Clim. Past Discuss., doi:10.5194/cp-2016-27, 2016
Manuscript under review for journal Clim. Past
Published: 23 March 2016

A universal error source in past climate estimates derived from tree rings

Juhani Rinne[1], Mikko Alestalo[1] and Jörg Franke[2,3]

[1]Finnish Meteorological Institute, P. O. Box 503, FI - 00101 Helsinki, Finland
[2]Institute of Geography, University of Bern, 3012 5 Bern, Switzerland
[3]Oeschger Centre for Climate Change Research, University of Bern, 3012 Bern, Switzerland

Correspondence to: H. J. Rinne (juhani.rinne@kolumbus.fi)

Article Review: pjk-Stockholm, SE

Introduction:

In this study (hereafter: RN()) the authors start out to prove that paucities and age-class gaps in the Torneträsk MXD chronology (Schweingruber 1988, Briffa et al., 1992) are responsible for a low frequency (red) bias. To prove this RN() have developed a method of MXD chronology construction that purportedly accounts for paucities and age-class gaps. They evidence their method's advantage by comparing a reconstruction of the historical Tornedalen temperature record (Klingbjer and Moberg, 2003) to a reconstruction produced by their method. As evidence for the low frequency bias in the RCS chronology produced by Briffa et al., 1992 (hereafter: BF92), the authors rely on a simple visual comparison.

It is clear that much of the motivation for this study relies on the theories proposed in Franke et al., 2013; hereafter: FK2013. One cannot fail to get this message from the rather brazen title and opening sentence of the Abstract (which requires references), and the last sentence (for which the study provides no evidence). Otherwise the implied goals are commendable. It would certainly be interesting to read about a new method of tree-ring climate reconstruction that does not involve transforming raw measurements into dimensionless indices, retaining the original units (e.g., Helama, 2015), and it would certainly be informative to learn what the "true color" of an MXD chronology is. However, I am not sure the method put forth in this study does any of this; in fact I am not exactly sure what this method does other than rescales averaged, variance adjusted MXD measurements. How the results presented here lead to the conclusion there is a spectral bias, *vis-à-vis* FK2013, in the 1992 Torneträsk MXD chronology (BF92), thereby corroborating FK2013, is beyond my ability to detect. There are no spectral analyses performed, no modeling of persistence, and above all no hypothesis testing with statistical rigor of any kind. There are only graphical comparisons (wiggle matches) between chronologies and reconstructions.

In this review I will argue the paper has no merit because *i*) the study is biased and provides no proof of significance, ii) the study is based on outdated data and does not contribute to advancing knowledge, *iii*) the proposed reconstruction method is does not produce a significantly different chronology, *iv*) the method does not account for inherent growth trends in MXD data, and *v*) the validation exercises are inconclusive.

I will conclude with a summary describing what I feel is the salvageable merit of this study, and a final comment on the gap-filling procedure described within.

Comments re: RN()

i) The scientific method and researcher bias

In science the null hypothesis defines a condition or relationship that an experimenter wishes to study and test. In a quest to find a difference between two conditions, A and B, the null hypothesis would be; there is NO difference. If the conditions in question have quantities that can be measured then statistical tests are used to decide whether to accept or reject the null hypothesis. The decision to reject the null hypothesis, ergo there IS a difference, is based on the probability (significance) that the observed difference cannot be explained by chance alone. In RN() the implied null hypothesis is: the BF92 MXD chronology IS different than that produced here; however, there is no evidence that the observed difference is significant.

There are libraries full of literature describing methods of signal processing and analysis of time series (e.g., Blackman and Tukey 1958, Percival and Walden 1993, Park 1992, Thomson 1982) that one can use to describe and compare the spectral properties of two time series (e.g., FK2013). Consider, in BF92 there is a plot of the power spectra of the MXD reconstruction produced using RCS (figure 9: BF92) (fig.1). Why couldn't the current authors have done the same? Or even push the concept further and computed the cross-spectral coherence between the two chronologies (http://www.spectraworks.com/web/welcome.html).

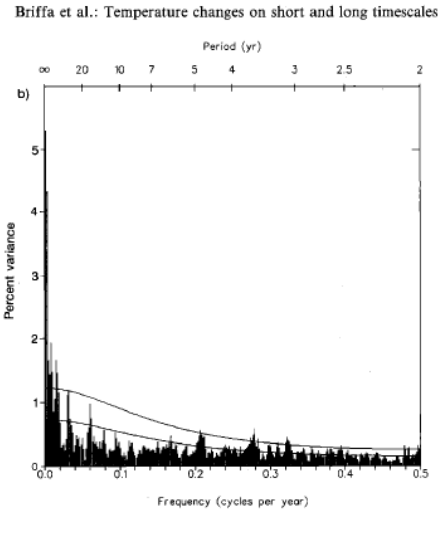

Figure 1 (BF92: fig.9b: reprinted without permission). The power spectra of the MXD reconstruction "based on the 300-lag autocovariance function and individual estimates have been smoothed with the Hamming window and have 12 degrees of freedom. The null continuum and 95% significance levels (for pre-defined peaks) are also shown".

Considering the ramifications of this study, and the overt claims of "erroneous bias", one would expect to find some statistical evidence for rejecting the unbiased, null hypothesis i.e., BF92 = RN(), but we don't. This is disconcerting to me for it means there really is no hypothesis testing and it is only the investigator's word that we must accept. Given the title and the exhausting use of *bias* and *erroneous* ("bias" is used 28 times; erroneous 10), it does not take a great deal of imagination to guess what the authors will conclude. The lack of a null hypothesis, and any attempt in applying statistical rigor to results, negates the significance of the conclusions.

ii) Why Briffa et al., 1992?

Since 1988 when the first Torneträsk MXD chronology was developed (Schweingruber et al., 1988) there has been tremendous effort and study invested in producing millennial length chronologies and reconstructions from the Scots pine trees in Fennoscandia, particularly those surrounding Lake Torneträsk, Sweden (Esper et al., 2014 and references therein). The most relevant of these is the most recent Melvin et al., 2013 (hereafter: MK2013).

With the exception of the last ~250 years when the Schweingruber et al., 1988 measurements are updated by the addition of predominantly faster growing, young living trees (Grudd 2002, 2008), the MXD data used in MK2013 are the same as those used in BF92, and this study. In other words, ignoring the post ~1650 CE period, MK2013 is essentially a re-analysis of BF92. That being the case then the real objective experiment would be to compare the chronology and reconstruction produced by RN() to that produced by MK2013. Why this was not done is again evidence to me that there is an a priori bias in RN(). So let us do it; let's compare MK2013 with RN() and decide which has more spectral bias.

Consider figure 3 in MK2013, reproduced here as figure 2. In panel b we see the two chronologies produced from the high MXD (red) and low MXD (blue) value trees, along with their average. In addition the authors have kindly provided us with information on where in the two chronologies the sample size falls below 4 (thick and thin line widths); the source of those egregious "*biases*" the present study attempts to correct. For all practical purposes the black curve in panel c (One RCS Chronology) is effectively the BF92 chronology, and the red curve is the new, improved MK2013 Torneträsk chronology.

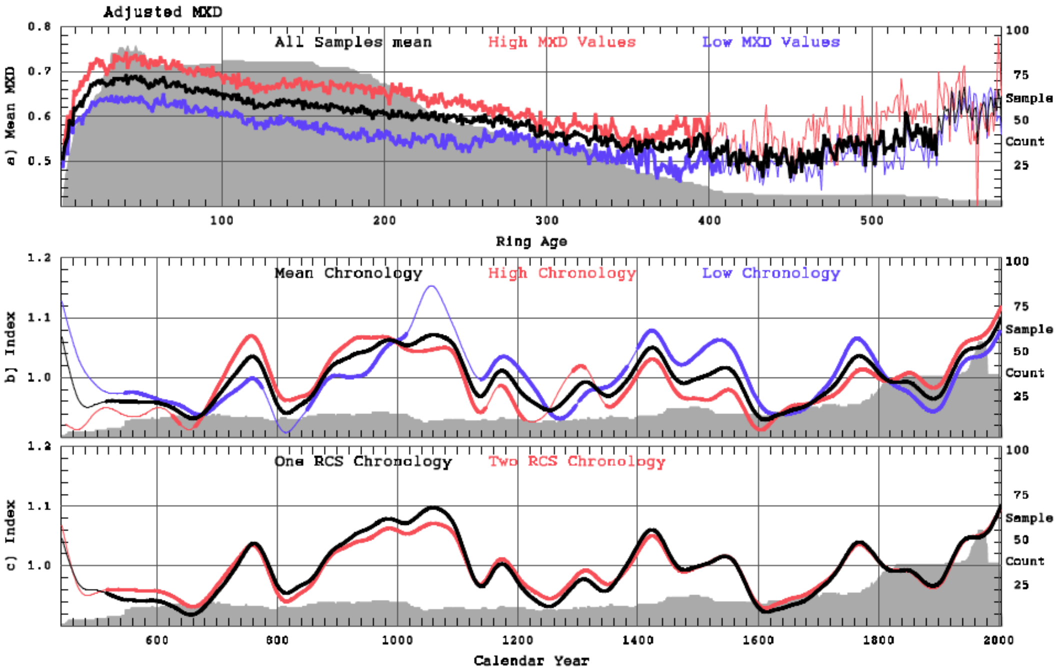

Figure 2. MK2013 main text figure 3 (reprinted without permission). "(a). The black curve is based on all samples and the curves in red and blue were built from samples with the highest and lowest values of MXD respectively, where sorting was based on comparison of mean signal-free MXD against that of a single RCS curve over their common period. b) shows mean chronologies created using two RCS curves; for high-MXD samples (red), low-MXD samples (blue), and the average of all samples (black). c) shows the chronologies created using a single RCS curve (black) and two RCS curves (red). Chronologies were low-pass filtered using a 100-year cubic spline. The thicker parts of the lines show sections of chronologies based on 4 or more samples and grey shading shows the sample counts over time." Melvin et al., 2013.

By simply comparing the red and black curves in Fig.2b one sees the new method proposed by MK2013 pays attention to temporal changes in sample depth particularly during the Medieval period where only the original Schweingruber et al., 1988 data are contributing to chronology. By visually comparing the black and red curves in Fig.2c one can imagine there is slightly less low frequency variation in red curve than the black.

iii) Where's the bias

"The reconstructions in Fig. 3a are astonishingly close to each other, in spite of the very different computational methods", RN(). So, is there a problem to fix? Let's consider the results of a comparison between BF92 chronology and the chronology produced by the method proposed here in RN() figure 3 reproduced below also as fig.3.

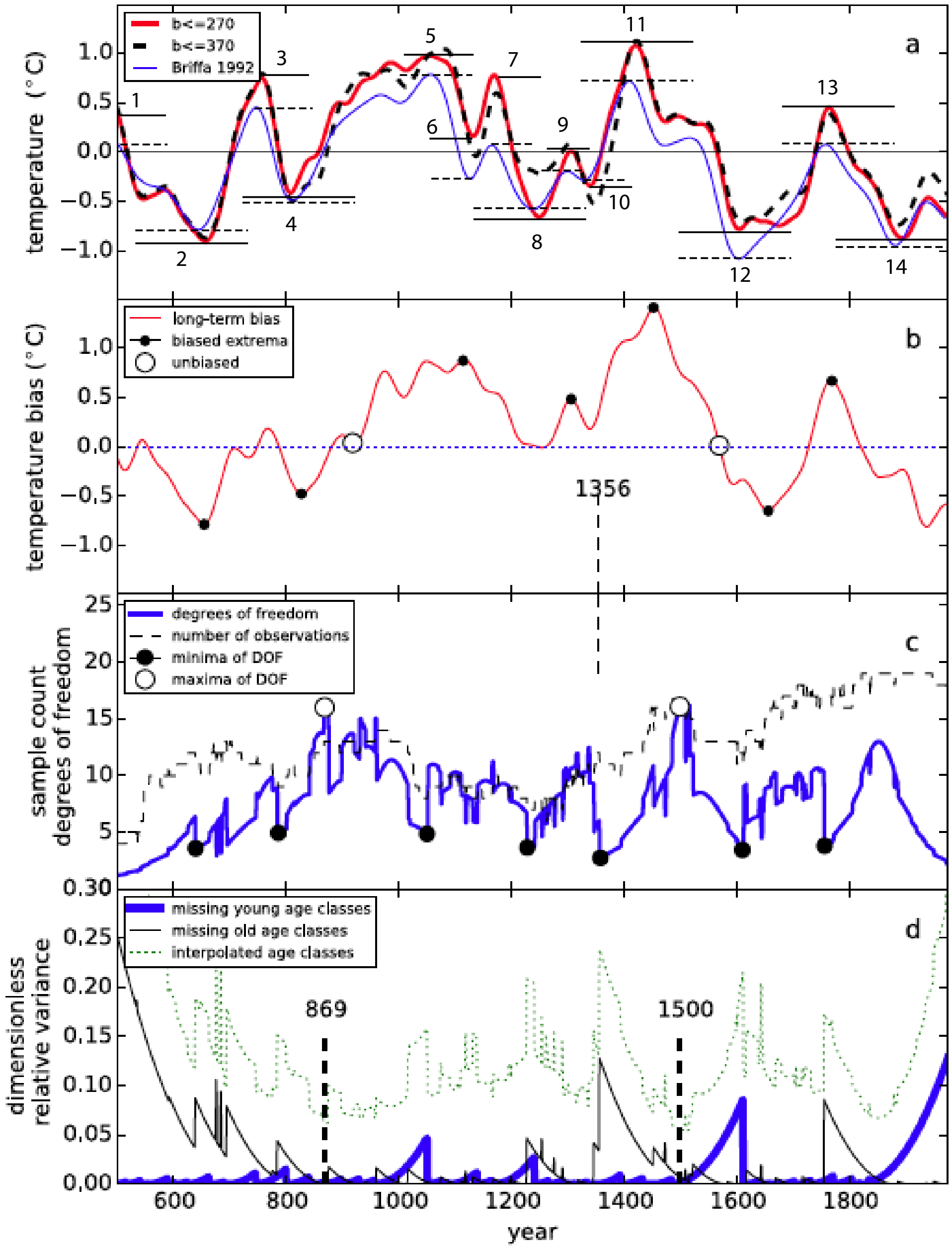

Figure 3. RN() Figure 3 with caption.

"Figure 3. The theoretical explanation of biased long-term oscillations between 500 and 1975 AD. Panel (**a**): Smoothed (85-year spline) temperature reconstructions derived from Torneträsk MXD data 441-1980. Shown are the estimate from Briffa et al. 1992 and the present ones both with b_{max}=270 and b_{max}=370. Panel (**b**): The smoothed (85-year spline) difference between the present temperature estimate (b_{max}=270) and that in Esper et al. 2012. Panel (**c**): Comparison of the sample count (number of observations) and degrees of freedom. Indicated are the seven and two cases of DOF<5 and DOF>14, respectively. Panel (**d**): components of relative variances. Some years mentioned in the text and shown in Fig. 1 are indicated in panels (**b**), (**c**) and (**d**)." RN().

In figure 3a I have marked the maximum and minimum levels of 14 major peaks and troughs in the two chronologies. For the purpose of example I used the b<=270 chronology to represent the "best" RN() chronology. It does not take a well-trained eye to see that the quasi-centennial scale variability in the b<=270 chronology is actually greater than in the BF92. For instance, between peaks 1 & 2 (fig.3a) the amplitude of the change in the b<=270 chronology is larger than in BF92, and between peaks 9 & 10, is larger still. Therefore, if there is a spectral bias in either of these two chronologies one could easily argue that it is greater in the b<=270 series.

iv) The fly in the ointment

MXD measurements have inherent growth trends and are commonly detrended by computing residuals, as opposed to ratios for tree ring-width (TRW) measurements, between the measured density and some mean biological growth function, viz. eq. 1.

$$I_t=R_t-G_t \qquad \text{eq.1}$$

where R(t) is the MXD measurement for a given year, G(t) the value of some function chosen to best model the overall trend in R, and I(t) the resulting index for year t for t=1,age. As described in BF92, G(t) is commonly a smoothed version of the mean age-aligned biological growth function. In general density measurements range from .5 to .7 density units, and in their raw form MXD measurements are much more homoscedastic then TRW measurements (Figure 4). As Figure 4 clearly shows, there is still an age-related trend in MXD measurements that is not climate related and must be accounted for. [Incidentally, please check all references. I am sure the placement of Cook and Peters 1997 is not correct; as it pertains to solely TRW data and the calculation of tree-ring indices as ratios, making it tangentially relevant to our discussion here but not where it is in the main text. I believe the more appropriate reference for the main text is Cook et al.,1995)].

The fact that MXD data have demonstrably less trend than TRW measurements is the reason RN() could simply average MXD measurements and produce a plausible chronology. I suspect that had RN() accounted for the non-climatic trend in the MXD measurements then their resulting chronology would have less “bias” and quite possibly look more like BF92.

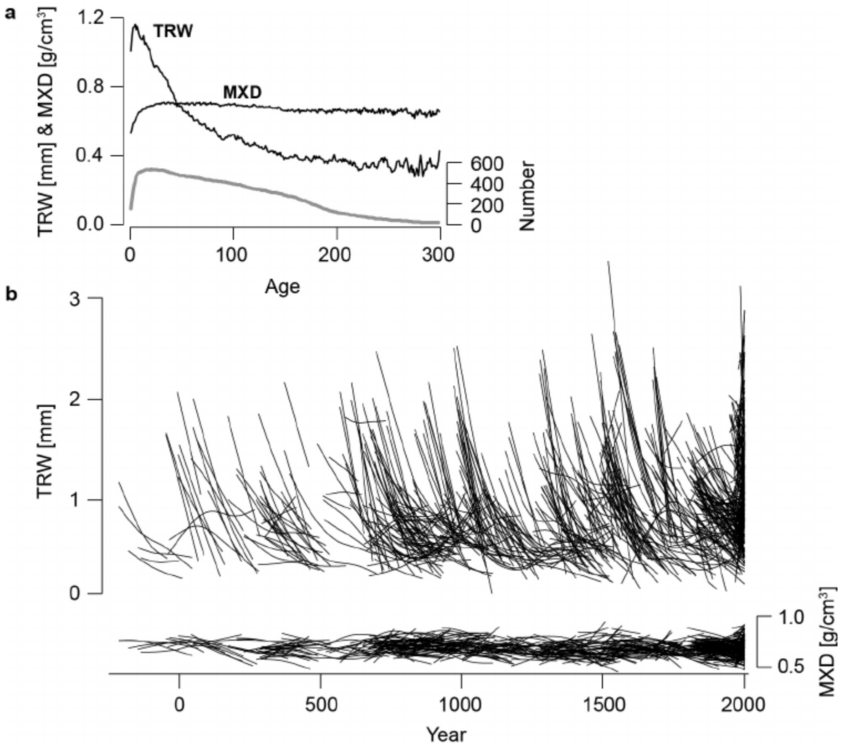

Figure 4. Growth trends in TRW and MXD measurements. Panel a, the age aligned trends: Regional Curves, panel b the calendar aligned trends (reprinted without permission from https://www.researchgate.net/figure/ 277853533_fig2_Figure-2-TRW-and-MXD-age-trends-a-Arithmetic-means-of-the-age-aligned-TRW-and-MXD)

v) Let's be fair

A perceived erroneous bias in the first multi-centennial MXD reconstruction based on RCS done 25 years ago does not demonstrate the existence of "universal error" in all RCS reconstructions. Such a statement is utterly fanciful and defamatory without proof.

The claimed illustration of bias in the Torneträsk reconstruction through comparison with the NSCAN MXD-RCS temperature reconstruction (Esper et al., 2012) in figure 3b also has problems that need to be at least admitted too. The assumption made is that the reconstruction of Esper et al. (2012) is in some sense more representative because it is based on a much larger sample size that is likely to be less affected by age-class gaps and paucities. The problem with this comparison is that the MXD data used in the Esper et al. (2012) reconstruction contains a significant amount of material from Finland, including sub-fossil lake material which the 1988 and 1992 Torneträsk chronologies do not. That is clearly indicated in Fig. 1 and Table S1 of Esper et al. (2012) and also in the main text of Esper et al. (2014).

In figure 3b, the use of completely different data, without a defendable argument for why this is okay, is not good science and undoubtedly contributes to some of the lack of agreement. In figure 3b a slightly better choice would be the N-Eur reconstruction Esper et al. (2014) as this reconstruction recognizes the affect of using trees of varying age classes.

Summary

Throughout this manuscript there were enough confusing comments and descriptions regarding previous work and dendrochronological methods that lead me go back and look at the author's affiliations. That's when it struck me as to what Rinne et al. () are trying to do! They are treating the Torneträsk MXD data as if the data were a group of meteorological records with periods of missing data, filling those gaps, and using the Tornedalen historical record to rescale the result. This sounds a lot to me like homogenizing data.

However, what strikes me as the most egregious non-scientific element in this work is found in following extract.

"Klingberj [sic] *and Moberg (2003) composited a series of instrumental observations made in the Tornedalen region, some 300 km from Torneträsk. Their construction begins with instrumental data from Övertorneå (1802-1838). The observation hours were not stated explicitly for 1826-38 and the authors had to assume them. Here we correct that assumption by increasing their JJA temperature reconstructions by 1°C. The procedure can be interpreted as if an unknown component in the instrumental observations had been deduced from the tree ring estimates. Nevertheless, temperatures before ca. 1850 may still contain biases (Melvin et al., 2013; Grudd, 2008)."*

If I understand this correctly, RN() are correcting an assumption using tree-ring estimates while at the same time claiming the tree-ring estimates are biased? This boggles my mind. Not only does this confirm my opinion on how well long-historical records can be used for the assessment of spectral bias (e.g., FK2013, Osborn and Briffa 2003), but nails the coffin shut on my opinion of the present work. It completely explains how RN() could again produce a plausible reconstruction using their method! There are so many vagaries is this story that I feel the need for a new word for bias.

I strongly disagree this work brings any new insights into the causes of spectral bias in climate reconstructions from tree rings. In fact I would argue it demeans the FK2013 argument. The appearance of long-memory in long tree-ring reconstructions is just as likely to reflect the fact that climate has varied on a variety of time scales over the past millennia (NRC 1991, Koutsoyiannis 2002) and that our extant historical records of climate are too short to model this condition. Accepting the fact that we are not likely to find any new historical records, we must do the best we can to explore the proxy record. This study does not do that at all.

The writing style and structure of the manuscript could certainly be improved. I am sure I have misunderstood a meaning or two here and there simply because I could not understand clearly what was written. The one potential merit is the interpolation/extrapolation method described in section 2.1 and illustrated in Figure 2:RN(). The question of how a non-stationary distribution of age-classes affects a chronology is one Dendrochonologists continually revisit (google any of the more recent RCS publications). I highly recommend starting with Esper et al., 2003 " *Test of the RCS method for preserving low-frequency variability in long tree-ring chronologies*". As the title implies, Esper et al., 2003 is obviously germane to the topic in hand and should have been referenced.

The questions I ask are, what do the simulated values look like, and is this step really necessary? I wonder how biologically relevant are the simulated values? This is certainly an area in dendrochronology that could benefit from further research, but more in the vein of Esper et al., 2003 rather than this.

Finally, something I just noticed, the age-class limit found optimal in this study (b_{max}=270) is remarkably close to the ideal age range used in Esper at al., 2014.

"For the final summer temperature reconstruction we only used tree rings of a certain biological age ranging from <=31 to >=306 years, i.e. removed the rings 30 years (here termed as the 'young rings') and >306 (here termed as the 'old rings') from the combined, and adjusted S88 + E12 dataset." (Esper et al., 2014)

In the quote above S88 is the Schweingruber et al., 1988 dataset, E12 is the 587 MXD collection used in Esper et al., 2012. The similarity between the two upper limits of age class restriction further convinces me the present work provides little insight in the way contemporary Dendroclimatology applies Regional Curve Standardization.

-pjk
Stockholm

References:

Blackman, R.B., and Tukey, J.,W., 1958. The measurement of power spectra from the point of view of communication engineering. Dover Publications, 190 pp.

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

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 paleoclimatic studies. Holocene 5, 229e237

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. Journal of Quaternary Science 29, 487-494

Esper J, Frank DC, Timonen M, Zorita E, Wilson RJS, Luterbacher J, Holzkämper S, Fischer N, Wagner S, Nievergelt D, Verstege A, Büntgen U., 2012. Orbital forcing of tree-ring data. Nature Climate Change 2, 862-866.

Franke, J., Frank, D., Raible, C. C., Esper, J. & Brönnimann, S. Spectral biases in tree-ring climate proxies. Nature Clim. Change 3, 360-364 (2013).

Grudd H, Briffa KR, Karlén W et al., 2002. A 7400-year tree-ring chronologyin northern Swedish Lapland: Natural climatic variability expressed on annual to millennial timescales. The Holocene 12: 657-665.

Grudd H., 2008. Torneträsk tree-ring width and density AD 500-2004: A test of climatic sensitivity and a new 1500-year reconstruction of north Fennoscandian summers. Climate Dynamics 31: 843-857.

Klingbjer, P. and A. Moberg, 2003. A composite monthly temperature record from Tornedalen in northern Sweden, 1802-2002, Int J Clim., 23, 1465-1494, 2003.

Koutsoyiannis, D., 2002. The Hurst phenomenon and fractional Gaussian noise made easy. Hydrol Sci J 47(4):573-595.

Mann, M.E. and J.M. Lees., 1996. Robust estimation of background noise andsignal detection in climatic time series, Clim. Change., 33, 409-445.

Melvin TM, 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.

National Research Council (Committee on Opportunities in the Hydrologic Sciences) 1991. Opportunities in the Hydrologic Sciences. 21. National Academy Press, Washington DC, USA.

Osborn, T. J., and K.R. Briffa, 2004. Climate: The real color of climate change? Science 306, 621_622 (2004).

Park, J., 1992. Envelope estimation for quasi-periodic geophysical signals in noise: A multitaper approach, in Statistics in the Environmental and Earth Sciences, edited by A.T. Walden and P. Guttorp, 189-219, Edward Arnold, London.

Percival, D.B., and A.T. Walden, 1993. Spectral analysis for physical applications--Multitaper and conventional univariate techniques. Cambridge University, 580 pp.

Schweingruber FH, Bartholin TS, Schar E et al., 1988. Radiodensitometric dendroclimatological conifer chronologies from Lapland (Scandinavia) and the Alps (Switzerland). Boreas 17: 559-566.

Thomson, D.J., 1982. Spectrum estimation and harmonic analysis, Proc. IEEE, 70, 1055-1096.