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

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

[a] Department of Geography, Johannes Gutenberg University, 55099 Mainz, Germany
[b] Department of Physical Geography and Quaternary Geology, 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] Centre for International Development and Environmental Research, Justus Liebig University Giessen, 35390 Giessen, Germany
[h] Università degli Studi di Padova, Dipartimento Territorio e Sistemi AgroForestali, 35020 Legnaro, Italia
[i] Tree Ring Laboratory, Lamont-Doherty Earth Observatory, Palisades, NY 10964, USA
[j] Department of Environmental Science, William Paterson University, Wayne, NJ 07470, USA
[k] Institute of Forest SB RAS, Akademgorodok, Krasnoyarsk, 660036, Russia
[l] Laboratory of Ecosystem Biogeochemistry, Siberian Federal University, Krasnoyarsk, 660041, Russia
[m] Institute for the Humanities, Siberian Federal University, Krasnoyarsk, 660041, Russia
[n] Natural Resources Institute Finland (Luke), Rovaniemi Unit, Rovaniemi, Finland
[o] Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland
[p] Laboratory of Tree-Ring Research, University of Arizona, Tucson, AZ 85721, USA
[q] Instituto Argentino de Nivología, Glaciología y Ciencias Ambientales, CCT-Mendoza, 5500 Mendoza, Argentina
[r] 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*. A new reconstruction-scoring scheme, based on these 5 characteristics, is designed and applied to 39 published, millennial-length temperature reconstructions from Asia, Europe, North America, and the Southern Hemisphere. By providing a more comprehensive set of criteria by which to evaluate tree-ring chronologies we hope to improve the development of large-scale temperature reconstructions spanning the past millennium. All the reconstructions and their corresponding scores are provided at www.url-here.de.

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 2015, 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, Mann et al. 2008, Schneider et al. 2015, Shi et al. 2013) and global scales (Mann and Jones 2003) 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 our mechanistic understanding of the influence of temperature on tree growth (Fritts 1976). The relative significance of tree-ring series increases back in time, as the overall number of annually resolved proxies rapidly declines towards the early centuries of the past millennium (Esper et al. 2004).

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 back over the entire past millennium if one or more individual trees is 1000 years or more in age. Such longevity, however, is rare and restricted to only a few locations (OldList at: www.rmtrr.org/oldlist.htm). Most millennium-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 buried under ground or preserved in lakes (sub-fossil samples). The successful combination of samples from living trees with historical/remnant/sub-fossil samples improves when the provenance of all samples is ecologically consistent. If not, sections of a millennium-length chronology can have different growth rates and climate signals than those sections dominated by samples from living trees (Tegel et al. 2010, Linderholm et al. 2014). 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 (Salzer et al. 2014b). Sub-fossil trees from a shallow lake are ideally combined with information from living trees surrounding the lake (Düthorn et al. 2013, 2015).

Combining samples from 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 from which position in the surrounding forests the samples were originally collected (Büntgen et al. 2006b). It is not uncommon that historical structures, particularly in alpine environments, contain recycled material of unknown origin, as a consequence of repairs and additions (Bellwald 2000, Kalbermatten and Kalbermatten 1997). Without detailed construction histories the researchers 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 combined 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 chronologies from only living trees, sampled at a single site.

Another important characteristic of millennium-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 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 TRW series is an important characteristic of a mean chronology, and is commonly computed to evaluate the temporal changes in the chronology's signal strength (Fritts 1976). It is rarely stable and can change, for example, 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 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 in 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 (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 increases 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).

The basic characteristics of TRC *Homogeneity*, *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, Pages 2k consortium 2013). Promoting these characteristics, and stimulating their consideration in addition to the classical calibration against instrumental climate data, is the main objective of this article. We review the characteristics of a total of 39 published temperature reconstructions reaching back to AD 1000 and use the results to rank the timeseries.

2. Data and Methods

2.1 Temperature reconstructions

A survey of tree-ring based temperature reconstructions, reaching back with a minimum replication of three TRW (or MXD) measurement series to AD 1000, returned 39 records (Table 1). 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, and Tasmania/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, millennium-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 millennium-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 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) resulted in vastly different reconstructed temperature ranges, varying by only a few tenths to several degrees Celsius over the past millennium (see the varying ranges of the 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 these 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 metrics of these characteristics are used in an ordinal scoring scheme that is understandable to non-specialists. In those instances when raw TRW and MXD data were publically available or contributed by the authors (raw data at: www.url-here.se), we have re-calculated the metrics of interest. When the original cross-dated measurements were not available (see last column in Table 7), we have estimated their 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*), were taken from the original articles. In the instances where no measure of calibration was 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 the ordinal scoring scheme to rank the reconstructions. To aid reconstruction comparison, results of the TRC scores are stratified into four groups: 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 highlighted 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 systematic was changed with *Data Homogeneity* (5 green, 9 light green, 16 light blue, 9 blue) to account for the larger number intermediate TRCs. The individual ranks for each characteristic (2.2.1 to 2.2.5) were finally summed into an overall score.

2.2.1 Data Homogeneity

Of the five characteristics introduced here, *Data Homogeneity* is the most descriptive, i.e. 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 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" for records composed of living plus relict (sub-fossil/remnant/historical) material, and "L" for 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 exemplary 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 ordinary 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 reconstructions (the black curves in Fig. 4), as well as a minor tendency towards reduced values back in time, particularly in some of the 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 detailed above, application of 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 living and relict trees, ideally including young and old tree-rings evenly distributed throughout the past millennium (Esper et al. 2014). TRCs that are 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 adjustment *(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-length TRCs' *Homogeneity*, *Replication*, *Growth Coherence*, *Chronology Development*, and *Climate Signal* is presented in Tables 2-6. The final ranking (Table 7), derived from the sum of these metrics, 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 *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 metrics. W-Himalaya is the only reconstruction scoring in class-D in four: *Homogeneity*, *Replication*, *Growth Coherence*, and *Chronology Development*. This low scoring, combined with the 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 *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 includes only living trees). These top-ranked records are followed by a group of nine reconstructions that were sampled from slightly less homogeneous conditions, including data at moist and dry micro-sites, from different elevations, and measured using different techniques, for example. Despite their less than ideal *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 only one, well-constrained site.

The reconstructions scoring well in *Homogeneity* are not necessarily top-ranked in *Replication* (Table 3). To appear in the top group in *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 doing 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 the skill of these timeseries 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 dramatically 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 a 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 values (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 *Homogeneity*, *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 (*Replication* metric), or the samples originate from different valleys (*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 sufficient instrumental data are available to calibrate over periods of 100 years and longer (Table 6). The three top-ranked records (Torneträsk (MXD), N-Scan, Alps (larch)) all correlate at ≥ 0.70 against instrumental temperature data, with only minor differences (<0.10) between calibration and verification periods. Other reconstructions, with calibration period correlations ≥ 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 robust 20th century climate signals, can result in an overestimation of statistical skill over the past millennium.

3.3 Ranking implications

Over the past decades a number of statistically valid methods have been developed to describe 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 the 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 *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 TRCs that emphasize decadal scale variance (scoring low on "Maximum frequency") when intending to reconstruct low frequency (centennial to millennial scale) temperature variance.

R2 Avoid overrating TRCs that average many TRW (MXD) measurement series in the 20th century, but only few series at the start.

R3 Pay attention to the sample composition of millennium-length TRCs. Do data sources change through time (different sites, buildings, valleys, etc.)?

R4 Consider TRC replication and Rbar changes when interpreting tree-ring based climate reconstructions.

R5 Differentiate between composite TRCs that integrate data from varying sources (living/remnant/historical/sub-fossil) and TRCs 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 metrics. 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), potentially indicating 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 millennium-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 elevational 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.

Finally, it is important to acknowledge the seasonality of the climate signal (typically related to the regional growing season, Table 1) and question the use of a seasonally restricted reconstruction, as reported in the original literature, in attempting to reconstruct mean annual temperatures (Mann et al. 1999, 2008). While we acknowledge that the calibration/verification statistics (Table 6), as well as means by which the chronology is transformed to measures of climate are also important (e.g. 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, Rutherford et al. 2005, von Storch et al. 2004, Zorita et al. 2003), we believe that a revised consideration of basic TRC characteristics as well as the development of new TRCs is key to improving the extensive network of large-scale reconstructions and our understanding of past climate variability.

4. Conclusions

Thirty-nine millennium-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: *Homogeneity*, *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 will support 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 TRC characteristics permitted ranking the 39 millennium-length temperature reconstructions into four groups (class-A to class-D) for each of the five metrics. No reconstruction scores top in all five metrics, 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 end of a ranking table does not mean they cannot be used for climate reconstruction purposes, but indicates that the users of these data need to be aware of potential weaknesses that may inadvertently affect their experiment. This review of millennial-length TRCs will be updated as new reconstructions are produced. Updates will be published online at: www.url-here.se

Acknowledgements

We thank all the tree-ring data producers for sharing their chronologies and measurement series. JL acknowledges the German Science Foundation project "Attribution of forced and internal Chinese climate variability in the Common Era".

References

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.

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, 2015. 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.

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.

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.

Rutherford, S., Mann, M.E., Osborn, T.J., Bradley, R.S., Briffa, K.R., Hughes, M.K., Jones, P.D., 2005. Proxy-based Northern Hemisphere surface temperature reconstructions: sensitivity to methodology, predictor network, target season, and target domain. J. Clim. 18, 2308– 2329.

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., 2014a. Five millennia of paleotemperature from tree-rings in the Great Basin, USA. Clim. Dyn. 42, 1517–1526.

Salzer, M.W., Larson, E.R., Bunn, A.G., Hughes, M.K., 2014b. Climate response in near-treeline bristlecone pine. Environ. Res. Lett. 9, doi: 10.1088/1748-9326/9/11/114007.

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.

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.

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 Millennium-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 *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 *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 mean age curve point 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 including 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 the *Homogeneity, Replication, Growth Coherence, Chronology Development, and Climate Signal* scores. Last column indicates which datasets are publicly available.

Fig. 1 Location of millennium-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 integrated in 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.