	0.37	0.14	46	●
Swiss/Austrian Alps	0.17	0.37	0.08	33	●
Alps (Larch)	0.20	0.32	0.07	33	●
Central Alps	0.20	0.50	0.10	20	●
S-Finland	0.20	0.50	0.05	20	●
Tatra	0.28	0.68	-0.07	11	●

🟢 Class-A ○ Class-B 🔵 Class-C ● Class-D

Table 5

	1. Dentrending method	2. Age range [yrs.]	3. Age trend	4. Maximum frequency	5. Chronology development
N-Scan	1	83	-0.6	1	Class-A
Finland	1	107	-0.9	2	Class-A
Torneträsk (TRW)	1	106	1.0	2	Class-A
Lötschental	1	281	-0.9	2	Class-A
E-Canada	1	66	2.1	2	Class-A
Taimyr	1	136	4.2	1	Class-A
Indigirka	1	200	1.9	2	Class-A
Yamal	1	78	3.6	2	Class-A
Polar Ural	1	142	3.3	2	Class-A
Tatra	1	142	4.5	2	Class-A
Torneträsk (MXD)	1	174	-4.6	2	Class-B
Swiss/Austrian Alps	2	172	2.6	2	Class-B
Jämtland	1	198	6.3	2	Class-B
Oroko Swamp	1	278	6.3	2	Class-B
Dzhelo	1	250	8.2	2	Class-B
Icefield	2	203	6.2	2	Class-B
Alps (Larch)	1	298	10.8	2	Class-B
Central Alps	1	370	10.0	2	Class-B
Lauenen	2	205	4.7	3	Class-B
Mongolia	2	267	8.2	2	Class-B
S-Finland	2	200	10.0	2	Class-C
S-Chile	2	250	6.6	3	Class-C
Mongun	2	172	8.2	3	Class-C
Tasmania	1	254	24.2	2	Class-C
French Alps	2	225	13.2	2	Class-C
Gulf of Alaska	1	340	14.9	3	Class-C
Tien Shan	2	268	18.1	2	Class-C
Southern Colorado	2	299	19.0	2	Class-C
Great Basin	2	285	20.8	2	Class-C
Qilian	2	279	28.1	2	Class-C
W-Himalaya	1	600	40.0	2	Class-D
Boreal Plateau	2	434	27.6	2	Class-D
Qamdo	2	495	28.9	2	Class-D
Karakorum	2	568	31.5	2	Class-D
Wulan	2	418	27.6	3	Class-D
Upper Wright	2	528	42.1	2	Class-D
Rio Alerce	2	510	43.6	3	Class-D
Dulan	2	625	61.0	2	Class-D
Crabtree	2	737	78.0	2	Class-D

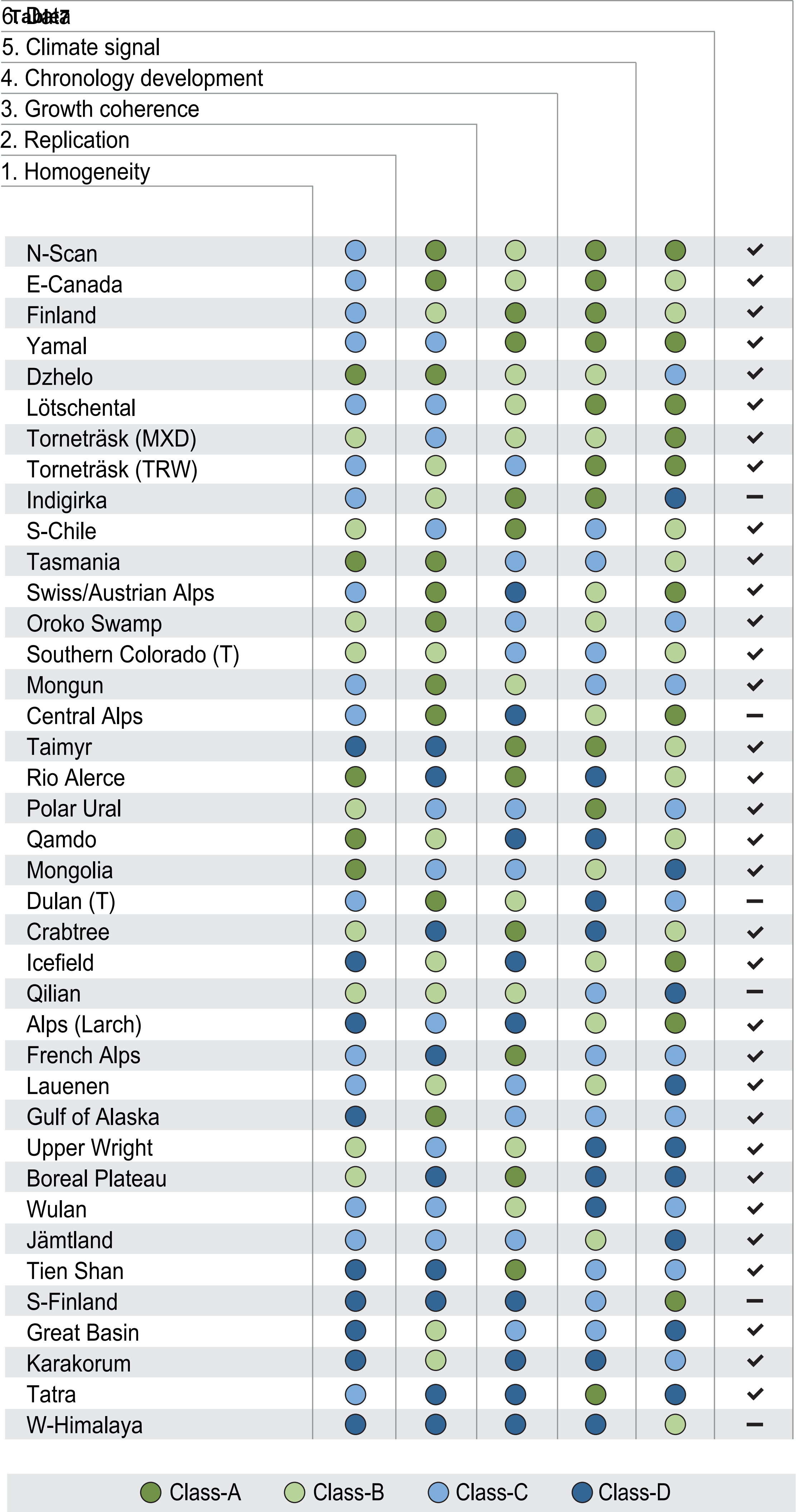
● Class-A ● Class-B ● Class-C ● Class-D

Table 6

	1. Length [yrs.]	2. Correlation	3. Calibration/verification difference	4. Truncation	5. Climate signal
Torneträsk (MXD)	147	0.84	0.02	1.0	🟢
N-Scan	131	0.77	0.03	1.0	🟢
Alps (Larch)	140	0.70	0.07	1.0	🟢
Lötschental	186	0.69	0.16	1.0	🟢
Swiss/Austrian Alps	139	0.64	0.03	1.0	🟢
S-Finland	253	0.63	0.18	1.0	🟢
Icefield	100	0.74	0.07	1.0	🟢
Central Alps	140	0.72	0.18	1.0	🟢
Yamal	123	0.67	0.10	1.0	🟢
Torneträsk (TRW)	147	0.60	0.10	1.0	🟢
Crabtree	116	0.58	0.03	1.0	○
Finland	128	0.66	0.15	1.0	○
Rio Alerce	77	0.65	0.04	1.0	○
E-Canada	102	0.61	0.11	1.0	○
Qamdo	53	0.71	0.03	1.0	○
S-Chile	78	0.61	0.05	1.0	○
Southern Colorado	86	0.68	0.15	1.0	○
Taimyr	57	0.77	0.18	1.0	○
Tasmania	106	0.68	0.25	1.0	○
W-Himalaya	104	0.65	0.25	1.0	○
Dulan	43	0.69	0.07	1.0	🔵
Tien Shan	102	0.45	0.06	1.0	🔵
Wulan	47	0.64	0.07	1.0	🔵
Polar Ural	123	0.52	0.21	1.0	🔵
Dzhelo	68	0.55	0.15	1.0	🔵
French Alps	102	0.42	0.10	1.0	🔵
Karakorum	115	0.31	0.01	1.0	🔵
Mongun	35	0.62	0.10	1.0	🔵
Gulf of Alaska	95	0.62	0.02	0.5	🔵
Oroko Swamp	79	0.64	0.03	0.5	🔵
Mongolia	112	0.44	0.20	1.0	●
Jämtland	86	0.63	0.09	0.5	●
Lauenen	76	0.34	0.08	1.0	●
Indigirka	45	0.61	0.10	0.5	●
Boreal Plateau	92	0.19	0.04	1.0	●
Great Basin	108	0.40	0.27	1.0	●
Upper Wright	92	0.17	0.03	1.0	●
Tatra	54	0.60	0.30	0.5	●
Qilian	53	0.76	0.63	1.0	●

🟢 Class-A ○ Class-B 🔵 Class-C ● Class-D

Table 7



	1. Homogeneity	2. Replication	3. Growth coherence	4. Chronology development	5. Climate signal	6. Data
N-Scan	C	A	B	A	A	✓
E-Canada	C	A	B	A	B	✓
Finland	C	B	A	A	B	✓
Yamal	C	C	A	A	A	✓
Dzhelo	A	A	B	B	C	✓
Lötschental	C	C	B	A	A	✓
Torneträsk (MXD)	B	C	B	B	A	✓
Torneträsk (TRW)	C	B	C	A	A	✓
Indigirka	C	B	A	A	D	–
S-Chile	B	C	A	C	B	✓
Tasmania	A	A	C	C	B	✓
Swiss/Austrian Alps	C	A	D	B	A	✓
Oroko Swamp	B	A	C	B	C	✓
Southern Colorado (T)	B	B	C	C	B	✓
Mongun	C	A	B	C	C	✓
Central Alps	C	A	D	B	A	–
Taimyr	D	D	A	A	B	✓
Rio Alerce	A	D	A	D	B	✓
Polar Ural	B	C	C	A	C	✓
Qamdo	A	B	D	D	B	✓
Mongolia	A	C	C	B	D	✓
Dulan (T)	C	A	B	D	C	–
Crabtree	B	D	A	D	B	✓
Icefield	D	B	D	B	A	✓
Qilian	B	B	B	C	D	–
Alps (Larch)	D	C	D	B	A	✓
French Alps	C	D	A	C	C	✓
Lauenen	C	B	C	B	D	✓
Gulf of Alaska	D	A	C	C	C	✓
Upper Wright	B	C	B	D	D	✓
Boreal Plateau	B	D	A	D	D	✓
Wulan	C	C	B	D	C	✓
Jämtland	C	C	C	B	D	✓
Tien Shan	D	D	A	C	C	✓
S-Finland	D	D	D	C	A	–
Great Basin	D	B	C	C	D	✓
Karakorum	D	B	D	D	C	✓
Tatra	C	D	D	A	D	✓
W-Himalaya	D	D	D	D	B	–

Class-A Class-B Class-C Class-D

Figure1

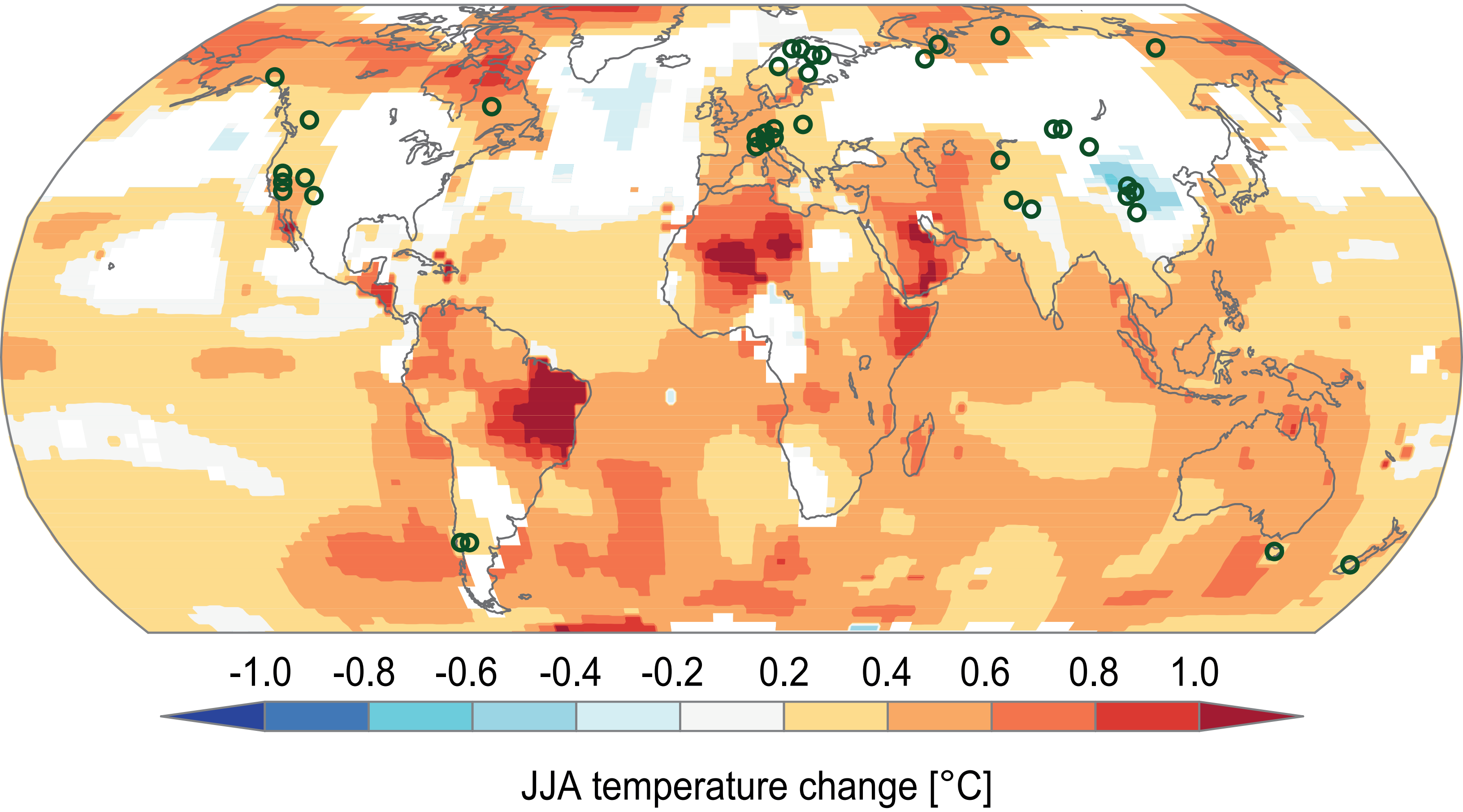
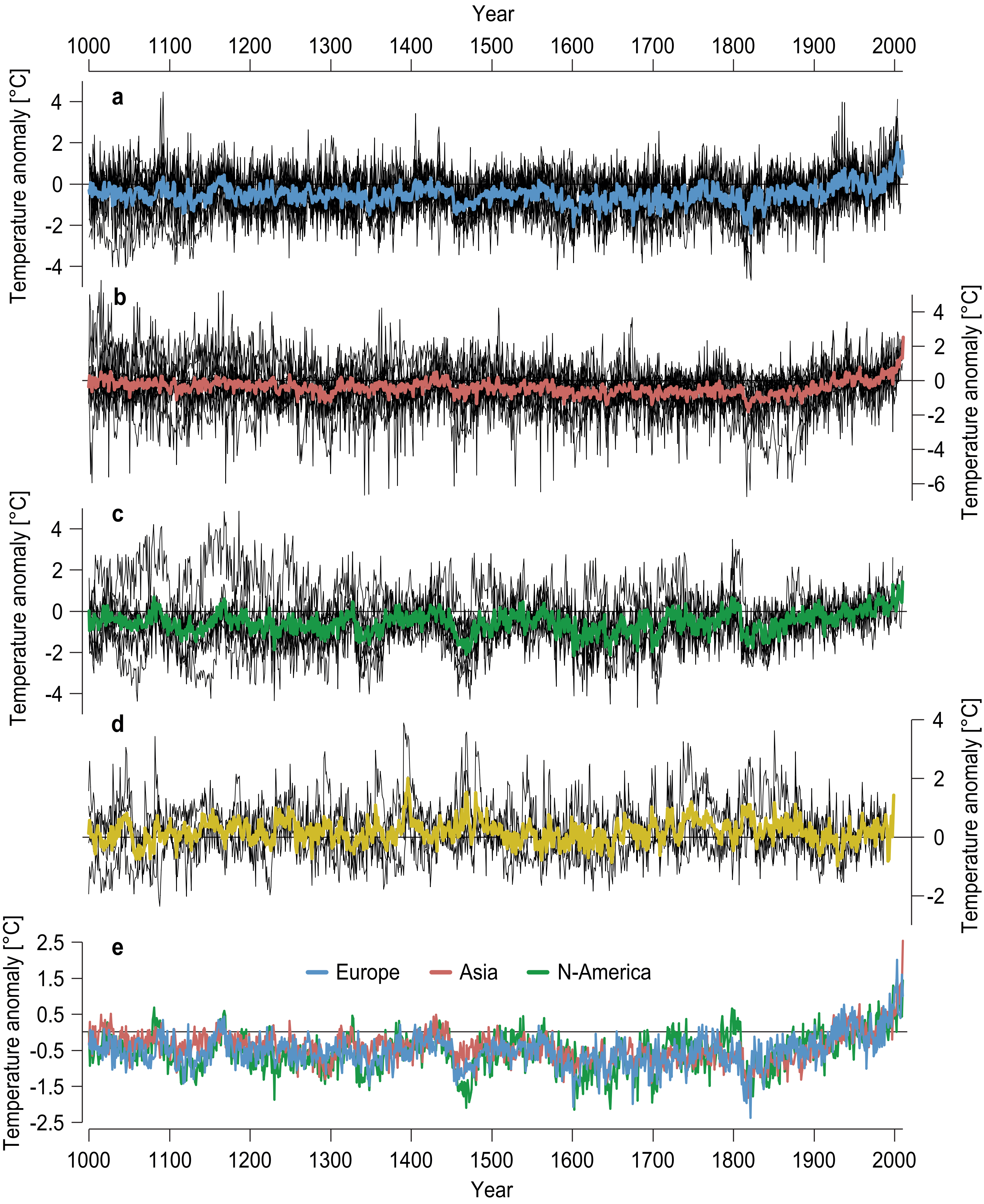


Figure2



Year
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
a
Temperature anomaly [°C]
4
2
0
-2
-4
b
Temperature anomaly [°C]
4
2
0
-2
-4
-6
c
Temperature anomaly [°C]
4
2
0
-2
-4
d
Temperature anomaly [°C]
4
2
0
-2
e
Europe
Asia
N-America
Temperature anomaly [°C]
2.5
1.5
0.5
-0.5
-1.5
-2.5
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
Year

Figure3

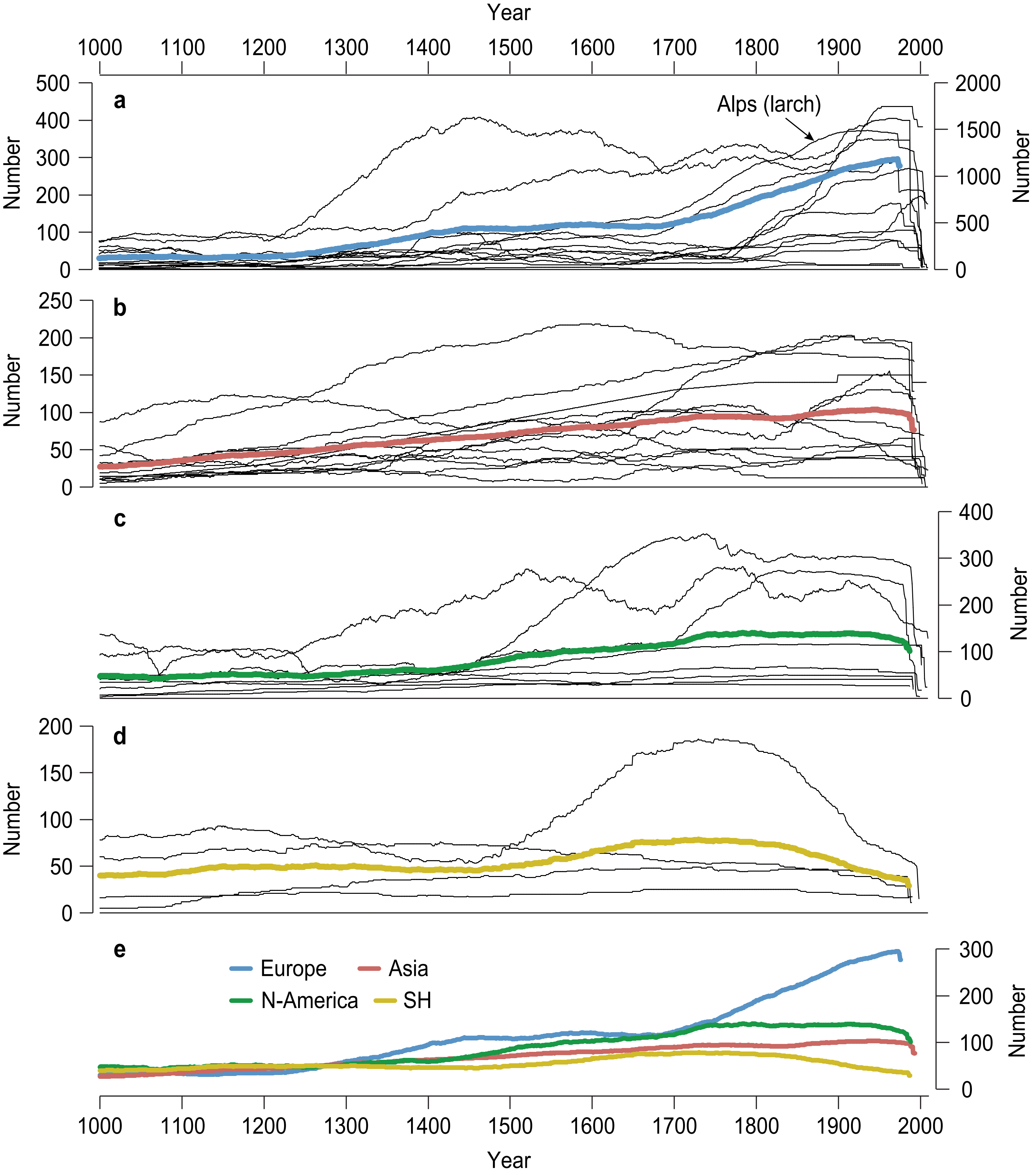

Year
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
a
Alps (larch)
Number
500
400
300
200
100
0
2000
1500
1000
500
0
b
250
200
150
100
50
0
c
400
300
200
100
0
d
200
150
100
50
0
e
Europe
Asia
N-America
SH
300
200
100
0
Year

Figure4

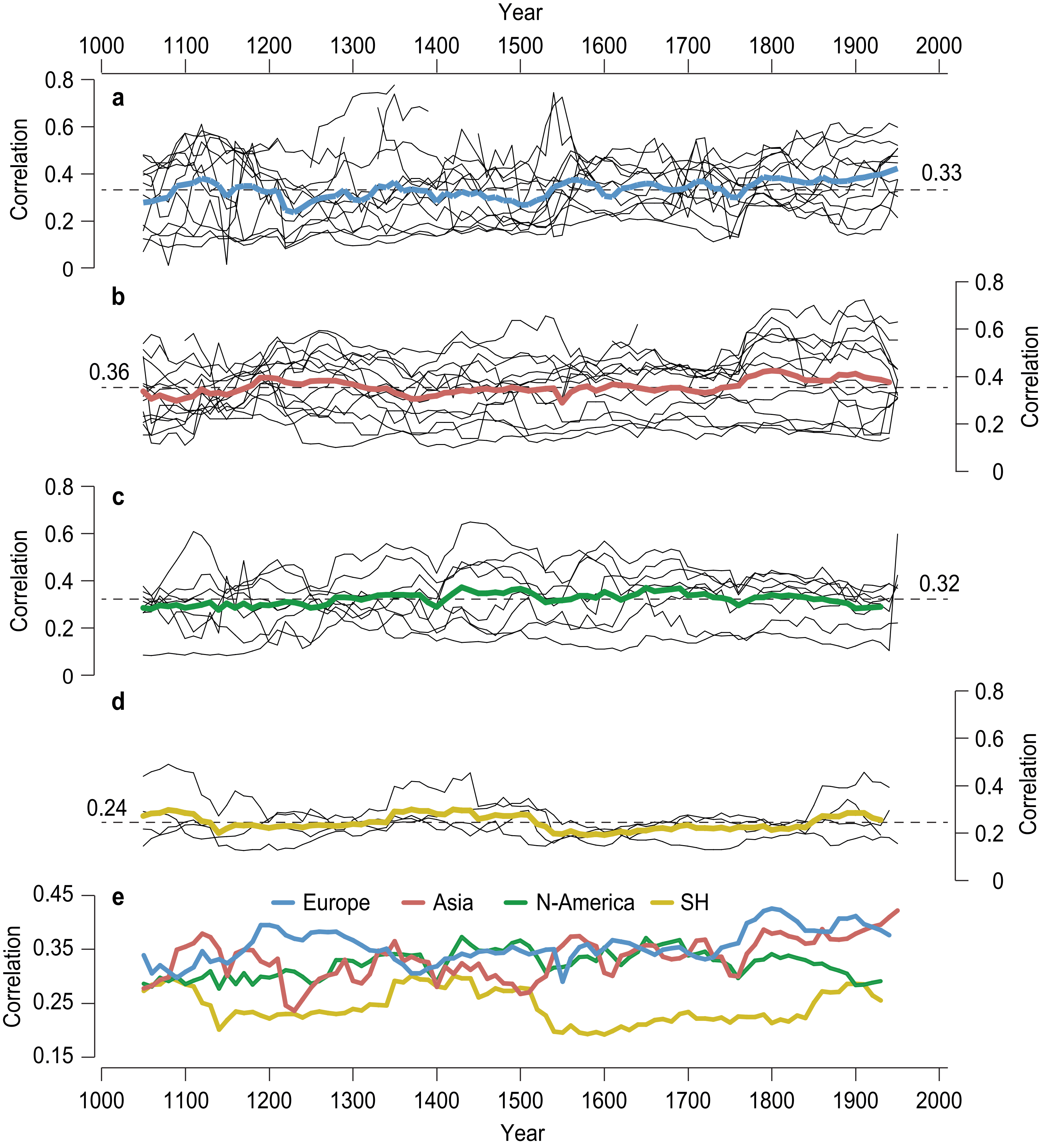
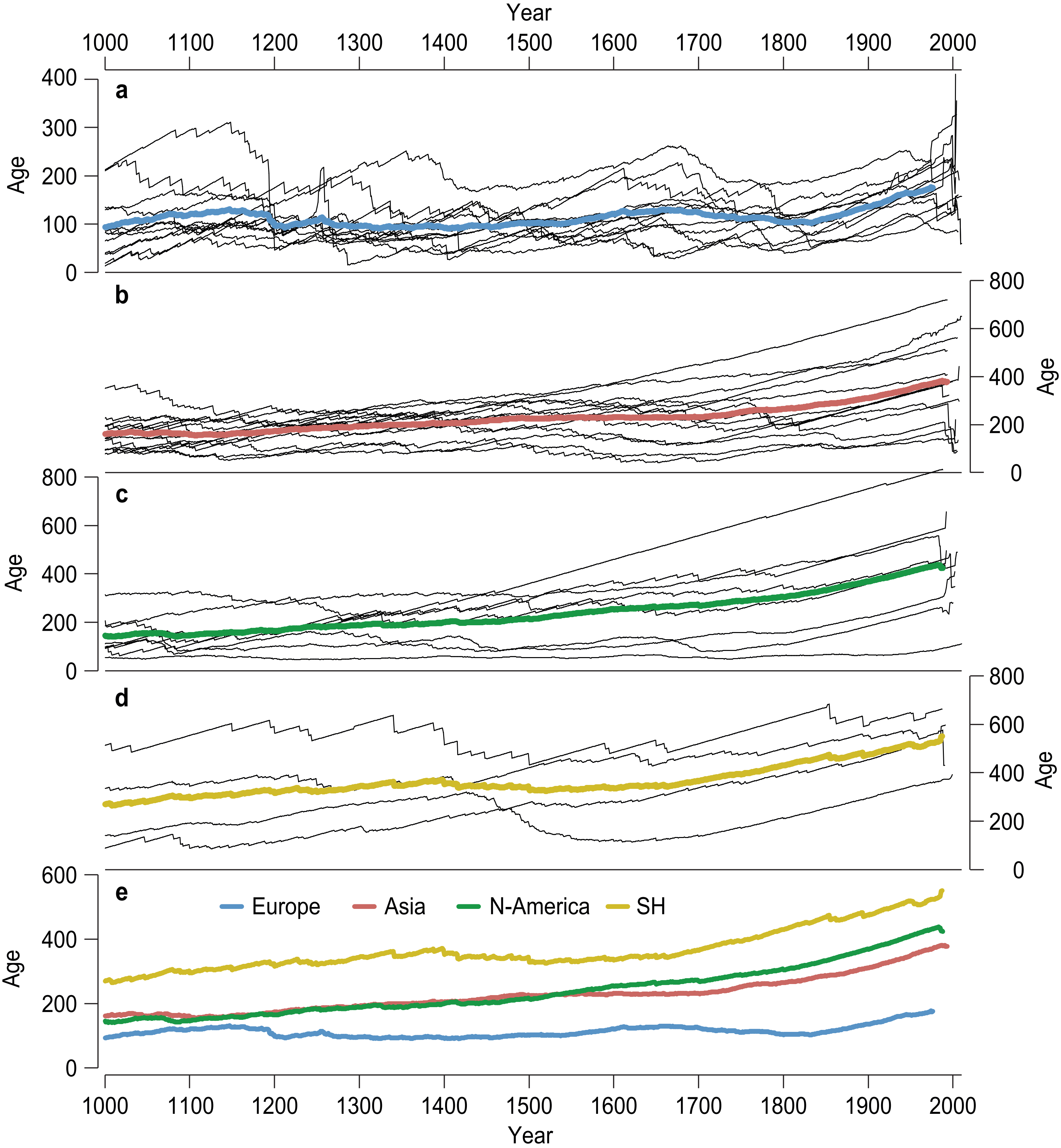

Year
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
a
Correlation
0.8
0.6
0.4
0.2
0
0.33
b
0.36
Correlation
0.8
0.6
0.4
0.2
0
c
Correlation
0.8
0.6
0.4
0.2
0
0.32
d
0.24
Correlation
0.8
0.6
0.4
0.2
0
e
Europe
Asia
N-America
SH
Correlation
0.45
0.35
0.25
0.15
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
Year

Figure5


Year
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
a
b
c
d
e
Age
400
300
200
100
0
800
600
400
200
0
600
Europe
Asia
N-America
SH

Ranking of tree-ring based temperature reconstructions of the past millennium

Highlights

Basic tree-ring chronology characteristics are detailed.

An ordinal scoring scheme understandable to non-specialists is proposed.

The scheme is used to rank 39 tree-ring based temperature reconstructions.

The ranking supports the development of large-scale temperature reconstructions.

Updates will be published online at: www.blogs.uni-mainz.de/fb09climatology

***Response to Reviewers**

Point-by-points response

Esper et al.: Review of tree-ring based temperature reconstructions of the past millennium

Reviewer #1

In this review, the authors make the case that tree-ring records (and tree-ring based temperature reconstructions) should not be solely judged by the correspondence between tree growth and temperature, but also need to consider other aspects of the underlying tree-ring data. To that end, they present a 5-point scheme to evaluate the overall quality of the data that feeds into temperature reconstructions. Although their article deals squarely with temperature-sensitive trees, the same ranking scheme could also be applied to moisture-sensitive trees and, with some adjustment, other proxies.
–/–

I agree it would be very useful to remind the broader field of paleoclimatology of the complexity of tree-ring records, particularly given the central role these data play in proxy estimates of past temperatures. Most of my comments on this draft are intended to help the authors refine their argument for the relatively broad audience of Quaternary Science Reviews. Towards that end, the current Introduction needs the most attention. As written, it spends too much time outlining the five characteristics of tree-ring data (only four of them, really) and doesn't present adequate evidence to make the case that this assessment is needed. If the authors can provide specific instances where attempts to reconstruct temperature screened tree-ring data based solely on climate correlations, or where these other qualities were neglected, the motivation for their own assessment would be much more clear.
We re-structured the introduction and added detail on the data screening including further, recently published examples: Neukom et al. 2014 (“We use an extensive Southern Hemisphere palaeoclimate data network from more than 300 individual sites yielding 111 temperature predictors ... The predictors for the reconstructions are selected based on their local correlations with the target grid ... A proxy record is included in the predictor set if this local correlation is significant (p<0.05).”). Mann et al. 2008 (“The screening process requires a statistically significant (P < 0.10) correlation with local instrumental surface temperature data during the calibration interval.”). Stoffel et al. 2015 (“For each individual chronology, we carefully tested its sensitivity to JJA temperatures and thus excluded series with ambiguous climate response. The remaining 233 chronologies ... represent one of the most complete data sets of exclusively temperature-sensitive TRW and MXD data sets ever used in a millennium-long NH temperature reconstruction.”). Xing et al. 2015 (“If the correlation exceeds the 95% threshold estimated by the Monte Carlo method, the TR component is screened out for the final reconstruction.”). We, however, do not intend to add technical details and/or quotes from the literature as this seems unnecessary.

Additionally, I recommend the authors take more care to distinguish between tree-ring chronologies (records) and temperature reconstructions based on tree-ring data. I appreciate that is a difficult task, as the reconstructions are derived from one or more tree-ring chronologies, but the current manuscript jumps back-and-forth between these two types of information too casually.
We added “tree-ring based” in places to avoid confusion, and included a statement at the beginning of the Data and Methods section saying that „reconstructions are derived from TRCs typically by applying a linear transfer function or simple scaling”. This should also account for the title where we refer to “temperature reconstructions” rather than “tree-ring chronologies” (something we would like to keep).

The title seems a bit off the mark. To me, it implies the article will review tree-ring based temperature reconstructions, but instead of focusing on the reconstructions themselves, it actually reviews their underlying structure using a set of five criteria. I recommend a revised title that emphasizes the assessment scheme, not the subject of the assessment.
While we were not fully convinced about this recommendation, we changed the title to: “Ranking of ....” instead of “Review of ...”. We did not want to switch to something like “Review of the quality of ...” as was suggested by the second reviewer though.

As written, the Abstract describes succinctly the motivation for this assessment, but does not include statements describing what was learned by applying these criteria to a suite of temperature reconstructions. Because QSR allows longer abstracts, I recommend adding text to highlight the specific findings of these tests.
We added information on the final ranking, and mention that each TRC performs differently in different categories.

The authors write that tree-ring chronologies are often incorporated into climate reconstructions based solely on the strength of their correlation with instrumental temperature observations, but state that this criterion ignores several other important characteristics possessed by tree-ring chronologies. Because 'climate signal' is the fifth of those characteristics, the authors need to make more clearly (both in the Abstract and in the main text), how 'strength of correlation' is different than 'climate signal'. I think the language here needs to be revised so readers don't wonder how focusing on climate correlation can lead to ignoring the climate signal.
We added a brief statement in the Abstract and edited the text, though believe that readers understand that correlation is just one (of several) metrics of the climate signal characteristic.

Page 3, line 11: Please clarify how the significance of tree-ring data increases back in time. I think that point will not be obvious to readers, who might assume the opposite given that old trees are less common than young ones.
Done. The reference also provides context here.

Page 3, line 17: Oldlist only describes the very individual trees within a species, so it can't be cited to provide an indication of the number of 1000-year old trees around the world.
Why not? As there is no better reference, and the site is highly informative, we prefer keeping this information in the paper.

Page 3, line 25: Why is Salzer et al (2014b) cited to support a comment about blended chronologies from the Alps?
Replaced with Neuwirth et al. 2004.

Page 3, line 27: Can you provide a counterfactual for the case of the subfossil trees recovered from a lake? I'm not sure most readers will follow the reasoning as to why such trees ought to be combined with trees from the same area.
Changed to: "...growing around the lake, as opposed to drier inland locations".

Page 4, line 14: In this section, the authors lay out the main rationale for their assessment; specifically, that the construction of large-scale paleoclimate networks mainly (or exclusively) choose to include or exclude individual tree-ring records based solely on their correspondence with instrumental weather records. Because this issue provides the core motivation for the following work, the authors need to expand this section (to make that argument more clear) and (more importantly) provide specific examples where instrumental calibration is the only criterion used to build the proxy network.
See above. We re-structured the introduction and added detail on the data screening including further, recently published examples.

Page 4, line 23. What is the basis for the statement that the between-tree correlation within one site (or many) is "rarely stable" or exhibits "gradual trends"? I'm not aware of any systematic review of this property in tree-ring records, or even any common assumptions about its behavior through time. It is an interesting question though!
There is no systematic review of this property, i.e. to our knowledge, we here provide the first such overview (Figure 4). The changes are particularly obvious in some of the records from Europe that

include dendroarchaeological data, but the circumstances are actually outlined on page 4 and the changing Rbar is then detailed and considered in the Growth Coherence characteristic (2.2.3). "The inter-series correlation is rarely stable and can change at (i) the transition from living trees to series from historical/remnant/sub-fossil material, or (ii) from a cluster of measurement series of a certain building to another building, or (iii) by the proportion of juvenile, mature, and adult growth rings (Cook and Kairiukstis 1990)."

Page 5, line 2: I think it would be fair to acknowledge Ault et al. (2014) here, as the agreement across frequencies between tree rings and weather data appears to depend in part on geography and the choice of climate parameter.
Reference included.

Page 5, line 14: I agree RCS often gives the best chance to preserve low-frequency variability, but it might be useful to take this point further, and emphasize that most tree-ring chronologies do not include enough samples or have a equal distribution through time, and so RCS should not be applied in most cases.
But this is exactly what we are stating in the manuscript (?!): "However, RCS demands a large number of TRW (MXD) measurement series and requires the underlying data to represent a combination of short segments (trees) distributed more or less evenly throughout the entire chronology (Esper et al. 2003a)." We even go into more detail and state: "For example, if a TRC is composed of only very old living trees, the chronology's biological age will steadily increase towards the present. This causes the biologically younger rings to be concentrated at the beginning of the past millennium and the older rings in the modern period.", and later use exactly these criteria to evaluate TRCs. So, nothing changed here.

Page 5, line 16: Here the authors list four of their five criteria to assess tree-ring temperature reconstructions (but do not mention the fifth; 'climate signal'). I would like to see the language tweaked here, so that the terms used to describe these criteria match those listed in the Abstract. And although here the authors cite the Fifth Assessment Report and the recent PAGES compilation, I'm not sure I agree that the reconstructions incorporated into each of these sets didn't consider any of the criteria listed by the authors (especially the PAGES synthesis, which includes several tree-ring based temperature reconstructions developed by authors of this paper). I agree these criteria are absolutely important and examining the quality of tree-ring records beyond their correlation with climate is crucial. But if the rationale for this review is that these other qualities are being ignored by our field, the authors need to provide specific examples of studies that have neglected these issues. And overall, the Introduction would be more effective if it spent more time discussing the motivation or need for the assessment, and less space dedicated to defining these five qualities.
We adjusted the wording in the Abstract and Introduction to avoid confusion, and added references to studies where proxies were selected based on calibration (see above). We didn't remove the Pages and IPCC reference though, as these are important publications that include reconstructions based on calibration screening. We do not state that "each" reconstruction did not consider "any" of the here listed characteristics, but rather that "they are not usually recognized" and "rarely, if ever, considered". Section extended and partly re-written.

Page 5, line 28: In this section, the authors use 'reconstructions' and TRC (tree-ring chronologies; also 'records') interchangeably, which is a problem since the former is produced from the latter. These descriptions need to be revised to readers can more clearly separate the products from the materials.
See above. We added "tree-ring based" in places to avoid confusion, and included a statement at the beginning of the Data and Methods section saying that „reconstructions are derived from TRCs typically by applying a linear transfer function or simple scaling". This should also account for the title where we refer to "temperature reconstructions" rather than "tree-ring chronologies" (something we would like to keep).

Page 6, line 28: But don't the temperature reconstructions also have different ranges because they are aiming at different targets? The variability of a global temperature reconstruction will be much different than a reconstruction attempting to replicate temperatures at a single station.
For the regional reconstructions, the targets are effectively the same: regional temperatures. This would of course be different, if regional and large-scale reconstructions are compared (with the latter containing substantially less variance), but this is not was is done here. We added "regional" to the first sentence of this paragraph to make the point clearer: "... have either been calibrated against regional instrumental climate data..."

Page 7, line 1: Yes, the correspondence is encouraging, but I'm not sure a coral person would be astonished by a regional correlation of 0.4. This discussion of continental composites seems to be overreaching a bit here.
Sure not. The coral person, as anybody else, is not expecting an inter-continental correlation of 1. The values reported here are higher than the ones obtained from the continental Pages 2K reconstructions, for example, and we already started working on a paper exploring this difference between multi-proxy and tree-ring only reconstructions at the continental scale. This is not about regional, but inter-continental correlations.

Page 7, line 27: Missing 'in'.
Corrected.

Page 7, line 33: Revise the sentence to remove improper usage of 'i.e.', better to state only that data homogeneity is based on a combination of qualitative traits.
Done.

Page 8, line 1: There are only a few possible subfossil sources, so better to list them than use 'et cetera'.
Not done.

Page 8. line 22: I think 'exemplar', rather than 'exemplary'.
Done.

Page 9, line 3: Here the difficulty to distinguish between reconstructions and the underlying tree-ring data is particularly important. It's only possible to calculate the inter-series correlation on sets of individual tree-ring records from the same location. But a reconstruction (which is a temperature estimate) can't itself possess a 'between-tree correlation'; that value is only relevant to the tree-ring data which are used to produce the reconstruction. So I think the authors need to explain the difference between these two types of information more explicitly, and avoid situations where the properties of one type is ascribed to the other.
Changed to "TRCs". See above, statements on TRCs and derived reconstructions included.

Having said that, Figure 4 is pretty neat (but may belong in the later 'Results' section). I'm not sure I see the same trends towards weaker agreement back in time, but the common signal is clearly weaker in the small set of South American records. Do the authors think that's a meaningful difference, or just a product of small sample size? And I'm a little surprised the between-tree agreement in all four regions is so low.
The lower Rbar values of the SH records might result from a combination of factors, though is seems not constructive to speculate about potential reasons. Otherwise, good to see people being surprised by the data.

Page 9, line 18: Worth noting that RCS also requires substantial numbers of samples? I know it might not need to be said, but I see too many instances where RCS is applied to a dozen trees (that germinated in the 1800s)…

Fair enough. So, stated again here.

Page 12, line 1: In this section, the authors highlight those tree-ring records that exhibit the strongest agreement between neighboring trees. Can you provide the RBAR values for these top cases, and compare them with any standard for this metric? Looking at the Oroko Swamp record, I wonder whether that case illustrates the point that, because it's so difficult to develop millennial-length records from temperature sensitive trees (especially in the Southern Hemisphere), our community is required to set a lower standard for this characteristic than would be acceptable for drought reconstruction, for example. I think that difference could be interesting to QSR readers.
This is not fully clear as we do provide information on these particular Rbar values ("... do not fall below Rbar=0.20 at any time over the past millennium ..."). Otherwise, there is no standard for this metric, as this is obviously difficult to establish, and certainly case-sensitive. We also do not intend to add all the Rbar values to the Tables, also because this might open up a discussion between dendrochronologists on how this should be calculated (e.g. among all trees, among all radii, between trees and the mean, using different windows, etc.). This discussion, while important, should take place in a more targeted journal (Tree-Ring Research, Dendrochronologia).

Page 12, line 31: Capitalize 'table'.
Done.

Page 12, line 35: I worry this statement could be interpreted to mean that these six reconstructions are particularly 'good' at capturing decadal variability, when the authors intend to say they are unsuitable to estimate temperature variability at timescales longer than decades. Also, I'm not sure what is meant by the parenthetical reference to Mann et al. (2008) - I think that point needs to be made more clear or the citation removed.
Not changed, as this is exactly what we want to point out here. The sentence reads: "These records, as well as some of the individually detrended TRCs, should not be used with the objective of reconstructing the full spectrum of temperature variance over the past millennium (e.g. Mann et al. 2008)." We are not saying that these records are particularly good at capturing higher frequency variability, and we do not intend to include such a statement, as this seems unjustified. Also, Mann et al. (2008) included exactly such records ("individually detrended TRCs"), so we reference this work here. It seems important that these limitations are better considered in future reconstruction attempts.

Page 13, line 37. I'm not sure I agree with this example. The authors argue that the Alps (larch) record might (essentially) fool people into including it as a predictor in a millennial-length reconstruction because its strong correlation with local climate might lead researchers to overlook its poor replication earlier on. But this weakness (limited coverage during the earlier period) would be address by standard approaches to cut tree-ring series at the point where small sample depth and noise serve to overwhelm the stand-wide signal. The problem described would only arise in cases where a sample depth or signal strength criterion is not used to truncate the series. Because this approach is standard procedure in dendroclimatology, I think the authors need to produce at least one or two examples where reconstructions included the entire length of tree-ring records as temperature predictors, including the early noisy sections of the record.
This seems slightly optimistic. Which "standard approach" is referred to here? The by far most commonly used approach is to truncate at some minimum replication, mostly 3 or 5 series (radii or trees!). Sure there are many examples where EPS is used as a criterion. However, this statistic is fairly unknown to non-dendrochronologists, and an EPS value of 0.8 can be reached with only a few trees (radii) if the Rbar value is high. This being said, we would like to keep this extreme example, also because it was published by some of the co-authors of this current contribution, and it feels easier (and perhaps more convincing) to criticize our own work rather than pointing to colleagues here.

Page 14: I agree wholeheartedly with the sentiment behind these recommendations, but as written, they are somewhat informal, rely heavily on acronyms defined elsewhere in the paper, and are terse and somewhat imprecise ('et cetera' again). Because these statements form the core 'takeaway' of the article, they should be revised to be more complete and more specific.
Changed accordingly.

Page 15, line 27: The reference to 'seasonality' appears somewhat suddenly, so I think this subject needs to be defined here and propped up with a few citations. When reading this section, it seems the authors are almost making the case for a better understanding of tree physiology and forest ecology but instead couch their discussion about tree-ring 'characteristics'. I'm also not sure how the body of the paper (evaluating chronologies used for temperature reconstruction) connects with the last statement, which describes the authors belief that we need more tree-ring records.
Section removed.

Page 16, line 12: This statement, which directs me to an earlier part of the paper, makes for a weaker conclusion. Better to summarize or present again the recommendations. And I'm not sure how this article will be updated (will other articles be published, or this one revised)? If the authors intend for this synthesis to be followed with subsequent updates online, they should say that directly.
Not changed. The article will be accompanied by a website where new reconstructions will be added (see last sentence of the Abstract and Conclusions.). We are currently working on this website, and one of the co-authors (Oliver Konter, now permanently employed that the University in Mainz) agreed to maintain the website.

Page 16, line 32: Change the capitalization on 'BY'.
Done.

Reviewer #2
This paper is very timely, not only because of the increasing number of proxy-climate reconstructions being produced but because other fields are reviewing the quality of their own findings (e.g. testates; Payne et al., 2016). Connected to this, I feel the title is slightly misleading - it is not a review in the broader sense of what the 39 series have achieved, contributed or compared to other proxies, but rather it is focused on an assessment of their quality. So I would suggest amending the title to something like: "Review of the quality of tree-ring based temperature reconstructions spanning the last millennium".
Title changed to "Ranking of…". See above, response to reviewer #1.

Suggest adding key words: quality control (please note, I have not suggested quality assurance i.e. prevention; rather I suggested quality control i.e. detection).
Not done, as seems not applicable to the content of this contribution.

I believe the manuscript would be strengthened by the inclusion of some discussion about quality control and quality assurance of the temperature reconstructions. The differentiation into the two types of quality management might help with the wider adoption of their criteria.

Quality Assurance (QA)	Quality Control (QC)
The process of managing for quality	Used to verify the quality of the output
A strategy of prevention	A strategy of detection
* Site selection	* Homogeneity
* Proxy (i.e. species) selection	* Replication
* Climate data selection	* Growth coherence
* …	* Chronology development
	* Climate signal

This seems to be beyond the scope of the paper.

This then leads on to the question - should all the different categories (or quality control measures) be weighted equally? How volatile/stable is the final ranking? Or put another way, are some of the categories really telling us the same thing - was one category more discriminatory than the others? Would seem to me that a discrimination index could be explored given that the authors have a matrix of 5 categories for 39 temperature reconstructions.
This is clearly addressed in the paper. See first paragraph on page 15 of the submitted manuscript: "We acknowledge that some of the metrics presented here contain partly redundant information, e.g. ...

I like the concept of this paper and feel the six key recommendations offer important advice. However I am slightly disappointed that given the eminent list of authors there is only an intuitive ranking system applied. Are there plans for the development of more validated metrics such as the suggestion of minimum thresholds - like that done with the 0.85 EPS and SSS thresholds by Wigley et al., 1984.
No. Some of these metrics are mentioned in the manuscript, but we also clearly state that the methods are inapplicable to other proxies and that we particularly attempt to bridge the gap between communities (page 14). Note that the 0.85 EPS "threshold" has no statistical basis, and has not been recommended as such by Wigley et al. 1984. It just became fashionable to use this "threshold" after it was considered in the late 80s and early 90s in some papers.

In relation to the more widespread adoption of the quality control criteria, I would like to see some specific discussion about how this might be integrated into data archives such as the International Tree-ring Data Bank (ITRDB) or the PAGES 2K network.
This is beyond the scope of this paper (also beyond the control of the authors).

Just a final more minor thing, most if not all the suggested categories were discussed by Boswijk et al. (2014) in relation to New Zealand kauri (Agathis australis) and should be included in this review.
Boswijk G, Fowler AM, Palmer JG, Fenwick P, Hogg A, Lorrey A, Wunder J (2014) The late Holocene kauri chronology: assessing the potential of a 4500-year record for palaeoclimate reconstruction. Quaternary Science Reviews 90, 128-142. doi:10.1016/j.quascirev.2014.02.022.
Payne RJ, Babeshko KV, van Bellen S, Blackford JJ, Booth RK, Charman DJ, Ellershaw MR, Gilbert D, Hughes PDM, Jassey VEJ, Lamentowicz Ł, Lamentowicz M, Malysheva EA, Mauquoy D, Mazei Y, Mitchell EAD, Swindles GT, Tsyganov AN, Turner TE, Telford RJ (2016) Significance testing testate amoeba watertable reconstructions. Quaternary Science Reviews 1-5. doi:10.1016/j.quascirev.2016.01.030
We included Boswijk et al. (2014).