Global temperature changes of the last millennium

***Ronan Connolly* [1], *Michael Connolly* [1]**

[1] *Connolly Scientific Research Group. Dublin, Ireland.*

Abstract

A review of the various global (or hemispheric) millennial temperature reconstructions was carried out. Unlike previous reviews, technical analyses presented via internet blogs were considered in addition to the conventional peer-reviewed literature.

There was a remarkable consistency between all of the reconstructions in identifying three climatically distinct periods. These consisted of two relatively warm periods - the "Medieval Warm Period" (c. 800-1200 AD) and the "Current Warm Period" (c. 1900 AD on) - and a relatively cool period - the "Little Ice Age" (c. 1500-1850 AD). Disagreement seems to centre over how the two warm periods compare to each other, and exactly how cold, and continuous the cool period was.

Unfortunately, many of the assumptions behind the reconstructions have still not been adequately justified. Also, there are substantial inconsistencies between the different proxy data sources, and between proxy-based and thermometer-based estimates. Until these issues have been satisfactorily resolved, all of the current millennial temperature reconstructions should be treated with considerable caution.

Citation:
R. Connolly, and M. Connolly (2014). *Global temperature changes of the last millennium*, Open Peer Rev. J., 16 (*Clim. Sci.*), ver. 1.0.
URL: http://oprj.net/articles/climate-science/16

Version: 1.0
First submitted: January 8, 2014.
This version submitted: October 11, 2014.

1 Introduction

In recent decades, there has been considerable interest in trying to accurately quantify how globally-averaged surface temperatures have changed over the last millennium or so.

Some groups, e.g., the University of East Anglia's Climate Research Unit (CRU)[A1, A2] have attempted to estimate global surface temperature changes from thermometer records at various weather stations across the globe. Such analyses have suggested an almost continuous "global warming" trend since at least the late 19th century. However, these estimates only stretch back to the mid-to-late 19th century, as there are only a few longer thermometer records (mostly European).

In the absence of direct temperature measurements before the 19th century, researchers have attempted to estimate past temperatures using *"temperature proxies"*. A temperature proxy is any measurable occurrence or process, that temperature is a function of, and which can be dated (either exactly or approximately). Many different temperature proxies have been used, from Japanese records of the dates when cherry blossom trees bloom[A3] to changes in pollen species in lake or peat bog sediments[A4].

By calibrating temperature proxies to local thermometer records in the period for which they overlap, estimates of long-term temperature trends for the region can be made. These estimates of *local* temperature trends can then be combined with other estimates from different locations around the world to construct *hemispheric* or *global* estimates for the last millennium[A5–A31]. The term *temperature reconstruction* is often used.

The studies considered here predominantly rely on just a few types of proxy. The most frequently used have been:

1. Tree-rings (either widths or maximum density thicknesses)

2. Isotopic analyses of various depositional

*Corresponding author: ronanconnolly@yahoo.ie. Website: http://globalwarmingsolved.com

substances, e.g., speleothems (i.e., stalactites/stalagmites/etc.), ice cores and lake sediments

Some proxies have an annual resolution, e.g., tree rings, while others are less precise and can only be used for studying changes on time-scales of tens (or even hundreds) of years, e.g., sea sediment cores. The higher resolution proxies are usually favoured. But, since the proxies are generally used for estimating long-term trends, the low resolution proxies are also useful[A17, A24, A32–A34].

Early proxy studies, e.g., Lamb, 1965[A5] suggested that over the last millennium, global temperatures varied substantially on ten to hundred year time-scales. It was thought that sometime between c. 800-1200 A.D., there was a "Medieval Warm Period"[A5], while sometime between 1500-1850 A.D., there was a cold period known as the "Little Ice Age"[A35]. In this view, we have recently entered another warm period[A36], which we will call the "Current Warm Period".

In the late 1990s, a few studies suggested that the Current Warm Period was substantially warmer than the Medieval Warm Period, and that recent temperatures were unprecedented in the last millennium[A9–A12]. A 1999 study by Mann, Bradley & Hughes, which extended a 1998 study (sometimes called "MBH99"[A11] and "MBH98"[A10] respectively, after the author initials and year of the studies) was particularly striking.

The Mann et al. studies (Figure 1) suggested that global temperatures had remained fairly constant over most of the last millennium, other than a gradual cooling from the Medieval Warm Period to the Little Ice Age, but that at the start of the 20th century, temperatures had begun to rise dramatically[A11]. The study's graph of northern hemisphere temperatures of the last millennium became known as the "hockey stick graph", due to its similarity in shape to an ice hockey stick[B1], and henceforth we will refer to the Mann et al., 1998 and Mann et al., 1999 studies collectively as *"the hockey stick study"*.

This iconic image had a very powerful political and social impact as it appeared to vindicate the theory that much of the 20th century global warming suggested by the thermometer-based estimates was due to "man-made global warming". This is a theory which suggests that increasing atmospheric carbon dioxide (CO_2) concentrations from fossil fuel usage is leading to unnatural global warming.

Before the *hockey stick study*, critics of the man-made global warming theory argued that if the Medieval Warm Period had occurred naturally then there was no reason to assume the recent global warming was related to CO_2[A37, A38]. Meanwhile, many supporters of the theory agreed that much of the global warming of the Current Warm Period was "natural global warming" but argued that man-made global warming would dominate over natural trends *in the future*, if CO_2 concentrations continued to increase[A39].

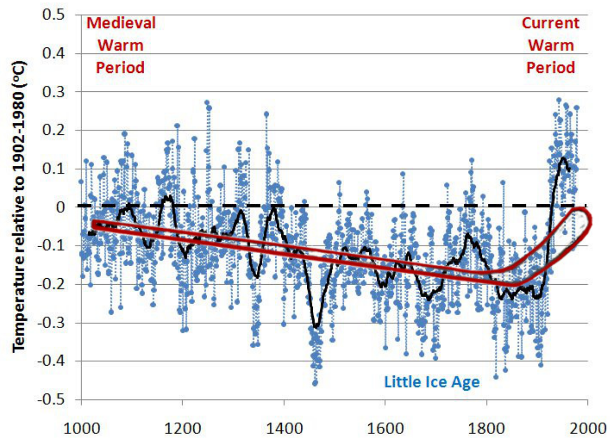

Figure 1: *The Mann et al., 1999 proxy-based estimates of the temperature trends of the last millennium, relative to the 1902-1980 mean, commonly referred to as "the hockey stick graph". Data taken from World Data Center for Paleoclimatology. Solid black line is the 31 year running mean. Red lines show a schematic outline of an ice hockey stick.*

The *hockey stick study* initially appeared to discredit both arguments as it implied that the recent global warming was unprecedented in the last millennium, and seemed to be correlated with the increases in CO_2 since the Industrial Revolution. The hockey stick graph featured prominently in both scientific reports[A40] and popular public presentations[B2], and generated considerable scientific and public concern over atmospheric CO_2 concentrations.

However, since then, a number of flaws in the *hockey stick study* have been highlighted[A37–A39, A41–A48]. In addition, many subsequent studies have suggested considerably more temperature variability over the last millennium[A13, A14, A17, A21, A24], even from the authors of the *hockey stick study*[A22].

This topic has become highly contentious. On one side of the debate, some contend that the *hockey stick* *study* is non-scientific and politically motivated[B3, B4], while on the other side, some contend that *criticism* of the *hockey stick study* is non-scientific and politically motivated[B2, B5, B6]. In this review, we will try to present the arguments from both sides.

A considerable amount of relevant analysis has occurred on "non peer-reviewed" internet weblogs (or "blogs"). However, this analysis has been overlooked in the "peer-reviewed" forums, including recent literature reviews[A34, A49, A50]. For many researchers, this may be due to a lack of awareness of the analyses, but in some cases it appears to be due to a belief that "non peer-reviewed" analyses have no relevance. This is unfortunate as the merit of an idea or argument does not depend on its source. Hence, we will consider analysis from both forums - references from peer-reviewed sources are denoted with the prefix "A", and non peer-reviewed sources with the prefix "B".

Some blogs have been critical of the *hockey stick study*, e.g., Climate Audit, The Air Vent, Bishop Hill, or Watts Up With That?. Some have defended the *hockey stick study*, e.g., Real Climate, Skeptical Science, or Open Mind. Others have tried to avoid a partisan approach, e.g., Climate Etc., Die Klimazwiebel, The Blackboard, or Collide-a-Scape.

The format of this article will be as follows: In Section 2, we review the theoretical basis and different reconstruction methods used for the current global temperature proxy estimates. In Section 3, we will discuss some of the problems involved with the proxies used in the studies. In Section 4, the specific criticisms of Mann et al.'s *hockey stick study* are reviewed. In Section 5, the different proxy-based temperature estimates are compared and contrasted with each other. Finally, in Section 6, conclusions are offered on what the current scientific information tells us and does not tell us, and how future investigation should be approached.

2 Methods used for global temperature reconstructions

The first step in generating a global (or hemispheric) temperature trend estimate from temperature proxies is to decide what proxy dataset to use.

This decision is often very subjective, and different researchers will often disagree over which proxies to include or exclude. For instance, some studies only used tree ring proxies[A13, A14, A18], while Loehle, 2007 specifically *avoided* tree ring proxies[A21]. For some studies it was important to only use proxies that have annual resolution, e.g., Shi et al., 2013[A28], while other studies intentionally included some "low frequency" proxies since a primary goal is to study long-term trends, e.g., Moberg et al., 2005[A17]. Several studies tried to ensure that all of the proxy series used covered most of the reconstruction period[A12, A17, A20, A24, A26, A28], while other studies attempted to use as many proxies as possible[A10, A11, A22, A25].

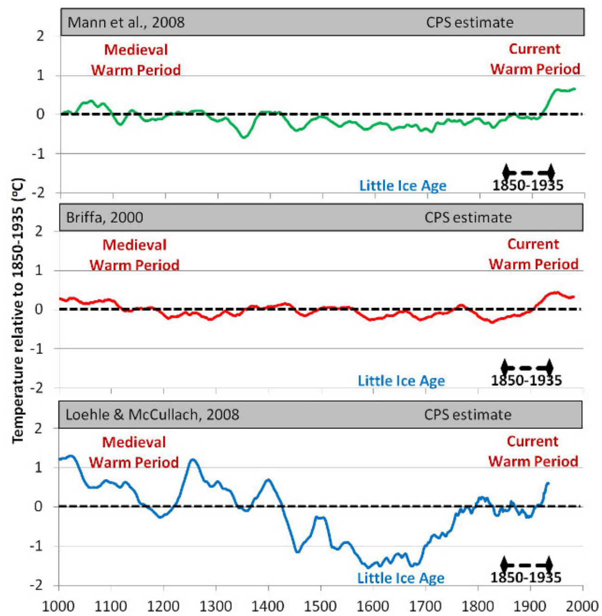

Figure 2: *Three different temperature proxy reconstructions all of which used the Composite Plus Scale ("CPS") method. All plots are rescaled and smoothed as described in Section 2.6.*

While researchers often offer valid arguments for how they constructed their proxy dataset, these decisions can have a very pronounced influence on the final results. We will discuss this in more detail in Section 3, but we can get some indication of this from Figure 2. Figure 2 shows three different proxy-based estimates each of which uses a different proxy dataset. The three estimates (Mann et al., 2008 "CPS"[A22]; Briffa, 2000[A13]; and Loehle & McCullach, 2008[A21, A51]) each suggest a different description of temperature trends of the last millennium, yet all three use essentially the same reconstruction method, i.e., "Composite Plus Scale" (CPS) - see Section 2.4.

In order to select a reasonable proxy dataset, it is important to understand the theoretical basis behind why individual proxies are thought to have a "temperature signal". So, in Sections 2.1-2.3, we will briefly summarise some of the key concepts. Specifically, in Section 2.1, we will use tree rings as a case study for illustrating why and how a specific temperature proxy can be constructed. In Section 2.2, we will provide some discussion and recommendations on how the temperature signal of an individual proxy could be tested and quantified. Then, in Section 2.3, we briefly highlight some of the problems associated with some common assumptions that have been used by most of the reconstructions discussed in this paper.

Once an appropriate proxy dataset has been compiled, a reconstruction method is needed to combine the individual proxy series into a global (or hemispheric) estimate. In Section 2.4, we briefly discuss some of the reconstruction methods which have been used for the various estimates described in this paper.

Many of these methods involved keeping/discarding (or up-weighting/down-weighting) individual proxy records on the basis of how well they correlated to the thermometer-based records in the calibration period. While this might initially seem like a reasonable way to ensure only the "best" proxies are used in the estimate, statisticians have shown in other disciplines that this "data-mining" approach actually makes the reconstructions *less* reliable[A52–A54]. In Section 2.5, we explain why, and strongly urge researchers to abandon this approach.

In this paper, we will be comparing 19 different millennial temperature estimates which are not directly comparable as originally archived. Hence, in Section 2.6, we outline various techniques, assumptions and approximations that we applied to the original estimates to allow for direct comparison.

2.1 Case study: Tree rings as temperature proxies

Like most plants, the growth of a tree depends on a number of factors: e.g., the age and species of tree; the amount of rain the area receives, i.e., soil moisture; nutrient availability; the amount of sunlight during the growing season; the amount of competition from neighbouring trees (for sunlight and/or nutrients and/or water); temperature during the growing season. Insect infestations and fires can lead to scars in tree rings.

If one of these factors is exclusively limiting the growth of a particular tree at a particular time, then it is plausible that changes in the tree rings from year to year can be used as a proxy for changes in that factor. This is the reasoning behind their use as temperature (or similarly precipitation) proxies. However, it is important to remember that, just because a tree's growth might be temperature-limited over one period, it might not always have been.

Dendroclimatologists try to maximise the temperature (or precipitation) signal by selecting trees which should, on average, be predominantly temperature-limited. For instance, the growth of trees at high latitudes (subarctic or "boreal") or at high altitudes near the tree-line ("alpine" after the European Alps), which receive adequate precipitation, and are sparsely populated, may be predominantly temperature dependent[A55–A57][B7]. On the other hand, trees growing in a drought-sensitive region may be precipitation dependent, while other trees may be limited by competition for soil nutrients.

In order to construct a useful proxy series from tree rings, dendroclimatologists extract cores from as large a selection of trees (living and/or sub-fossil) in a given area as possible. Ideally, more than one core is taken per tree, since tree growth is not always symmetric around the trunk and a core taken from one part of the tree might be different from that from another part. Different cores are then lined-up with each other ("cross-dating") and averaged together to construct a regional tree-ring time-line ("chronology"), which can then be used as a proxy series.

A major difficulty in the construction of a chronology is the problem of "standardization". As a tree ages, its growth rate may change (in general, ring growth tends to slow as a tree gets older). But, since it is changes in growth rate which are being used as the temperature proxy, it is important to remove those age-related trends.

A number of standardization techniques have been developed in an attempt to resolve this problem, but removing age-related trends, *without also* removing temperature-related trends, is a difficult challenge. So, each technique has its critics and supporters[A18, A41, A58–A62][B8–B14].

One approach which has become quite popular is "Regional Curve Standardization" (RCS) - see Esper et al., 2003 for a review[A63]. First, all the tree ring data for a specific species and region is aligned together according to the age of the tree rings (as opposed to their date). Then, an average curve is fitted to the data. This "Regional Curve" is assumed to represent the average age-related component of the tree ring growth for that species and region. Therefore, this curve is subtracted from the data for each core, and the remaining trends are assumed to be non-age-related. The standardized data for each core is then re-aligned according to date, and a chronology is constructed.

Bouldin has recently written a series of posts for his blog arguing that Regional Curve Standardization will give seriously misleading results for most of the current archived chronologies[B12–B14], although he does suggest that the problems would be substantially reduced if tree ring *areas* were analysed instead of tree ring widths.

Advocates of Regional Curve Standardization acknowledge that the assumptions in the technique are very crude, and that there are potential problems with it. However, they argue that *some* form of standardization is needed, and that it is one of the best *currently* available[A14, A18, A63–A65]. Nonetheless, it is important to be conscious of these potential problems, and treat the results cautiously.

For instance, Yang et al., 2011 have shown that the standard Regional Curve Standardization introduced a spurious positive trend in their Dulan chronology constructed from long-living junipers on the Tibetan Plateau[A66]. They suggest modifying the standardization to take into account the fact that each tree can have its own growth rate due to local growth factors[A66]. Cecile et al., 2013 also recently proposed a similar modification[B15].

Once a chronology is constructed, it can then be used to generate the desired proxy series[1]. Ideally, to create a temperature proxy, the tree ring growth should be calibrated against the local temperature records. But, sometimes, they are calibrated directly against regional (or even global) thermometer-based temperature estimates.

There are many different possible approaches which could be used for calibrating the proxies. However, for most of the proxy-based estimates reviewed here, proxy records were calibrated by simply rescaling the record so that they had the same mean value and standard deviation as the thermometer records over the calibration period, e.g., Briffa et al., 2000[A13].

In our opinion, this type of calibration is overly crude and problematic. It assumes the temperature response of the proxies is linear (which Loehle, 2009 has noted is unlikely for tree ring proxies[A67]). It also assumes that the signal-to-noise ratio of the proxy is very high, i.e., that the proxy trends during the calibration period are all climatic. Also, it does not offer an estimate of the signal-to-noise ratio of the proxy.

Instead, we recommend statistically "fitting" the proxy data to the temperature data, instead of simply "scaling" the proxy record. A typical engineering approach might be to compile a table of annual ring widths and the mean local temperature for the corresponding year, or perhaps just for the growing season. A simple model (e.g., linear or a polynomial) could then be fitted to the data for the calibration period ("training data"), and the annual ring width values of the chronology could then be converted into modelled temperatures.

There are several advantages to this engineering-style approach to proxy calibration. It does *not* assume the temperature response of the proxy is linear, e.g., if the function is non-linear and better modelled by a polynomial or some other simple function, this can be determined during calibration. Also, it does not assume the signal-to-noise ratio is very high.

Moreover, it actually provides an estimate of the signal-to-noise ratio for that proxy record, i.e., the statistical significance of the fitting function. If no obvious (statistically significant) fit can be determined for a particular proxy record, then this can be immediately recognised, and the record can be identified as unreliable.

A caveat should be mentioned on the statistical fitting approach to proxy calibration. If the fitting functions used are overly complex and/or there are very few calibration data points available, then there is a danger of "overfitting", e.g., see Refs. [A68–A70]. However, once all of the relevant statistical information is provided with the calibrated proxy record (see Simmons et al., 2011[A70]), future users of the proxy data are in a position to make an informed assessment of the reliability of the proxy record. The "scaling" approach to proxy calibration does *not* provide this information. If "overfitting" is considered a high potential risk in the proxy calibration, then Bayesian statistical inference models (with sensible priors) could offer an alternative approach that is much more robust to overfitting[B16].

At any rate, the above introduction should provide the reader with sufficient background to appreciate the basic logic behind using tree rings as temperature

[1] This process is basically the same, regardless of whether the chronology is to be a proxy for temperature, precipitation or for some other factor.

proxies. Other types of temperature proxies also have their own issues that need to be similarly considered.

2.2 Testing individual temperature proxies

We saw in the previous section that dendroclimatologists believe that tree ring growth for sparsely-populated trees at high altitudes ("alpine") and/or high latitudes ("boreal") are strongly influenced by local temperatures. Other types of temperature proxy might have a different theoretical basis. For instance, Lauritzen & Lunberg, 1999 constructed a temperature proxy record from the oxygen isotope ratios in a Norwegian speleothem using a temperature-dependent theoretical model for calcite precipitation[A71].

However, it is important to remember that a data series does not necessarily work as a temperature proxy just because a theoretical basis has been proposed for it. That is, in order to use a particular series as a genuine temperature proxy, it is essential to ensure that the theoretical basis behind it is valid. If it is not, then regardless of how well-grounded the theory might be, it might turn out to be just another example of what Huxley, 1870 referred to as *"...the great tragedy of Science - the slaying of a beautiful hypothesis by an ugly fact"*[B17].

Moreover, most "temperature proxies" are influenced by multiple factors as well as temperature. We mentioned a number of different factors which influence tree ring growth other than temperature in Section 2.1. Other proxies are similarly affected by multiple factors, e.g., see McDermott, 2004 for a review on the use of speleothems as climate proxies[A72]. For this reason, it is unrealistic to treat a proxy record as a perfect "temperature record". Instead, it is important to determine some statistical estimate of the Signal-to-Noise Ratio (SNR) of the proxy record.

A generalised method for testing and quantifying the relationship of a proposed proxy to temperature is outlined below:

1. Construct a hypothesis for the relationship of your proposed proxy to temperature. This could be done either a) on theoretical grounds or b) by analysing a sample of available data.

2. Compile a large data sample for checking the hypothesis. This data can be freshly collected specifically for the check. Alternatively, if there is already a lot of data available, a sample of previously unanalysed data could be used instead. However, if your hypothesis was formed by studying the available data, it is essential that the data you used for forming the hypothesis is *not* used for checking the hypothesis - e.g., see Anderson et al., 2001[A52].

3. Compare the temperature relationship predicted by your hypothesis to the actual relationship with the new data sample.

    - If the predicted relationship did not hold, then your hypothesis does not work. If this hypothesis has been considered by others in the scientific community, it may be important to notify them of your findings.
    - If the actual relationship was different than predicted, it might be worth modifying your hypothesis and then repeating Step 2.
    - If the predicted relationship holds and is statistically significant, it is important to estimate the approximate Signal-to-Noise Ratio for that relationship, e.g., what are the error bars associated with the fit of the data sample to the hypothesised relationship?

There are several different approaches that could be used in the actual comparisons between the proxy samples and temperature:

- Measure how the proxy responds to a range of temperature conditions. Depending on the type of proxy, one way to do this might be to obtain samples from a range of different climatic regions, e.g., Weckström et al., 2006 sampled 64 Finnish lakes to calibrate their lake sediment proxy[A73]. Alternatively, if the proxy can be mimicked under laboratory conditions, laboratory measurements could be made at different temperatures.

- Compare the proxy record with local (or regional) historical temperature records from nearby weather stations during the period of overlap. As we discuss elsewhere[B18, B19], weather station records are often affected by non-climatic biases, e.g., station moves, urbanization bias, changes in instrumentation, and the "homogenization" algorithms used to correct for these biases are often problematic[B19–B21]. So, apparent trends in the weather records should always be treated cautiously.

- Compare the proxy record to other temperature proxy records for the region, which have been previously validated.

None of these approaches are perfect, but each has their own advantages and disadvantages:

The first approach has the advantage that it does not rely on historical data, and if there is not enough data, more can be collected. However, it requires assuming that the current observed temperature relationships held throughout the proxy record.

The second approach has the advantage that the proxy behaviour can be directly compared to recorded changes in local (or regional) temperature. However, the longest weather station temperature records are only a century or two, and so these comparisons can only be carried out for a small portion of the proxy record.

The third approach has the advantage that the analysis can be carried out over most of the proxy record (depending on the length of the proxy records it is being compared to). However, it is only an indirect comparison, since the other series is only a proxy for temperature. Hence, there is a danger of circularity. If the previous validation of the comparison series was inaccurate, then it might not have a strong temperature signal. In that case, the comparison would tell us nothing about the reliability of the new series.

Ideally, all three approaches should be taken.

We should stress that the above approaches to validating/calibrating individual temperature proxies have in general **not** been taken. Indeed, as we will discuss in Section 3.3, for *many* of the commonly-used proxy series, there is no documentation of what theoretical basis (if any) the researchers had for assuming it is even a temperature proxy, let alone an estimate of its Signal-to-Noise-Ratio. Instead, researchers using these proxies typically make a number of *assumptions* about the proxy's relationship to local temperature, which as we discuss in Section 2.3 are problematic.

This should be a very serious concern for the paleoclimate community. If we are to have any reasonable confidence in our proxy-based millennial temperature estimates, it is essential that the individual proxies used each have a statistically significant relationship to local temperature. For this reason, we believe top research priorities for the community should be (a) the testing of the theoretical basis for specific types of temperature proxies and (b) the testing/calibration of individual temperature proxy series.

2.3 Problems with common proxy assumptions

Unlike thermometer measurements, temperature proxies only give indirect estimates of temperature, at best. Palaeoclimatologists hope that, by calibrating the proxies with actual thermometer records, the proxies can provide a reasonable approximation of temperature trends. However, as the thermometer records are not available outside of the calibration period, their accuracy cannot be directly tested. Furthermore, in calibrating (or training) the proxies, some of the following problematic assumptions are often made:

1. The thermometer-based data used for calibration is assumed to be accurate and reliable.
2. Proxy records which have been identified by a researcher as being a "temperature proxy" are assumed to contain a strong temperature signal.
3. Assumption of "uniformitarianism", i.e., the current relationship between local temperatures and proxy values existed for the entire proxy record.
4. The relationship between local temperatures and proxy values is assumed to be linear.

Unfortunately, all of the above assumptions are problematic:

- In a series of companion papers, we show that there are a number of serious biases which have not been adequately handled in current thermometer-based estimates[B18–B21].
- Proxy records are *supposed* to be chosen on the basis that they contain a temperature signal. However, the rationale and justification for this basis is not always given. In some cases, the researcher may merely have selected proxies which they believe are *likely* to contain *some* temperature signal. Therefore, some records which are nominally "temperature proxies" might not have **any** actual temperature signal.
- Many temperature proxies could have non-linear temperature responses[A67]. This is particularly problematic if temperatures before the calibration period are believed to have been substantially cooler or warmer than temperatures in the calibration period, as the proxy might not be adequately "trained"[A46–A48].

- If temperature is considered the "limiting factor" for a given proxy, e.g., tree ring growth, then if another factor (precipitation, sunlight, nutrients, etc.) became the limiting factor at some stage over the proxy record, the temperature relationship would have ceased.

- It is quite likely that the "noise" in the proxy record varies over time[A74], therefore the signal-to-noise ratio would similarly vary. This is of particular concern if the noise in the calibration period is substantially smaller or larger than at other stages[A75].

2.4 Deciding on reconstruction methods

Some of the early proxy-based temperature trend estimates (e.g., Jones et al., 1998[A9], Briffa, 2000[A13]) used a fairly straightforward method for estimating global (or hemispheric) temperatures from the individual proxy series. Essentially, all of the proxies were rescaled to have the same variance and mean as the thermometer-based estimates over a common calibration period. Then the mean value of all available proxies for each year was calculated. This method has come to be known as the "Composite Plus Scale" (CPS) method.

If all of the proxy records a) had a strong temperature signal; b) were evenly distributed around the world; and c) implied fairly similar trends, then in principle this method should give a reasonably accurate estimate of global temperature trends. However, unfortunately, as we will discuss in Section 3, none of these conditions applies.

For this reason, over the last 15 years or so, several groups have started using more complex multivariate statistical analysis techniques for their reconstruction methods, e.g., see Jones et al., 2009[A50] or Lee et al., 2008[A76] for a review. They hope that these complex techniques will be able to extract a more meaningful temperature signal from the noisy and inconsistent proxy data. In addition, one of the first attempts to do this (the *hockey stick study*[A10]) claimed that their method was not only able to estimate the average Northern Hemisphere temperature for a given year, but could also provide annual temperature anomaly maps showing which regions were hotter or colder than the hemispheric averages. That is, they claimed to have generated a "Climate Field Reconstruction" (CFR) method.

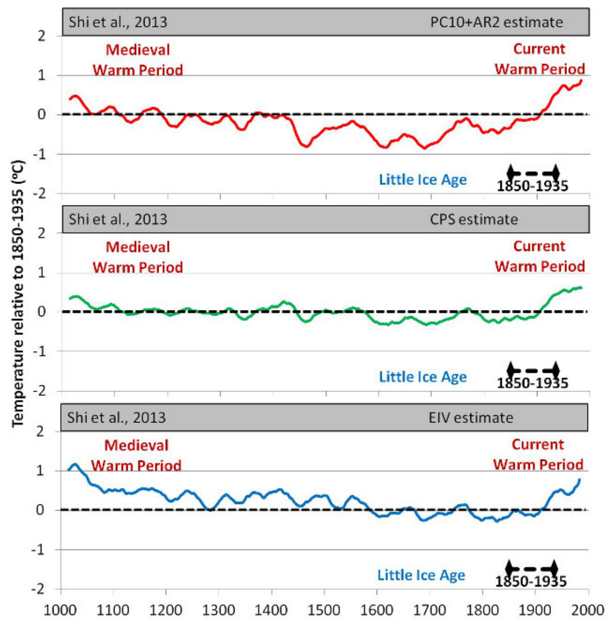

Figure 3: *The three different temperature proxy reconstructions by Shi et al., 2013[A28]. All reconstructions used the same proxy dataset, but each used a different reconstruction method. All plots are rescaled and smoothed as described in Section 2.6.*

A major problem with using complex ("sophisticated") multivariate methods for this sort of analysis is that, unless considerable caution is taken, the analysis can often yield spurious artefacts as "results" with apparently high statistical significance[A52, A77]. For instance, Rexstad et al., 1988 showed how a naïve application of several popular multivariate statistical analysis techniques to a dataset constructed from completely unrelated observations (e.g., greeting card prices, street addresses, package weights of hamburger) yielded apparently significant results[A77].

This is not to imply that multivariate statistical methods are useless. On the contrary, if the researcher knows how to analyse the appropriate diagnostics associated with their chosen method, and they analyse their results critically, these sophisticated methods can be very useful in extracting significant information from complex data[A78–A80]. They can be particularly useful during the early exploratory stages of research[A80]. However, if they are used uncritically (as is often the case), they can easily produce spurious results[A52, A79].

At any rate, with the introduction of multiple different possible reconstruction methods, the question has arisen - which (if any) of the current reconstruction methods are the most accurate, and how accurate are they? Figure 3 shows the three different Shi et al., 2013 estimates. All three estimates use the same proxy dataset. The only difference is the reconstruction method used[A28]. Yet, they each provide a different estimate of the temperature trends of the last millennium. Presumably, other methods could yield even more estimates from the same dataset.

Bürger et al. have argued that for just one proxy dataset, they could come up with at least 32 plausible reconstruction methods, each giving a slightly different result[A46–A48]. So, how do we know which ones are the most reliable and/or how accurate they are?

A problem with all proxy-based temperature estimates is that we do not know if the method of reconstruction actually works. After all, the purpose of developing such estimates is to try and figure out what past temperatures were. But, since we do not know what the past temperatures actually were, we cannot check how accurate our estimates are.

One approach to overcoming this problem has been to use computer simulations of temperature changes of the last millennium (for instance). Of course, we do not know if the simulated temperature changes are at all representative of the real temperature changes which occurred over the last millennium. But, unlike the real world, in our simulated world, we can check *with 100% accuracy* the exact *simulated* temperatures at *any* time or place during the simulation. So, if we can construct realistic mimics of our real proxies from our simulation results ("pseudoproxies"), we have at least one test of the reliability of our reconstruction *method* which we can check.

We can do this by withholding the "true" *simulated* temperature changes and then directly comparing them to our pseudoproxy reconstructed estimate. "True" is in quotes because, although we know the exact values *of the simulation*, we do not know how closely the simulation reproduces the actual temperatures.

Nonetheless, if our reconstruction method is unable to accurately approximate the mean temperature trends of the simulation (which we know exactly), then we at least know that it will not do any better for describing the temperature trends of the real world.

In other words, pseudoproxy simulations can give us a simple "validation test" for our reconstruction method. If our reconstruction method passes the test, this does *not* prove that the reconstructions are accurate. However, if the reconstruction method *fails* the test, then we know for certain that any reconstructions generated by this method are unreliable.

The use of pseudoproxy simulations for testing proxy reconstruction methods is relatively new. Its popularity seems to have arisen mainly out of interest in the *hockey stick study*. Because the *hockey stick study* purported to offer a reliable climate field reconstruction, in the early 2000s, Zorita, von Storch et al. decided to test its "MBH" reconstruction method using the results from a 1000 year Global Climate Model simulation ("ECHO-G")[A43, A81]. Zorita et al., 2003 was quite complimentary of the MBH method[A81]. However, a follow-on study – von Storch et al., 2004[A43] – was highly critical of the MBH approach, suggesting that it substantially underestimated temperature variability during the "handle" of the "hockey-stick".

As we will discuss in Section 4.2, von Storch et al., 2004's findings were hotly disputed by the authors of the *hockey stick study* and their supporters, e.g., Mann et al., 2005[A82] and Wahl et al., 2006[A83]. This led to considerable debate in the literature[A43, A46–A48, A82–A96]. Partly as a result of this debate, the use of pseudoproxy analysis has now become a quite popular test for comparing and devising new reconstruction methods, e.g., see Refs. [A76, A97–A106][B22]

As we mentioned above, if the temperature signals in the available proxies were as strong and consistent as is often implied, we would expect that *all* reconstruction methods would give essentially the same results. In that case, it would probably be sufficient to use the simpler Composite Plus Scale method. However, the fact that different methods yield different estimates (e.g., see Figure 3) indicates that this is *not* the case.

Therefore, it may be that Composite Plus Scale methods are not sophisticated enough to extract a meaningful signal from the current proxy data. However, we should remember that the increasing complexity of some reconstruction methods does *not* in itself lead to greater accuracy. Indeed, it is possible that it might introduce spurious artefacts[A77] and thereby *reduce* its accuracy.

> *"Everything should be made as simple as possible, but no simpler"* - attributed to Albert Einstein by Roger Sessions[B23]

2.5 Data mining and the spurious regression problem

As we discussed in the previous sections, merely calling a data series a "temperature proxy" does not in itself mean that it has a strong relationship with local temperature, or even that it has a temperature signal at all. In Section 2.2, we offered several suggestions as to how researchers could test whether specific types of proxies have a genuine relationship to local temperature, attempt to quantify what that relationship is, and estimate the "Signal-to-Noise Ratio" (SNR) of that relationship.

Unfortunately, such systematic approaches to statistical testing do not currently appear to be well-known within the paleoclimate field. Instead, paleoclimatologists have tended to rely on the problematic assumptions described in Section 2.3. We have already outlined some of the problems inherent in those assumptions, and we recommend that, in the future, researchers should use approaches similar to the ones we outline in Section 2.2 when constructing proxy series and/or proxy datasets.

One approach that has been frequently used by paleoclimatologists to distinguish between the "reliable" and "unreliable" proxies in their proxy dataset is to either screen or apply different weights to individual proxies on the basis of how well they correlate to the thermometer-based estimates in the overlap period. Intuitively, this might *seem* a reasonable approach, and it is widely adopted.

For instance, Esper et al., 2003, have argued[A63],

> *... this does not mean that one could not improve a chronology by reducing the number of series used if the purpose of removing samples is to enhance a desired signal. The ability to pick and choose which samples to use is an advantage unique to dendroclimatology.* - p92 of Esper et al., 2003[A63]

while D'Arrigo has apparently claimed that *"cherry-picking"* of proxy series is acceptable, if you want to *"make cherry pie"*[B24]. Hence, D'Arrigo et al., 2006 *"screened"* all of their proxies *"... by comparisons with instrumental (local and larger scale) temperature data to ensure that the temperature signal in the final reconstructions was as strong as possible..."*[A18].

Similarly, Lee et al., 2008 have claimed that this is a desirable practice,

> *The correlation based weighting scheme has the advantage of minimizing the influence of potentially unreliable proxy series on the composite record.* - Lee et al., 2008[A76]

As a result, 10 out of the 19 proxy-based estimates reviewed in this paper used either screening or weighting to promote the influence of those proxies with the highest correlation to the thermometer-based estimates. 7 of them used screening to exclude proxies from their datasets: D'Arrigo et al., 2006[A18]; Mann et al., 2008 "CPS"[A22]; Christiansen & Ljungqvist, 2011[A26]; Christiansen & Ljungqvist, 2012[A27] and the three Shi et al., 2013 estimates[A28]. 2 of the estimates explicitly weighted the proxies in their datasets based on correlations to the thermometer-based estimates: Mann & Jones, 2003[A15, A16]; and Hegerl et al., 2007[A19]. The reconstruction method of the *hockey stick study* implicitly weighted proxies on how they correlated to the thermometer-based estimates[A10, A11].

In an attempt to overcome the shortage of available temperature proxies (see Section 3.2), Mann et al., 2008 took this approach to an extreme by intentionally relaxing their requirements over what constitutes a "temperature proxy" to increase the number of proxies in their dataset to 1209, and then discarding those proxies ($\sim 40\%$) which showed very poor correlation to the thermometer-based data[A22]. In other words, rather than selecting temperature proxies on theoretical grounds (as recommended by Frank et al., 2010[A34]), Mann et al., 2008 essentially defined their data series as "temperature proxies" based on their correlation to the thermometer-based data.

Unfortunately, however reasonable this approach might intuitively seem, as soon as you start "picking-and-choosing" between your samples, you undermine the statistical independence of your dataset, and the assumptions required for statistical inference break down[A52, A68–A70]. As a result, counter-intuitively this apparently "reasonable" approach actually *reduces* the reliability of your analysis.

It is true that we would expect a temperature proxy with a strong temperature signal to be well-correlated to the local thermometer-based record for that area - where "correlation" is often quantified in terms of the Pearson correlation coefficient, R^2 (or r^2 in some disciplines), which varies from a value of 0 for non-correlated series to 1 for perfectly correlated series. However, as Yule, 1926 pointed out, "nonsense correlations" occur surprisingly often[A107]. This phenomenon is well-known in the field of econometrics as "spurious regression", e.g., see Granger, 1974[A108]; Ferson et al., 2003[A53]; or Ventosa-Santaulària, 2009[A54], where many market analysts have learned to recognise that spurious trends in a random non-stationary series have no predictive value. Hence, a high R^2 value does *not* necessarily indicate that the proxy has a strong temperature signal.

On the other hand, we discussed in Sections 2.1-2.4 how the temperature signal of the currently available temperature proxies seems to be weak, at best. Therefore, it is quite likely that a given temperature proxy would coincidentally have a low R^2 value when compared to local thermometer records, yet still be one of the better temperature proxies.

Sorting through the proxies in a proxy dataset according to their R^2 values is a form of "data-mining", i.e., analysing large collections of data for underlying "patterns". This is a technique which has been studied in detail in several fields[A52]. In particular, in stock market analysis, the ability to accurately predict future stock prices would obviously be financially lucrative. Stock market analysts have even larger datasets to work with than paleoclimatologists have proxies in their proxy datasets. So, these techniques have been well studied in econometrics[A53, A54, A108] and data-mining has been consistently shown to lead to spurious results when *then* combined with statistical inference techniques.

By only choosing proxies with a high R^2 with the thermometer-based data, you are "training" your data to increase the apparent fit to the thermometer-based data. This may well seem desirable. However, it also leads to the danger of "over-fitting" – see Babyak, 2004[A68] or Hawkins, 2004[A69] for an overview of the problem. That is, your estimate will match quite well to temperature during the "training period", i.e., the period of overlap between the thermometer and proxy records, but it will have little (or no) "predictive power" outside of the training period:

> *If the standard instruments in the literature arise as the results of* [data-mining], *they may have no predictive power in the future. ... The spurious regression and data mining effects reinforce each other. If researchers have mined the data for regressors that produce high* [R^2 values] *in predictive regressions, the mining is more likely to uncover the spurious, persistent regressors* - Ferson et al., 2003[A53]

By the way, in the context of paleoclimate, the "predictive power" of a proxy-based estimate refers to its ability to "hind-cast" past temperatures, as opposed to forecasting future results. However, since the temperatures before the thermometer-based records are otherwise unknown, the use of the term is still analogous to its use in econometrics.

We should stress that data-mining *can* be a very useful technique during the early, "exploratory stage" of research, i.e., during hypothesis-formation[A52] - see Section 2.2. However, statistical inference is based on the assumption that your statistical samples are randomly selected. If you start sorting your samples for some criteria (e.g., correlation to thermometer records), this assumption breaks down, and the apparent statistical significance of your results can become meaningless - see Simmons et al., 2011 for an entertaining demonstration of this[A70].

Therefore, we recommend that the practice of "sorting" proxies in a proxy dataset by correlation to thermometer records to *"minimiz*[e]*the influence of potentially unreliable proxy series"*[A76] be abandoned.

2.6 Techniques used in this article for comparing estimates

Table 1 lists all 19 proxy-based global or hemispheric temperature estimates for the last millennium which had been published at the time of writing. However, as archived, many of these estimates are not directly comparable. For this reason, we have applied various analytical techniques to the data before comparison.

There are a number of reasons why the original estimates are not directly comparable, e.g., the different estimates (i) cover different regions of the globe, (ii) have different standard deviations, (iii) have been scaled to different mean values and (iv) cover different time periods - see Table 1. To deal with these problems, we will take the following crude approaches, but offer some justifications and caveats:

- We will consider the various "northern hemisphere" and "extra-tropical northern hemisphere" ("NH" and "ext-NH" respectively in Table 1) and "global" estimates to all be equivalent. "Arctic" estimates, such as Kaufman et al., 2009[A23] or Hanhijärvi et al., 2013[A30] are not considered here, although they are similar.

At first glance, a "global temperature estimate" might seem quite different from an "extra-tropical northern hemisphere temperature estimate". However, there is actually a considerable overlap between the proxies used in the various studies.

Proxy-based millennial temperature estimates:

Model name †	Period covered	Region covered	Season	1850-1935 mean	1850-1935 σ
Jones et al., 1998[A9, A109]	1000-1991	NH[(2)]	Summer	-0.35°C	0.23°C
Mann et al., 1999[A11]	1000-1980	NH	Annual	-0.19°C	0.19°C
Briffa, 2000[A13, A110]	1-1996	ext-NH	Summer	0.12°C	0.52°C
Crowley, 2000[A12, A111]	1000-1965	NH	Annual	-0.04°C	0.09°C
Esper et al., 2002[A14, A112, A113]	831-1992	ext-NH	Summer	1.07°C	0.07°C
Mann & Jones, 2003[A15, A16]	200-1995	NH[(2)]	Annual	-0.28°C	0.07°C
Moberg et al., 2005[A17, A114]	1-1979	NH	Annual	-0.23°C	0.15°C
D'Arrigo et al., 2006[A18] "RCS"	713-1995	ext-NH	Annual	-0.41°C	0.17°C
Hegerl et al., 2007[A19] "long"	946-1960	ext-NH	Annual	-0.16°C	0.14°C
Juckes et al., 2007[A20] "union"	1000-1980	NH	Annual	-0.11°C	0.12°C
Loehle, 2007[A21, A51]	16-**1935**	Global	Annual	-0.08°C	0.07°C
Mann et al., 2008[A22] "CPS"[(1)]	200-1995	NH[(2)]	Annual	-0.33°C	0.13°C
Ljungqvist, 2010[A24]	1-2000	ext-NH	Annual	-0.25°C	0.11°C
McShane & Wyner, 2011[A25]	1000-1998	NH	Annual	-0.34°C	0.11°C
Christiansen & Ljungqvist, 2011[A26]	1000-1975	ext-NH	Annual	-0.46°C	0.44°C
Christiansen & Ljungqvist, 2012[A27]	0-1973	ext-NH	Annual	-0.50°C	0.48°C
Shi et al., 2013[A28] "PC10+AR2"	1000-1998	NH	Annual	-0.32°C	0.09°C
Shi et al., 2013[A28] "CPS"	1000-1998	NH	Annual	-0.37°C	0.18°C
Shi et al., 2013[A28] "EIV"	1000-1998	NH	Annual	-0.34°C	0.13°C
Thermometer-based estimate:					
CRUTEM3[A1, A2]	**1850**-now	NH[(2)]	Monthly	-0.33°C	0.22°C
Central Europe[A74, A115]	1760-2007	C. Eur.	Monthly	-0.71°C	0.65°C

Table 1: *Means and standard deviations (σ) over the common period 1850-1935 of the various proxy-based millennial temperature estimates and two thermometer-based estimates. Region covered is either Northern Hemisphere ("NH"); extra-tropical Northern Hemisphere ("ext-NH") or global. Data taken from World Data Center for Paleoclimatology, except Briffa, 2000 (Climate Research Unit); Juckes et al., 2007 (Climate Audit website); Loehle, 2007 (National Council for Air and Stream Improvement); McShane & Wyner, 2011 (Article supplementary materials) and CRUTEM3 weather station-based temperature estimate taken from Climate Research Unit.*
† The updated versions of Mann & Jones, 2003[A15] (Jones & Mann, 2004[A16]) and Loehle, 2007[A21] (Loehle & McCullach, 2008[A51]) were used. But, the original version of Esper et al., 2002[A14] (as the Frank et al., 2007[A112, A113] estimates were unarchived) was used.
(1) *Mann et al., 2008[A22] did not archive their post-1850 "EIV" estimate.*
(2) *Global and/or southern hemisphere estimates were also available.*

All of the estimates have a strong contribution from the extra-tropical northern hemisphere, i.e., the region north of the tropics. In contrast, the southern hemispheric contribution is typically small, and as a result, the nomenclature is somewhat arbitrary. For instance, only 3 of the 18 proxies used in Loehle, 2007[A21]'s "global" estimate are from the southern hemisphere, while 4 of the 12 proxies used in Mann et al., 1999[A11]'s "northern hemisphere" estimate were ironically from the southern hemisphere.

- All proxy-based estimates are rescaled so that they have the same mean and standard deviation as the CRUTEM3 thermometer-based estimates in the common period of overlap (1850-1935).

We should point out that some of the 19 proxy-based estimates used different versions of the thermometer-based estimates than others for calibration. However, typically, a version of one of the Climate Research Unit's datasets was used, and the other thermometer-based estimates which have been used are quite similar. So, we simply used a recent version of the Climate Research Unit's estimate (CRUTEM3)[A1, A2].

Rescaling the proxy-based estimates to the same mean and variance allows us to directly compare them to each other. However, it also introduces statistical artefacts which can be misleading. For instance, rescaling different estimates to have the same mean over a specific period, misleadingly implies greater agreement during that period (1850-1935 in our case) and disagreement outside that period[A116].

Also, rescaling different estimates to have the same variance (standard deviation) over a specific period, can be particularly problematic if that period was one with unusually high or low variability. For example, in Figure 2 of Briffa et al., 2000[A13], it can be seen that the period which was chosen for normalisation (1601-1974) was one with unusually low variability for the Tasmania chronology.

It should also be noted that some estimates were constructed with methods which were not *"scale-invariant"*[B25]. As a result, they may lose some of their meaning by rescaling.

- For comparison purposes, estimates will be "smoothed" before plotting by using a 31-year running mean.

This is merely for visual clarity. It should be remembered that all "smoothing" processes remove information, and there is no guarantee that this information is all "noise". Sometimes, unwary researchers may be misled by the apparent clarity of smoothed data into thinking that it has a higher "signal-to-noise" ratio. This is not necessarily the case[B26]. Running means can artificially introduce apparent "trends" which may not exist.

The various proxy-based and thermometer-based estimates (the archived, the rescaled and the smoothed versions) used in this article are included in the Supplementary Information.

3 Lack of consistency

As mentioned in Section 1, a common palaeoclimatic view maintains that there have been three main climatically distinct periods over the last millennium - the Medieval Warm Period[A5], the Little Ice Age[A35] and the Current Warm Period. However, since the 1990s, a few groups have questioned this view. Bradley & Jones, 1992[A117] pointed out that researchers often disagreed over exactly when and where the Little Ice Age occurred, as well as how long it lasted and how severe it was. This raised the question that researchers may have been using confirmation bias[A70, A118] to "identify" a global Little Ice Age in their studies.

Hughes & Diaz, 1994[A119] noted similar problems for the Medieval Warm Period. They also noted a few proxy studies which did not show Medieval Warm Periods. They suggested that the Medieval Warm Period was a "regional" phenomenon confined to areas such as Europe and Greenland. Several studies have since argued that point[A120–A124]. However, a number of other studies have found evidence of a strong Medieval Warm Period in many locations across the world, suggesting that it was a global phenomenon[A21, A33, A37–A39, A51, A125][B27].

Loehle, 2007 has suggested that one reason why the dates of the Medieval Warm Period are not always consistent could be due to dating errors with the proxies[A21]. But, there are other possible explanations, e.g., the temperature "signal" of the proxies may vary over time[A67], or the proxies may show considerable "noise" due to non-temperature related changes.

One part of the controversy seems to arise out of the inconsistencies between different proxies. Sometimes inconsistencies even exist between different versions of the same proxy series.

For example, Briffa et al., 1995[A126] developed

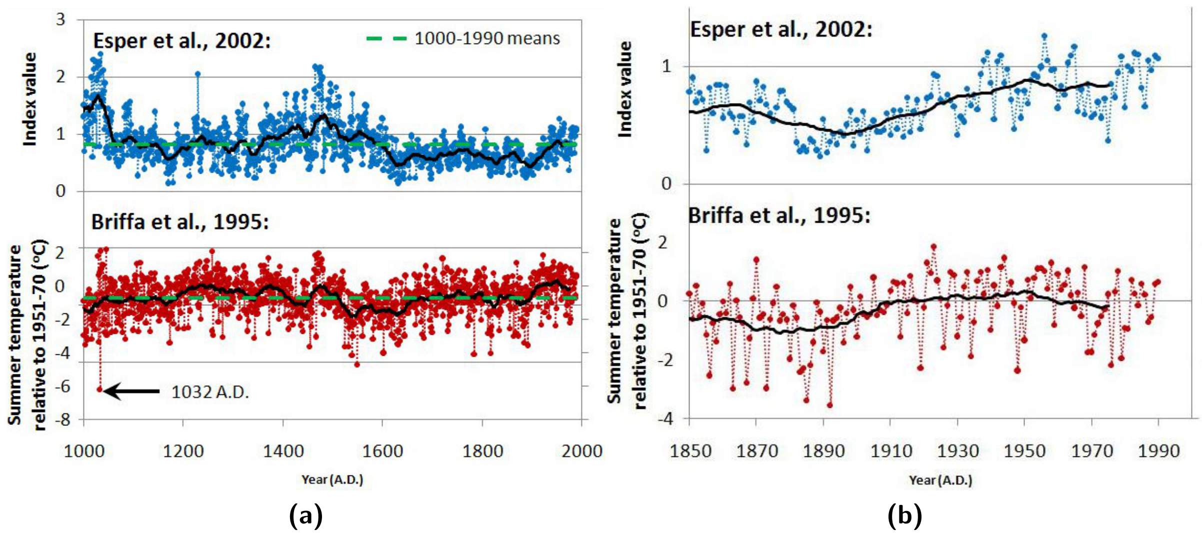

Figure 4: *Two conflicting Polar Urals chronologies. Solid black lines correspond to 31-year running means. (a) 1000-1990. (b) 1850-1990. Data for Briffa chronology taken from* http://www.climateaudit.info/data/briffa/briffa.raw.txt*. Data for Esper chronology taken from* http://www.climateaudit.info/data/esper/*.*

a Polar Urals chronology which was used in several of the early proxy-based temperature estimates[B28]. But, another version[A127] has been used by Esper et al., 2002[A14]. Both chronologies provide considerably different contexts for the Current Warm Period[B29].

The differences between the two Polar Urals chronologies are immediately apparent in Figure 4. The Briffa chronology implies a cold Medieval Warm Period and even suggests that 1032 A.D. was the coldest year of the millennium. In contrast, the Esper chronology suggests that the Medieval Warm Period was considerably warmer than the Current Warm Period. In addition, it suggests there was a second warm period from about 1400-1600 which was also warmer than the Current Warm Period.

Both of these versions show similar trends since the mid-19th century when the weather station-based estimates begin, so it is difficult to distinguish between them on this basis (Figure 4b). There do appear to be problems with how the Briffa chronology was constructed[B28, B30, B31]. However, some have argued that the Esper chronology also has problems[B32]. More recently, a third chronology from the area (the Yamal chronology) has become popular in proxy-based temperature estimates. But, as we will discuss in Section 3.4.2, this chronology suggests a different context still.

If the Briffa Polar Urals chronology is accurate, then perhaps there was no Medieval Warm Period in that area[A126]. But, if the Esper chronology is accurate, then the Medieval Warm Period was considerably warmer than the Current Warm Period in that area. Perhaps neither is accurate.

Another example is that of the Torneträsk tree ring chronology. While the original chronology used in a number of the estimates suggests a very warm Current Warm Period[B33], Grudd, 2008's updated chronology[A128] suggests the Medieval Warm Period was warmer[B34]. A third version by Melvin et al., 2012 suggests that both periods were of a similar warmth[A64].

While it is true that some proxies fail to show a Medieval Warm Period, the same could be said of the Current Warm Period. If researchers preferentially select proxies which show strong correlations with the thermometer-based data, i.e., show a warm 20th century, then this would introduce an artificial bias towards an apparently more "homogeneous" Current Warm Period, but not the Medieval Warm Period[A129].

For instance, McIntyre has pointed out that, by intentionally selecting proxy series with a pronounced "Medieval Warm Period", he was able to construct

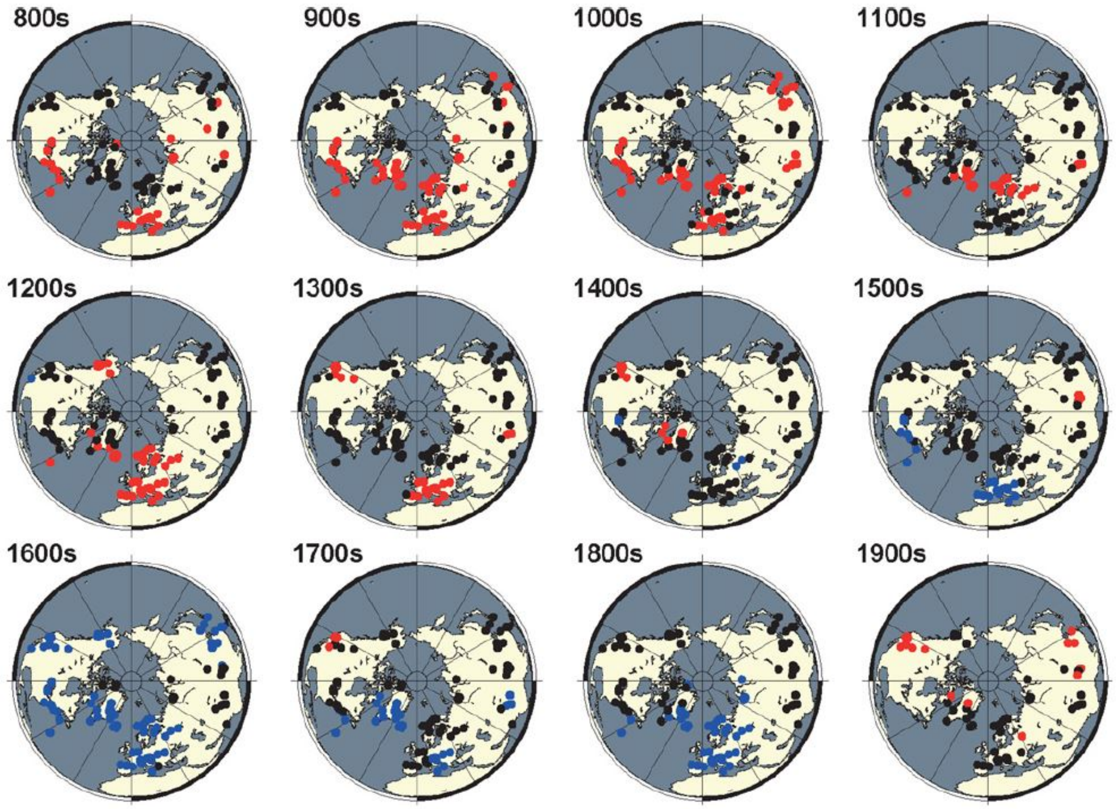

Figure 5: *Reproduction of Figure B1 in Ljungqvist et al., 2012[A31] under Creative Commons Attribution 3.0 Licence. Coloured circles indicate regions where multiple proxies are of the same sign, where red circles = warmer than average and blue circles = cooler than average. Black circles indicate regions where there is no significant agreement between proxies on the sign.*

a "reconstruction" which implied that the Medieval Warm Period was considerably warmer than the Current Warm Period[B35]. The apparent statistical correlation of his "reconstruction" to the thermometer-based estimates was comparable to some of the 19 estimates discussed in this paper (i.e., those in Table 1). The point of this exercise was *not* to claim that his "reconstruction" was "right", but rather that if a researcher was affected by confirmation bias, they could easily "find" whatever result they were expecting. He described the "methodology" for his "reconstruction" as follows,

> *Here I've picked 8 series from my files not randomly, but because I knew that they had elevated* [Medieval Warm Period] *values, scaled them and made an average (which is more or less what* [the Composite-Plus-Scale methodology] *is.) If I wanted to change the scaling properties of the series, there are proxy series with whatever noise properties that you want. This is my first run. So it is picked, but not tuned. The number of series ... in the 11th century portion of* [Jones et al., 1998[A9]] *is only 3 and* [Moberg et al., 2005[A17]] *uses only 11 series for* [their] *low-frequency portion. I could add a couple and make 11 and it wouldn't change the point.*
>
> *... I haven't tuned all the bells and whistles. For example, I haven't done a calibration-verification exercise yet. But you're starting off with something that you can tune to have a terrific RE value if it doesn't already.*
>
> *... I'm not saying that this is an alternative reconstruction of temperature. The point is that cherry pie is not only thing that you can*

make from the proxy orchard.

- Steve McIntyre, 11th March 2006[B35]

Even if the selection bias is not carried out by the compiler[A122], it may exist with the researchers who constructed the individual series. If a researcher is *expecting* to find a "Medieval Warm Period"[A119], a "Little Ice Age"[A117] or a "Current Warm Period" then confirmation bias could bias them into prematurely accepting or discarding their results.

If rigorous testing of the individual proxies is carried out along the lines of the approaches discussed in Section 2.2, this should not be a problem. However, if not, proxies giving an "unexpected" result may remain unpublished, leading to the "file-drawer problem"[A130–A132]. When the results do match with the researcher's expectations, they might be more inclined to publish them. If the reviewers and editors considering the researcher's work have similar expectations, those results are more likely to be published. Together, these processes will increase the amount of published data which apparently agrees with those expectations, and decrease the amount of published data disagreeing with those expectations. This, in turn, will reinforce those expectations among the paleoclimate community, aggravating the problem.

3.1 Importance of rigorous proxy substitution experiments

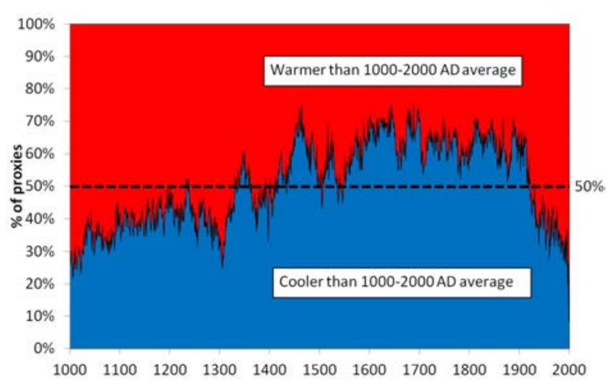

Figure 6: *Analysis of the 69 publicly archived temperature proxy records in the Ljungqvist, 2009[A32] dataset (`ftp://ftp.ncdc.noaa.gov/pub/data/paleo/contributions_by_author/ljungqvist2009/ljungqvist2009recons.txt`). Proxies are sorted for each year depending on whether they are above or below that proxy's 1000-2000 mean value.*

Ljungqvist et al., 2012 compiled a relatively large proxy dataset of Northern Hemisphere proxies with data for most of the last millennium[A31]. As part of their analysis, they looked at how similar the trends of proxies from the same region were.

Figure 5 is a reproduction of Figure B1 from Ljungqvist et al., 2012[A31]. For each map, red circles indicate that the majority of the proxies for that region had above average values for that century, i.e., were "warmer than average". Similarly, blue circles indicate that the majority of the proxies had below average values, i.e., were "cooler than average". Black circles indicate regions where there was no consistency between the proxies for that century.

From these maps, Ljungqvist et al., 2012 noted that, for those regions with coloured circles, the 800s-1300s were mostly warmer than average, as were the 1900s, but the 1500s-1800s were mostly cooler than average. They argued that this was in keeping with the expected Medieval Warm Period → Little Ice Age → Current Warm Period trends.

However, we note that for all centuries, many of the circles are actually black. That is, the proxies for many regions fundamentally disagree over whether or not a given century was warmer or cooler than average. This suggests a serious inconsistency between proxies, as we mentioned in the previous section.

We can again see this inconsistency in Figures 6 and 7 which were generated from all 69 of the publicly archived proxy series in Ljungqvist, 2009[A32]. In Figure 6, the percentages of proxies which are above/below their 1000-2000 mean are shown for each year. For most of the 20th century as well as the 11th-14th centuries, the majority of proxies are warmer than average. This is as expected for the Medieval Warm Period and Current Warm Period. But, surprisingly, for most of these years, more than 30% of the proxies are actually *cooler* than average. Similarly, for the 15th-19th centuries, when we would expect the proxies to be cooler than average (corresponding to the Little Ice Age), more than 30% of the proxies are actually warmer than average for most years.

Figure 7 also shows the same general lack of consistency between proxies. However, it provides more detail by grouping proxies, for each year, depending on *how much* warmer or cooler than average they are, i.e., whether they are $< 1, 1-2$ or ≥ 2 S.D. above or below their 1000-2000 mean.

With this in mind, when assessing the reliability of a given proxy-based estimate, it is important to

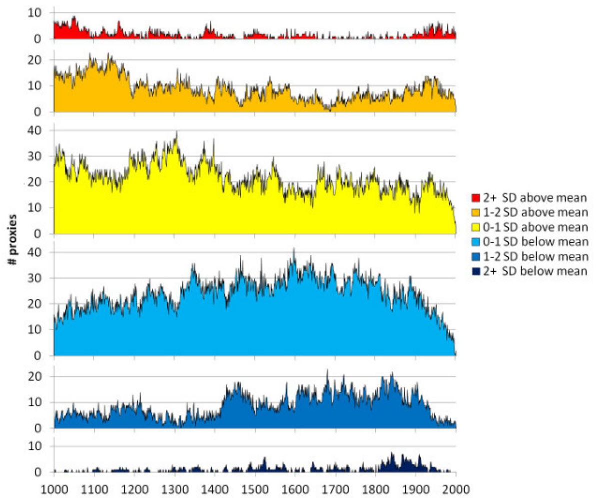

Figure 7: *Analysis of the same proxy records in Figure 6, sorted for each year depending on whether they are* < 1, < 2 *or* ≥ 2 *standard deviations (SD) above or below that proxy's 1000-2000 mean value.*

see how dependent (or "sensitive") it is to the inclusion/exclusion of individual proxy series. There are several simple "sensitivity" experiments that could be carried out for most of the estimates, e.g.,

1. Systematically remove each proxy series from the proxy dataset, one-at-a-time, and recompute the estimate using this modified dataset. Then, compare all of the recomputed estimates to the original estimate. If one or more of these recomputed estimates is noticeably different from the original estimate, then that indicates the "missing" proxy has a strong influence on the estimate, i.e., the estimate is "sensitive" to the inclusion/exclusion of that proxy. However, this test will fail to detect problems when two or more proxy series have similarly anomalous trends.

2. Systematically remove *several* proxy series at a time, or randomly generate several smaller subsets of the original dataset, and then test the results as above. This should detect problems if two or more proxy series are similarly anomalous. However, if the full proxy dataset is very small, then care should be taken that the subsets are not too small, e.g., for the Jones et al., 1998 northern hemisphere estimate, only three of the proxies used had data for the entire reconstruction period[A9].

3. Carry out "proxy substitution" experiments, by substituting one version of a proxy for another, e.g., if the original estimate included one version of the Torneträsk tree ring chronology, recompute the estimate using the other available versions[A64, A128][B33]. If any of these substitutions has a noticeable effect, it indicates that the estimate is sensitive to the inclusion/exclusion of that proxy version. If only one version of the proxy series exists, then other proxies that are similar could be substituted instead, e.g., maybe the different versions of the Polar Urals tree ring chronologies (Figure 4) could be alternately substituted for the Yamal chronology (see Section 3.4.2).

All three of the above types of experiments should be relatively straightforward to implement for most of the proxy-based estimates, and would provide a simple test of its robustness. For this reason, we recommend that future proxy-based estimates routinely carry out such sensitivity experiments as a basic check.

In cases where the second type of experiment is impractical because the proxy dataset is too small, the third experiment should provide similar information. For instance, in 2008[B36], McIntyre was able to substitute the versions of two of the three millennial proxies used by Jones et al., 1998 (Polar Urals and Torneträsk) with other published versions. These simple substitutions substantially altered the Jones et al., 1998 temperature estimates - suggesting a Medieval Warm Period considerably warmer than the Current Warm Period - the opposite of Jones et al., 1998's original conclusions[A9]. This indicated that the Jones et al., 1998 northern hemispheric estimate was highly sensitive to the choice of proxy.

However, remarkably, of the 19 estimates discussed in this review, the only studies which carried out explicit sensitivity experiments were Moberg et al., 2005[A17]; Loehle, 2007[A21]; Juckes et al., 2007[A20]; Mann et al., 2008[A22]; and Shi et al., 2013[A28].

Moberg et al., 2005 carried out the first type of sensitivity experiment for 11 of their 19 proxy series, i.e., their "low-frequency" proxy series[A17]. The results of these experiments were positive, indicating that their estimate was not overly sensitive to any one of the 11 low-frequency proxies. However, they did not carry out any other sensitivity experiments.

Loehle, 2007 carried out the first two types of sensitivity experiments[A21]. Both sets of experiments were successful, indicating a reasonable consistency between the 18 proxy series used in the original estimate. However, no substitution experiments were carried out. This may have been because Loehle had only been able to identify 18 series which met the requirements of the study (i.e., non tree-ring proxies with at least 20 dates over the last two millennia).

Juckes et al., 2007 carried out the first type of sensitivity experiment by systematically dropping each of their 13 proxies and studying its effect[A20]. The results were positive, and they concluded that their estimate was not overly affected by any *one* proxy series. However, their sensitivity experiments would have been too restrictive to detect anomalous proxies if two or more were similarly anomalous. McIntyre noted for an early draft of the study[B37] that more than one of the 13 proxies used *was* potentially problematic, e.g., two of the proxies were bristlecone/foxtail pine chronologies (see Section 3.4.1) and one was the Yamal chronology (see Section 3.4.2). So, it seems that the sensitivity experiments carried out by Juckes et al., 2007 were inadequate.

Mann et al., 2008[A22, A133, A134] claimed to have carried out *"sensitivity studies"*, and shown that their estimates were not unduly affected by any individual problematic proxy. However, as will be discussed in Section 3.4.3, the Mann et al., 2008 sensitivity studies were very poorly devised.

Mann et al., 2008 used the largest proxy dataset of all 19 estimates, with a total of 1209 proxy series. Hence, all three types of sensitivity experiment could be easily implemented. However, instead of carrying out any of the three types of experiments described above, they limited their analysis to two crude experiments. In one, they removed *all* tree ring proxies. 1035 of their proxies were tree ring proxies, so in this first experiment, they were removing more than 85% of their data - a rather extreme experiment. In their other experiment, they included all the tree ring proxies, but removed a new set of 7 other potentially problematic proxies. That is, they only removed 0.6% of their proxies - a rather minimalist experiment.

It later transpired that the Mann et al., 2008 estimates relied heavily on including *either* bristlecone/foxtails *or* another proxy series, known as the "Tiljander lake sediments", which were known to be problematic for the Current Warm Period, e.g., see Ref. [B38]. Both of these sets of proxies contained similarly anomalous "hockey stick" trends. Since the bristlecone/foxtail proxies were tree ring proxies, they were excluded by their tree ring proxy removal experiment. Similarly, the four Tiljander proxies were among the 7 other proxies which Mann et al., 2008 had flagged as potentially problematic, and so they were excluded in the second sensitivity experiment. However, neither of their experiments removed *both* sets of proxies.

As we will discuss in Section 3.4.3, a later sensitivity experiment which combined both experiments substantially altered the Mann et al., 2008 estimate, indicating that the estimate was strongly influenced by these problematic proxies. If more rigorous sensitivity experiments had been carried out, this unreliability could have been identified from the beginning.

Shi et al., 2013[A28] also carried out some sensitivity experiments, but they were again rather limited. Their proxy dataset contained 45 proxy series, 34 of which were tree ring proxies and 11 of which were not. They generated two subsets - one containing only tree ring proxies ("dendro") and the other excluding all tree ring proxies ("no-dendro").

Shi et al., 2013 claim that the estimates from both subsets and the full dataset were *"all fairly similar, with no distinct differences in the cold/warm phases of the reconstruction results at multidecadal timescales"*, i.e., they claimed the results from the experiments were positive. However, from visually examining their Figure 2, it seems to us that the relative magnitudes of the cold/warm periods are substantially different. Hence, it seems that the various Shi et al., 2013 estimates are quite sensitive to proxy selection. Moreover, they only carried out two subsetting experiments, so it is quite possible that more rigorous sensitivity experiments would identify even more problems.

Although Ljungqvist, 2010[A24] did not carry out any sensitivity experiments, Condon used the proxy dataset of Ljungqvist, 2010 to do so at the blog The Air Vent[B39]. Condon created a large number of different proxy-based estimates by randomly selecting different subsets of Ljungqvist's proxy network. All of the subsets were relatively similar to the original Ljungvist, 2010a estimate, suggesting that none of the proxy series unduly influenced that estimate.

In addition to the above sensitivity experiments, there are also other analytical techniques which could provide further insight into the inconsistency of proxies. These could be a useful supplement to the basic sensitivity tests.

For instance, in a May 2011 blog post, Eschenbach suggests how the similarities and differences between individual proxy series in a large proxy dataset can be identified by the construction of a cluster dendrogram[B40]. Eschenbach used this cluster analysis to study the 95 full-length proxy series used by the Mann et al., 2008 estimate[A22] mentioned above (and later by McShane & Wyner, 2011[A25]). He found the 95 proxy series could be grouped into 12 different clusters. He then calculated the average trends for each of these clusters and compared them visually.

Most of the cluster averages showed little long-term trends, but four of the clusters showed quite distinctive trends. One cluster ("Speleothems & Lake Sediments") implied a general "cooling" trend over the entire record, however three of the clusters implied a sharp upward "hockey-stick" shape for the Current Warm Period. One of these "hockey-stick" clusters consisted mostly of the "Graybill" bristlecone pine proxies discussed in Section 3.4.1. A second cluster comprised the four "Tiljander" proxies discussed in Section 3.4.3. The remaining cluster ("Misc. Asian Tree Rings") only contained three proxy series (*mongolia-darrigo, recjj-yy1* and *tornetrask*). In other words, Eschenbach's cluster analysis suggested that most of the full-length proxies in the Mann et al., 2008 did not show much of a trend, and hence most of the "trends" in the final Mann et al., 2008 estimate arose from only a small minority of proxy series[B40].

3.2 Shortage of long proxy records

One problem which is often unappreciated outside the palaeoclimatology community is the shortage of long temperature proxy series which are available. As a result, many of the same proxy series are re-used in different proxy-based temperature estimates[A41]. This means that different "independent" studies are not as independent as might be first thought. This was particularly problematic for early studies, e.g., Jones et al., 1998 only had 3 millennial proxies for their northern hemisphere estimate and 3 for their southern hemisphere estimate[A9]. Even today, there are still only a few long records [A32].

The *hockey stick study*[A10] attempted to reduce this problem by constructing their estimates in a stepwise manner, and thereby including large numbers of shorter proxy series (Mann et al., 2008 also adopted a similar approach[A22]). As discussed in Section 1, the *hockey stick study* consisted of two parts - Mann et al., 1998[A10] and Mann et al., 1999[A11]. The Mann et al., 1998 part of the *hockey stick study* divided up their analysis into several steps: 1820-1980, 1800-1820, 1780-1800, 1760-1780, 1750-1760, 1730-1750, 1700-1730, 1600-1700, 1500-1600, 1450-1500 and 1400-1450[A135]. The Mann et al., 1999 part of the *hockey stick study* then extended the 1400-1980 estimates with an additional 1000-1400 step[A11]. When estimating temperatures for each step, *all* the series that had data for that step were included. Hence, Mann et al., 1998 estimated 1820-1980 temperatures using the complete selection of series (159), but only used 22 series for estimating 1400-1450 temperatures[A135]. For the 1000-1400 extension[A11], they only used 12 series.

In effect, Mann et al. created several different "mini-reconstructions", each only spanning a few decades or centuries (depending on the step). These mini-reconstructions were then stitched together to create a much longer reconstruction of 600 years[A10], 1000 years[A11] or longer[A22]. Although this might initially appear a useful way of incorporating more information into the estimates, it actually leads to a *less informative* estimate. This is because *direct* comparisons are only meaningful for temperatures *within* a given step. For instance, while the temperature estimate for 1400 could be directly compared to the one for in 1450, it could not be directly compared to the one for 1460, since the 1460 temperature was estimated from a statistically different proxy network.

This approach also led to problems when proxy series in one step were replaced with different series. For example, Mann et al., 1998 had used bristlecone/foxtail tree ring proxies which were believed to be affected by non-climatic effects (Section 3.4.1). To address this concern, Mann et al., 1999 decided to apply an *ad hoc* adjustment (Section 4.1) to a series based on those proxies in their 1000-1400 extension[A11]. However, since they only applied the adjustment to the new step, the series was different when used for the 1000-1400 step than for the 1400-1980 steps. This substantially altered the apparent millennial temperature trends[B41].

A major problem with bulking up the total number of proxies in the proxy dataset by combining short proxies with long proxies, is that different proxies can imply different magnitudes of temperature variability. If a short proxy shows more variability than its longer counterparts, then averaging the two proxies together will increase the apparent temperature variability during the overlap period, but leave the earlier period unaltered.

For instance, McIntyre pointed out that one of the three northern hemisphere millennial proxies used by Jones et al., 1998[A9] showed little variability over its entire record, i.e., it did not suggest either a Medieval or Current Warm Period, or for that matter a Little Ice Age. This was the Greenland ^{18}O isotope "Crete" record - see top panel of Figure 8. Perhaps the reason for this is that the proxy was not particularly temperature sensitive, or maybe these hypothesised climate changes were not as pronounced in that part of Greenland as the rest of the world.

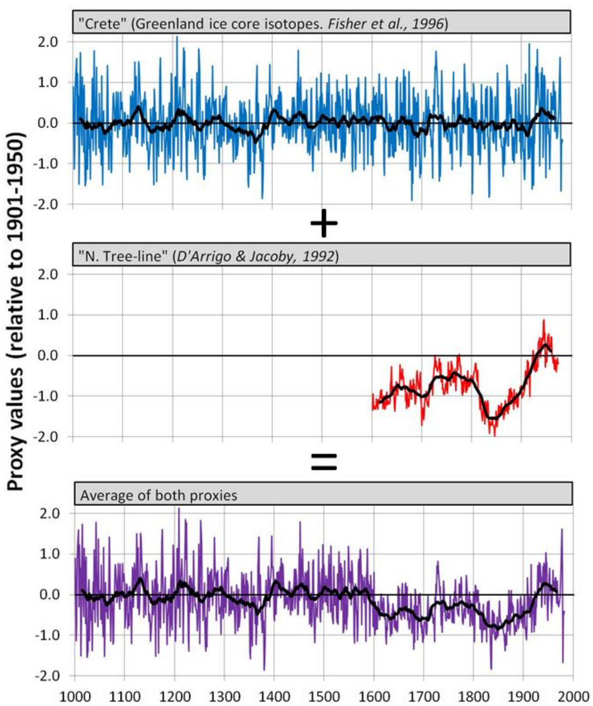

Figure 8: *Effects of averaging together a long, low variability proxy series ("Crete") and a short, high variability proxy series ("N. Tree-line"). Thick black lines correspond to the 31-year running means.*

Whatever the case, one might initially suppose that the use of the "Crete" proxy series should not alter the relative ratio of the Medieval Warm Period, Little Ice Age and Current Warm Periods in the global estimates, since they were all similarly dampened. However, Jones et al. also used several short proxies such as the Jacoby North American tree-line. This "N. Tree-line" proxy implied much more temperature variability than the "Crete" record, e.g., it implied that the Little Ice Age was *much* colder than the Current Warm Period - see middle panel of Figure 8.

If the higher variability of the "N. Tree-line" proxy is a more accurate representation of global temperature trends since 1600 AD than the low variability of the "Crete" proxy, then it is plausible that this higher variability *also* occurred over the 1000-1600 AD period. However, because the "N. Tree-line" record only covers the more recent period, when Jones et al., 1998 combined the two proxy records together, the "N. Tree-line" record made the temperature variability seem greater for the recent period, but left the pre-1600 period "flatter" - see bottom panel of Figure 8.

McIntyre noted that when a second short proxy with higher than average 20th century values ("Kameda melt"), this had the net effect of artificially making the Current Warm Period *appear* to be warmer than the Medieval Warm Period, even though neither of the short proxies were long enough to compare the Medieval and Current Warm Periods[B42].

Another problem with the currently-used proxies is that many of them are quite out of date, e.g., finishing in the 1970s or 1980s[B43]. As it is since the 1980s that temperatures are alleged to have become unprecedented due to man-made global warming[A10, A11], this is quite a serious short-coming. Mann has argued that this is because updating proxies is a costly, and labour-intensive activity[B44]. However, this seems to be an exaggeration[B43], as, in response, on a holiday visit to his sister, McIntyre (along with a reader of his blog, Holzmann, and his wife) was able to update one of the controversial Graybill bristlecone chronologies (Sections 3.4.1 & 4) critical for the *hockey stick study* and still have time for coffee[B45].

3.3 Problems of "grey data" and poor documentation

Often the authors of proxy-based studies provide little or no discussion of why they used certain proxies, why they discarded others, why they might have chosen an old version of a series in preference to more recent updates, or the basis for any adjustments/standardisations they may have applied to particular series[B46]. This means other researchers often have to do their own analysis with limited information[B47].

Also, researchers often appear surprisingly reluctant to archive the proxies and/or code they used[B48]. In the case of proxies, this is not always entirely the fault of the researchers. Unfortunately, within the dendrochronology/dendroclimatology community, there is a significant amount of "grey data"[B49]. The owners of this data do not want to make it public, but often allow researchers to *use* it, on the provision they do not pass it on. For instance, some of the proxies used by Esper et al., 2002[B50] or Moberg et al., 2005[B51] were "grey data" proxies. This obviously hampers the abilities of those researchers to archive all of their data.

Without having access to the data from which a study was derived, it can be very difficult to conclusively assess the study. Hence, unresolved discrepancies between different studies cannot be satisfactorily resolved[B52]. For this reason, perhaps it would be best if future proxy-based studies were only carried out using proxy series that the study authors are allowed to archive, i.e., no "grey data".

It is understandable that *in the past* open access to data was often unrealistic. However, with modern internet archives such as the World Data Center for Paleoclimatology, most of those arguments no longer apply. Indeed, it seems that when scientists make their data freely available, it not only helps alleviate suspicions over the validity of their research, but also encourages a better appreciation of their work[B53, B54]. Admittedly, where there are commercial applications for the data, or the research was privately funded, exceptions may be justifiable. But, this does not seem to be an issue for most of the palaeoclimate studies discussed here.

This could also help reduce the "file-drawer problem"[A130–A132] mentioned earlier. In their study of social science publications, Franco et al., 2014 found that researchers were less likely to publish the results of their experiments if they perceived the results as "null results"[A132]. Often this was *"... because they believed that null results have no publication potential even if they found the results interesting personally"*[A132]. Perhaps if the public archiving of *all* proxy results were more strongly encouraged, some researchers currently tempted to keep their results "in the file drawer" because they feel they would not "get a paper out of it" would archive their results anyway.

3.4 Over-reliance on controversial proxy records

As mentioned in Section 3.2, a surprisingly small number of long temperature proxy records have been used in multi-proxy temperature estimates. This problem is made more serious by the fact that there are known problems with some of the most widely-used proxies and composites, e.g., the Dunde ice cores[B55] and Yang's Chinese proxy composite[B56].

Considering the inconsistency between individual proxy series which we have discussed in this section, it is difficult to draw definitive conclusions from studies which rely heavily on any one series[A32]. So, it is a serious concern that almost all of the proxy-based temperature estimates rely heavily on at least one of two groups of problematic tree rings - bristlecone/foxtail pines (Section 3.4.1) or Briffa et al.'s Yamal chronology (Section 3.4.2) - see Table 2. If these groups are removed or even replaced with plausible alternatives, the relative ratio between the Medieval Warm Period and the Current Warm Period is often altered - specifically, the Medieval Warm Period becomes "warmer" and the Current Warm Period becomes "cooler"[A45][B10]. For instance, for the Shi et al., 2013 "PC10+AR2" and "CPS" estimates, their so-called "dendro" subset which includes both bristlecones and Yamal shows a much colder Medieval Warm Period than their "no-dendro" subset (see Figures 2 and 3 in Shi et al., 2013)[A28].

The use of these proxies does not in itself automatically alter the ratio. For example, although Moberg et al., 2005[A17] used both the Yamal chronology and two foxtail series, they only used the high-frequency components of those series, i.e., they removed the long-term trends and just kept the inter-annual variability. In addition, although the bristlecone/foxtail pines are all from a similar area (south-western North America), there are significant differences between different chronologies which have been constructed from them, e.g., Lloyd & Graumlich, 1997's foxtail chronologies[A136] suggest a warmer Medieval Warm Period than the Graybill & Idso, 1993[A137] bristlecone/foxtail chronologies.

Nonetheless, as we will discuss in Sections 3.4.1 and 3.4.2, both of these proxy groups have been controversial, so it is surprising that they have not been used with more caution. More importantly, if these specific proxies are critical in establishing the ratio of the two warm periods, then this has serious consequences for the robustness of the studies. Hence, it is worth briefly reviewing the controversy over these two specific proxy groups in Sections 3.4.1 and 3.4.2.

As we mentioned in Section 3.1, Mann et al., 2008[A22] argue that they obtain similar temperature estimates even if they exclude all their tree-ring proxies, provided they include the Tiljander lake sediment proxies. So, we will also briefly assess the Tiljander proxies in Section 3.4.3.

Millennial temperature estimate	# series	Bristlecones/foxtails		Yamal chronology
		Individual	"MBH PCs"	
Jones et al., 1998[A9, A109]	17			
Mann et al., 1999[A11]	12		√	
Briffa, 2000[A13, A110]	7			√
Crowley, 2000[A12, A111]	15	√		
Esper et al., 2002[A14, A112, A113]	14	√		
Mann & Jones, 2003[A15, A16]	13		√	√
Moberg et al., 2005[A17, A97, A114]	18	√		√
D'Arrigo et al., 2006[A18]	19			√(1)
Hegerl et al., 2007[A19]	14	√	√	√
Juckes et al., 2007[A20]	13	√		
Loehle, 2007[A21, A51]	18			
Mann et al., 2008[A22, A133, A134]	1209(2)	√(3)		
Ljungqvist, 2010[A24]	30			
McShane & Wyner, 2011[A25]	1209(4)	√		
Christiansen & Ljungqvist, 2011[A26]	40	√		√
Christiansen & Ljungqvist, 2012[A27]	32	√		√
Shi et al., 2013[A28]	45	√		√

Table 2: *The use of bristlecone/foxtail pines and/or the Yamal chronology in the various proxy-based millennial temperature estimates (listed chronologically). Bristlecone/foxtail series were either used as individual series, or indirectly through the use of Mann et al., 1999's 1st principal component of the North American tree ring database ("MBH PCs"), which was heavily weighted by bristlecone/foxtail pines (see Section 4.3). (1) D'Arrigo et al., 2006 used Yamal, but in the text they used the name and core counts of a neighbouring chronology - "Polar Urals (POL)"[B32]. (2) Mann et al., 2008 considered a large number of series, but only 59 of them extended back to 1000 A.D. Many of those series were discarded for showing poor correlation to the calibration data. (3) Mann et al., 2008 also carried out a sensitivity test where they excluded bristlecone/foxtails as well as other tree rings, but included the problematic Tiljander lake sediments - see Section 3.4.3. (4) McShane & Wyner, 2011 used the same dataset as Mann et al., 2008.*

3.4.1 Bristlecone/foxtail pine proxies

One family of trees which has been of considerable interest to climatologists is the bristlecone pine family. This consists of three closely-related five-needled pine species found at high altitudes in the California, Nevada and Colorado mountain ranges - the Rocky Mountains bristlecone pine (*Pinus aristata*); the Great Basin bristlecone pine (*Pinus longaeva*); and the foxtail pine (*Pinus balfouriana*).

The bristlecone pine trees are very long-lived - in some cases being several millennia old. It has been supposed that the highest altitude trees of these species are temperature-sensitive. These two factors initially suggest that they would make promising temperature proxies. However, LaMarche et al., 1984[A138] had noted unusual tree ring growth in bristlecone pines in recent decades, which had no relation to regional climatic trends.

LaMarche et al., 1984 suggested that the unusual growth was due to fertilisation from increasing concentrations of atmospheric CO_2, although this theory was controversial[A139]. In order to investigate this theory, Graybill & Idso, 1993[A137] sampled various bristlecone and foxtail pines. As well as the regular (*"full-bark"*) trees, *"strip-bark"* trees were also sampled. Strip-bark trees are pine trees where a lot of the bark has peeled off, leaving only strips of bark.

Graybill & Idso believed that the strip-bark trees would be more influenced by changes in CO_2. Indeed, they found a rapid increase in growth rate after the mid-19th century in the strip-bark trees, but not the full-bark. They agreed with LaMarche et al., 1984 that this dramatic growth was *not* related to local temperature changes, but was merely a consequence of CO_2 fertilisation.

Despite Graybill & Idso's explicit statement that the unusual growth rate of their strip-bark pines was non-climatic, the *hockey stick study* used the Graybill strip-bark chronologies as temperature proxies, contributing strongly to its "hockey stick" shape[A44, A45]. Before the *hockey stick study*, none of the proxy-based estimates used these proxies[B57] as it was generally agreed that their rapid 20th century growth was *not* due to temperature[B57–B59]. But, it can be seen from Table 2 that they have been heavily used since.

LaMarche et al.'s theory of CO_2 fertilisation was criticised because it had not been detected in other tree species[A140] or in the full-bark pines[A136], which appears a valid point. But, various other non-climatic explanations have been suggested for the unusual growth[A42, A136]. So, to justify the widespread use of bristlecone/foxtails in proxy-based temperature estimates, it is important to provide some evidence that its anomalous growth is related to local temperatures.

McIntyre specifically compared several of the Graybill pines to local temperatures, and found they were very poorly related[B60]. In addition, other tree ring studies in the area found the Current Warm Period to be comparable to the Medieval Warm Period[A136, A141, A142]. Indeed, when Ababneh carried out an update for her Ph.D. thesis[A143], on a Graybill chronology which had originally shown particularly strong 20th century growth, the 20th century growth no longer appeared unusual[B61, B62]. A recent isotopic analysis of several bristlecone trees also failed to identify anomalous 20th century climate change[A144]. After carrying out an update of another Graybill chronology, McIntyre noted that the recent sharp growth in strip-bark cores was often countered by reduced growth in other cores from the same tree. He suggested that the unusual growth may be related to the elliptical growth of strip-barked trees, rather than a climatic effect or CO_2 fertilisation effect[B63].

Bunn et al., 2005[A145] claimed that the unusual growth of the bristlecones in the 20th century was temperature-related. However, their entire basis for this claim was that the bristlecone growth was similar to the *hockey stick study*. This was effectively circular logic since the *hockey stick study* was itself heavily dominated by the Graybill pines[B64].

Later, Salzer et al., 2009[A146] claimed to have vindicated the use of strip-bark bristlecones as temperature proxies. They had updated several of the Graybill proxies on Sheep Mountain. They then compared the bristlecone growth rates to those of other tree ring measurements in a similar area - the "MXD" measurements of Rutherford et al., 2005[A147]. They found a reasonable match during the period 1630-1950, and therefore concluded that if the Rutherford MXD measurements were reliable, then so were their updated bristlecones. However, the Rutherford MXD measurements do *not* show the post-1900 "hockey stick" shape of the updated bristlecones (see Figure 5 of Salzer et al., 2009[A146]). Hence, that argument of Salzer et al. is limited to suggesting the bristlecones may have some merit *before* the contentious "hockey stick" rise.

Salzer et al., 2009 also argued that the characteristic "hockey stick" trend occurred in both the whole bark and the strip-bark pines - contradicting Graybill & Idso, 1993[A137]'s findings. They suggested that the contradiction was due to an inappropriate standardization used by Graybill & Idso. Hence, they compared the non-standardized chronologies of the whole-bark and strip-bark pines. They found no substantial difference between the two chronologies in the modern period[A146]. On this basis, they concluded that there was no divergence between the strip-bark and whole-bark. However, in Figure S4 of their Supplementary Information[A146], it is apparent that when they took this approach, there was a divergence *before* the 20th century. Hence, that particular argument appears very weak[B65].

Recently, Salzer et al., 2013 have put up another argument[A148]. When they compared their bristlecone chronology to three different Global Climate Model (GCM) simulations of the last 1000 years, one of the simulations ("ECHO-G2") showed a similar trend to their bristlecone chronology. However, we note that neither of the other two simulations ("MPI" or "CSM") showed this trend (see their Figure 4). So, we do not consider this a particularly compelling argument.

Bearing all of this in mind, there should be serious concern over the estimates which used bristlecone/foxtail pines. As can be seen from Table 2 this includes most of the millennial estimates. Even if part of the sharp 20th century up-tick in some of the bristlecone/foxtail pines is found to be due to temperature change[B66], considering the controversy over them, it is surprising they are so widely used.

3.4.2 The Yamal chronology

Briffa, 2000[A13] introduced the Yamal chronology, which showed dramatic growth in the 20th century. As can be seen from Table 2, it has been extensively

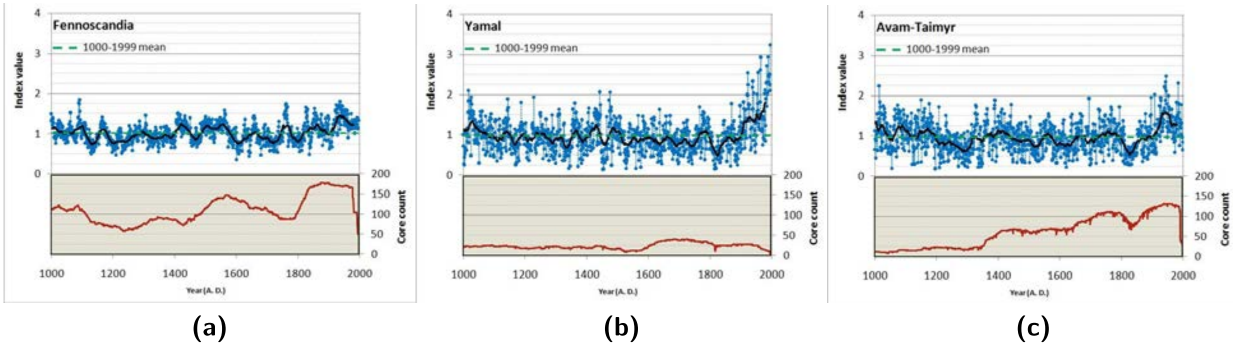

Figure 9: *The three northern Eurasian chronologies given in Briffa et al., 2008[B8], and the numbers of cores used for their construction. Data taken from* http://www.cru.uea.ac.uk/cru/people/melvin/PhilTrans2008/. *Thick black lines correspond to 31-year running means.*

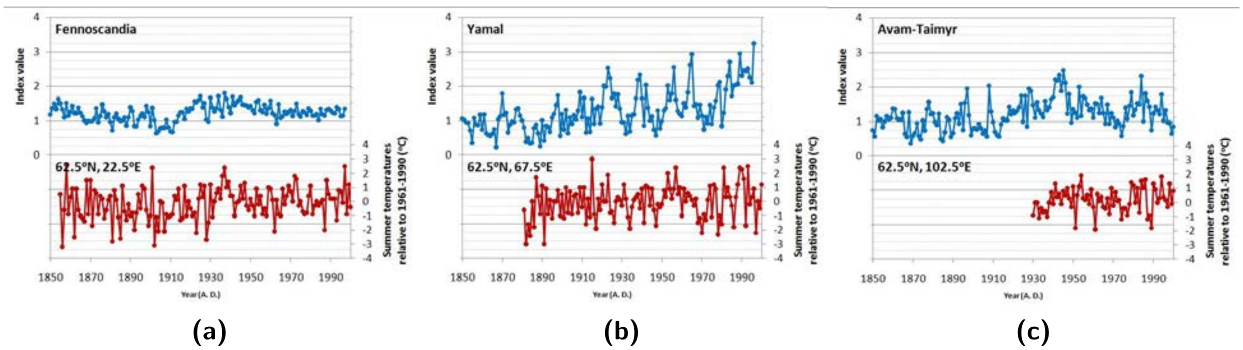

Figure 10: *Comparison of local gridded (weather station-based) summer temperatures (June-August) to the three chronologies described in Figure 9. Data taken from* http://www.cru.uea.ac.uk/cru/people/melvin/PhilTrans2008/.

used since. However, for such a widely used proxy, it has a number of problems.

Briffa et al., 2008[A149] revisited this Yamal chronology and created two other northern Eurasian chronologies - Fennoscandia and Avam-Taimyr. All three of these chronologies were located at around 62.5^{o}N, at different locations on the Eurasian continent. However, they each present rather different estimates for temperatures of the last millennium (Figure 9). If the strong 20th century growth rate of the Yamal chronology is genuinely representative of global temperatures, then it is hard to see why it is largely absent from the other two chronologies (from the same latitude and continent). Indeed, on the basis of the number of cores used for the construction of the chronologies (bottom panels of Figure 9), Yamal would appear the least reliable of the three.

Briffa et al. implied that all three chronologies showed a reasonable correlation with local summer temperatures (e.g., see Figure 1 of Ref. [A149]). However, from Figure 10, this is not immediately obvious. Certainly, the distinctive 20th century growth implied by the Yamal chronology appears to be absent from the corresponding local gridded temperatures (Figure 10b).

Following the publication of Briffa et al., 2008, Briffa finally archived the data for the Yamal chronology after several years of requests from McIntyre[B67]. McIntyre noted that only a few trees (17) were used for constructing the recent portion of the Yamal chronology[B68], i.e., the living samples. In addition, one of the trees, YAD061, showed 8 standard deviations of growth in the 20th century - a remarkable growth rate, which was not matched by any of the others. This had noticeably increased the 20th century average of the chronology[B69].

McIntyre carried out two sensitivity experiments for the Yamal chronology. In one experiment, he removed 12 cores and replaced them with 34 archived cores from the Khadyta River (which was in the Yamal area). In the other experiment, he added the 34 cores to the complete Yamal chronology. In the first experiment, the unusual 20th century growth was replaced with a decline. In the second experiment, the 20th century growth was higher than in the centuries immediately preceding it, but comparable to growth at various stages over the last two millennia, including the 11th and 15th centuries[B10].

Briffa et al. criticised these experiments[B70, B71] and suggested that the cores McIntyre had selected were anomalous and arbitrarily chosen. However, McIntyre argued that he had done a better job of justifying his selection than Briffa had for his selections[B72]. He also argued that the 17 living cores in Briffa's original chronology were inhomogeneous[B73], i.e., there was little consistency from core to core and between them and the sub-fossil cores, and that the Khadyta River cores showed better homogeneity.

Condon argued that the "hockey stick" shape of Yamal was an artefact of Briffa's age-related tree ring standardisation (see Section 2.1), and argued that other plausible standardisations yielded 20th century growth rates that were fairly average[B74].

Recently Briffa et al., 2013 has revised the Yamal chronology[A150]. Apparently, the new revision has reduced the magnitude of the "hockey stick" up-tick in the process[B75], although Bouldin argues that the tree ring age standardisation used is still inappropriate for the underlying data[B14]. At any rate, whether the Yamal chronology has any merit as a temperature proxy[B32] or not[B76], it is striking that its distinctively sharp 20th century growth is absent from the other Briffa et al., 2008 chronologies (Figure 9) as well as the two versions of the nearby Polar Urals chronology (Figure 4). It also fails to detect the strong Medieval Warm Period others have reported in the area[A151]. This suggests that it should only be used cautiously in proxy-based temperature estimates, if at all.

3.4.3 The Tiljander lake sediments

Following criticism[A38, A41, A42, A44, A45, A152, A153] of Mann et al.'s *hockey stick study*[A10, A11, A135, A154] for being highly dependent on the Graybill strip-bark pines described in Section 3.4.1, Mann et al., 2008[A22, A133, A134] claimed that their estimate was robust to the exclusion of either (a) tree rings or (b) a new set of 7 other potentially problematic proxies.

Four[B77] of these 7 non-tree ring problematic proxies were Tiljander et al., 2003[A155]'s Lake Korttajärvi sediment cores from Finland. Tiljander et al. had constructed a 3,000 year long chronology from lake sediments which suggested a strong Medieval Warm Period around 980-1250AD with several cool periods during the 16th, 17th and 18th centuries, possibly corresponding to the Little Ice Age.

However, after about 1720AD, the sediments appeared to have become increasingly contaminated by local human activity, e.g., wastewater run-off and bridge construction. This seems to have led to anomalously low apparent "temperatures". Therefore, Tiljander et al. had stressed that much of the post-1720 variability was strongly non-climatic.

Recognising that there was a problem with the post-1720 portion of the proxies, Mann et al., 2008 treated the Tiljander proxies as having *"potentially spurious features"*[A22]. However, Mann et al. still decided to use the complete proxies including the contaminated sections, anyway.

As an additional problem, Mann et al. effectively used two of the proxies in the opposite manner to that intended by Tiljander et al. suggesting a cold "Medieval Warm Period", mild "Little Ice Age" and a "hockey-stick" like warming for the Current Warm Period[B78].

Mann et al., 2008 created two separate sets of estimates - one using a composite-plus-scale approach ("CPS") and one using a climate field reconstruction ("EIV"). For the CPS estimates, the inversion of the Tiljander proxies from their intended interpretation appears to have been manually done.

For the EIV estimates, the inversion was an implicit feature of the algorithm which altered the sign of the proxy to yield the highest correlation with the weather station-based calibration temperatures of the Current Warm Period. Since the post-1720 portion of the proxy was non-climatic, any apparent correlation between temperature and the proxy in this period would be just a coincidence. However, the EIV algorithm does not consider this possibility. Hence, the sign of the proxies were adjusted by the algorithm so that the non-climatic portions appeared to show "warming" in the Current Warm Period[B79].

This second approach was also carried out in Mann et al., 2009[A123]. Kaufman et al., 2009[A23] also used these Tiljander proxies inverted in their Arctic analysis, in the same way Mann et al. had used them in the CPS estimate. However, when Kaufman et al. discovered that this was not how Tiljander et al. had intended them, they issued a correction to revert the sign back to the original interpretation[A156].

Mann et al., 2008 relied on the Tiljander proxies for their claim that their estimates were not dependent on the use of the bristlecone/foxtail pines[A22]. If they carried out a sensitivity analysis by removing all tree-ring proxies (including the bristlecone/foxtail pines), they obtained a similar estimate to their complete analysis. However, that "no-dendro" estimate included the four Tiljander proxies (with the contaminated portions) as well as another three proxies they had identified as potentially problematic. To test if they were a problem, they carried out a second sensitivity analysis by removing the 7 non-tree ring potentially problematic proxies, but leaving all the others (including the bristlecone/foxtail pines) in. This also yielded a similar estimate. On this basis, they concluded that their estimate was not biased by any particular proxy.

Strangely[B80], they did not carry out the simple test of *just* removing the 7 non-tree ring proxies they had identified as potentially problematic *and* the bristlecone/foxtail pines that the *hockey stick study* had specifically been criticised for using[A38, A41, A42, A44, A45, A152, A153]. Nonetheless, after much debate on the blogs over the reliability of the Tiljander proxies (see links at Ref. [B81]), Mann et al., 2009 included in Figure S8 of their S.I.[B82], results of an additional sensitivity analysis carried out for Mann et al., 2009[A123], which was equivalent to the EIV estimate of Mann et al., 2008. If both the tree ring proxies and the Tiljander proxies were excluded, then estimated temperatures for the period 1000-1850 were substantially increased[B83–B85]. However, the estimates failed verification before 1500 (possibly because they had excluded so many proxies).

At a later stage, Mann posted on his website[B86], a similar test for the CPS estimate. Again, this had significant effects, e.g., temperatures in the Medieval Warm Period reached higher values than in the 20th century. This suggests that the Mann et al., 2008 estimates were *not* robust to the proxies used, as had been claimed. Indeed, it again highlights the danger in relying heavily on questionable proxies, such as the bristlecone/foxtail pines discussed in Section 3.4.1, the Yamal chronology discussed in Section 3.4.2, or even the Tiljander lake sediments which were known to be problematic after 1720.

4 Criticism of the *hockey stick study*

As discussed in Section 1, the *hockey stick study* by Mann et al.[A10, A11, A154] was very influential, both politically and socially, due to its prominence in both scientific[A40] and popular presentations[B2][2]. Perhaps for this reason, despite a number of flaws having been identified with it[A37–A39, A41–A46], its dramatic claims that (i) global temperature change since the late 19th century have been strongly dominated by man-made global warming[A10], and (ii) current temperatures are unprecedented in the last millennium[A11] appear to be widely believed by the general public.

Part of this seems to be due to Mann and his supporters continuing to imply that it was a reasonably accurate study, e.g., see Ref. [B88] for a 2011 TEDxTalk by Mann, or Mann, 2012[B5].

Another factor also appears to be that many supporters of man-made global warming theory are reluctant to acknowledge that there may have been flaws with the iconic *hockey stick graph*. This appears to be due to a fear that if the public becomes aware of those flaws, they may become suspicious of other aspects of climate science. For example the Anonymous Reviewer #1 for Ljungqvist et al., 2011[B89] believes that there is a *"stubbornness by the sceptical community to accept very real environmental and climatic changes that more and more appear to be exceptional over the last 1000 years"* and worries that criticising previous palaeoclimatology studies might *"muddy the message"*.

This should be irrelevant for the reader who is trying to genuinely understand how climate has changed over the last millennium or so. However, the *hockey stick study* still seems to hold a strong influence on public thought. Hence, in this section, we will review the contentious debate over this one particular study. The reader who is uninterested in this outdated study may prefer to skip to Section 5.

[2]The *hockey stick study* appears to have been mistakenly labelled as "Dr. Thompson's thermometer" in Ref. [B2] - see Ref. [B87].

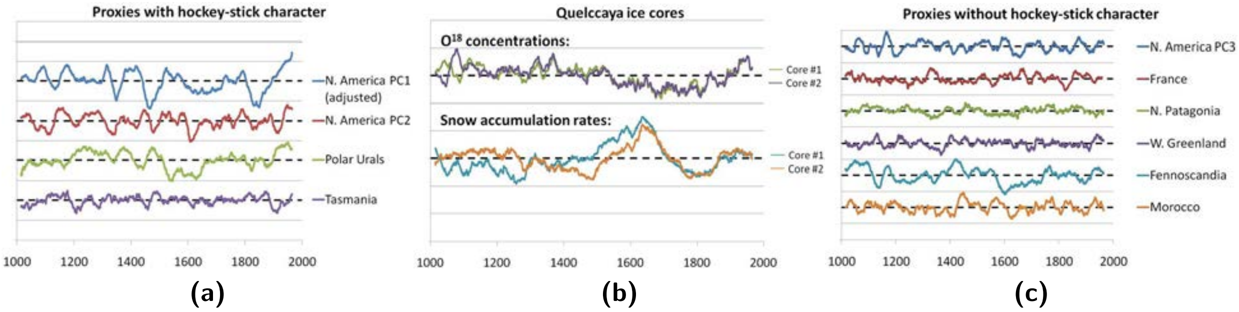

Figure 11: *All of the 12 proxy series used for the 1000-1400 step of the* hockey stick study. *All data has been smoothed with a 31 year running mean, for clarity. (a) includes the only series which show any sort of "hockey stick"-like uptick for the Current Warm Period. (b) There were two nearby ice cores taken from Quelccaya, and so the two series from there were each averages of two cores. (c) shows the other series. Proxy data taken from World Data Center for Paleoclimatology, except for the Morocco data which was taken from the Climate Audit website. Units were not provided with the archived data, and vary from proxy to proxy, so they are all plotted here in generic "proxy units". Each proxy was rescaled by dividing by its standard deviation and subtracting its 1000-1980 mean.*

4.1 Initial criticism and defence

As discussed in Section 3.2, while the *hockey stick study* used 159 proxies in total, for the critical 1000-1400 step they only used 12 proxies. With this in mind, the first point to note is how few of these 12 proxies (Figure 11) bear any resemblance to the final *hockey stick graph* (Figure 1).

Jones, 1998[A157] criticised Mann et al., 1998 for using long thermometer records as some of their temperature "proxies" and for failing to adequately discuss the problems associated with the various proxies they used. He also noted that other temperature estimates for 1400-1900s suggested different temperature histories. However, he later clarified[A158] that he agreed that those other estimates concurred with Mann et al., 1998 that the 20th century was warmer than the 15th-19th centuries.

Briffa & Osborn, 1999[A41] cautioned that the apparent agreement of the *hockey stick study* with others, was at least partially due to a substantial overlap in the proxies used by those studies (Section 3.2). They worried that Mann et al. had not paid enough attention to the problem of tree ring standardisation (Section 2.1). They also expressed concern over an adjustment Mann et al., 1999 had applied to one of their series which they relied heavily on - the first principal component (PC1) of a network of tree rings which was dominated by bristlecone/foxtail pines from western USA.

As discussed in Section 3.4.1, some researchers had argued that these trees were showing unusual 20th century growth due to CO_2 fertilisation[A137, A138]. To counter this concern, in their extension to Mann et al., 1998, Mann et al., 1999 had applied an *ad hoc* adjustment to this series (PC1 in Figure 11). However, there were a number of problems with this. First, the actual adjustment seems somewhat arbitrary, and not as simple as Mann et al., 1999 implies[B41, B61, B90–B93]. Second, from Figure 11, it does not appear to have worked since (even after applying this adjustment) the 20th century up-tick of the "PC1" series is still far sharper than the other series. Finally, they only applied the adjustment to the 1000-1400 step, so even if the adjustment did correctly remove a non-climatic bias, the bias still remained in the 1400-1980 steps[B41].

Broecker, 2001[A39] expressed concern over the apparent absence of a strong Medieval Warm Period in the *hockey stick study*. He argued that there was considerable evidence to counter this finding. Although Bradley et al., 2001[A124] pointed out that some studies failed to find a Medieval Warm Period, Soon et al.[A37, A38] pointed out a large selection of studies which did (see Section 3). Idso et al. have similarly found evidence for a globally distributed, strong Medieval Warm Period in their literature review[B27].

More specifically, Soon et al. only found a few studies (including the *hockey stick study*) which showed the Current Warm Period to be climatically anomalous in the last millennium (either in terms of temperature or precipitation). Indeed, they claimed the opposite, i.e., that *"the 20th century is probably not the warmest nor a uniquely extreme climatic period of the last millennium."*[A37], although this particular claim was strongly criticised by von Storch for being inadequately justified[B94].

Mann et al., 2003a[A159] criticised Soon et al. However, they do not appear to have carefully considered Soon et al.'s arguments since most of their criticisms had already been addressed in the Soon et al. articles[A37, A38]. Nonetheless, a brief debate was attempted[A160, A161].

McIntyre & McKitrick, 2003[A42] re-analysed Mann et al., 1998 using the data and methods provided to them by Mann et al.[B95]. They found that the data set Mann et al., 1998 had used was poorly organised and also contained a large number of errors. After accounting for those errors, and replacing some proxies with more up-to-date versions or comparable substitutes, their reanalysis suggested the 15th century was warmer than the 20th century. This contradicted the *hockey stick study*'s conclusion that 20th century temperatures were unusually warm, suggesting that the *hockey stick study* was not robust.

Rutherford et al., 2005[A147] suggested that some of the errors McIntyre & McKitrick, 2003 had noticed were due to them using an incorrect dataset. When McIntyre had asked Mann for the Mann et al., 1998 data, Mann had put him in contact with Rutherford who apparently gave McIntyre a slightly incorrect version. Ironically, this apparently incorrect version appears to have been the one used by Rutherford et al., 2005 and also later archived in Mann et al., 1998's 2004 corrigendum[B96]. The fact that even the authors of Mann et al., 1998 (who also co-authored Rutherford et al., 2005) were unclear over which dataset to use seems to have vindicated McIntyre & McKitrick's criticisms of the disorganised nature of the Mann et al., 1998 study.

However, Rutherford et al. also argued that McIntyre & McKitrick had taken a traditional approach to calculating the principal components of Mann et al., 1998's high density tree ring networks (see Section 4.3), rather than the undisclosed approach which it transpired Mann et al., 1998 had actually used. This apparently led to too strong an increase in the 15th century temperatures. McIntyre & McKitrick, 2005b[A45] applied the now-disclosed approach and the 15th century temperatures were indeed a bit lower than for McIntyre & McKitrick, 2003. Nonetheless, they were still comparable to the 20th century temperatures, and so the contradiction with the *hockey stick study's* conclusions remained.

4.2 "Pseudoproxy" analysis of the *hockey stick study*

In Section 2.4, we mentioned that one useful validation test that can be carried out on a temperature reconstruction method is to use pseudoproxy analysis.

Due to the high profile nature of the *hockey stick study*, a number of pseudoproxy studies have been carried out[A43, A46–A48, A82–A96] to investigate the reliability of its particular reconstruction method, henceforth referred to as the "MBH" method (after the initials of Mann, Bradley and Hughes, i.e., the authors of the *hockey stick study*).

A difficult challenge in this approach is in deciding how to construct realistic pseudoproxies. From a model simulation, it is relatively easy to generate pseudoproxies for the same locations as the proxy network used by the *hockey stick study*. This can be done by simply selecting the gridded simulated temperatures for those locations.

However, as we discussed in Sections 2.1-2.3, real proxy series contain a lot of "noise" from non-temperature factors. Also, the strength of the temperature response of the proxy could vary over time. To account for this proxy "noise", a simple first approximation in the construction of a pseudoproxy network is to introduce different amounts of random noise. In this way, pseudoproxies with different "signal-to-noise" ratios can be generated.

Ordinary random noise is considered "white". However, often noise has non-random properties. "Red" noise is noise whose value for one point has some dependence on the previous point, i.e., it is possible to have randomly occurring trends. Many temperature proxies, such as tree rings are thought to have more similarity to red noise than the trend-less white noise. In the case of tree rings, a previous years' growth can influence the next year's growth[A74, A75]. For example, a year of good growth could make the tree healthier, improving its growth for the next year.

As a first step, von Storch et al., 2004[A43] tested the *hockey stick study* reconstruction method on a pseudoproxy network constructed by applying varying amounts of white noise to the "Erik" simulation of the last millennium. They found that, even with white noise, the MBH method substantially underestimated the actual temperature variability of the simulation. Their results suggested that much of the apparent "flatness" of the "hockey stick handle" was merely an artifact of their reconstruction method.

The von Storch et al., 2004 study was quite controversial and led to considerable debate[A46–A48, A82, A83, A85, A96, A162]. Much of this debate was over the fact that they had used so-called "detrended" pseudoproxies[A84]. Before carrying out their analysis, they had temporarily removed the long-term trends of all their pseudoproxies and calibration data, so that they would achieve a better inter-annual calibration, and thereby a more realistic estimate overall. However, Wahl et al., 2006[A83] argued that this detrending should not be carried out. They showed that, if non-detrended pseudoproxies were used, the underestimation of the MBH method was somewhat reduced[A83]. In response, von Stoch et al. noted that, even using non-detrended pseudoproxies, the underestimation was *still* substantial[A84, A85].

Another criticism was that there were problems with the "Erik" simulation von Storch et al. had used[A162]. In particular, the simulation had been insufficiently equilibrated, and so it had suggested a warmer Medieval Warm Period than other simulations. However, for the purposes of pseudoproxy tests, this was irrelevant, since they were merely assessing how successful the MBH method was at reconstructing the simulated temperatures, *not* how accurate the simulated temperatures were[A86]. Indeed, similar results were found for the MBH method when an improved simulation ("Erik II") were used[A96].

Rutherford et al., 2005[A147] applied a new method, called "RegEM", to the same proxy network as the *hockey stick study* and achieved a similar result. When Mann et al., 2005[A82] carried out their own pseudoproxy analysis on this new method, the RegEM method appeared to be very successful at reconstructing simulated temperatures.

Initially, this seemed to suggest that the conclusions of von Storch et al. were invalid, leading to some debate[A82, A86, A87]. However, it later transpired that Mann et al., 2005 had made a serious error in their analysis. Before applying the RegEM method, they had standardised all their pseudoproxies over the entire simulation period, rather than just over the calibration period[A76, A88, A89, A163]. This meant that all of their pseudoproxies already roughly approximated the simulated temperature over the entire simulation. In the real world, the pre-instrumental temperatures are unknown - after all, that is why proxy-based studies are being carried out. After correcting for this, the RegEM method also significantly underestimated the actual simulated temperatures[A88].

Mann et al., 2007c[A89] tested a new version of RegEM, called "RegEM TTLS" (the older version is now known as "RegEM Ridge"). This method did not show as much underestimation as the older version (or the original MBH method), and when this method was applied to the *hockey stick study's* proxy network, it again yielded a similar reconstruction to the original *hockey stick study*.

This initially appears puzzling[A92, A164]. Although Smerdon et al., 2008b[A90] noted that Mann et al. had been using a badly corrupted version of a computer simulation for their 2005 and 2007 analyses, this did not affect Mann et al., 2007c's essential conclusion[A90, A91, A164]. Even though the RegEM methods still showed underestimation[A76, A98, A165], they did appear to give more realistic results than the original MBH method[A76]. However, when applied to the *hockey stick study's* proxy network, they all yielded essentially the same result[A89, A90, A164].

A likely explanation is that while there were problems with the original MBH method, coincidentally, there were also serious problems with the proxy network itself. As we will see in Sections 4.3 and 4.4, this is the case.

The quite technical and seemingly continuous back-and-forth nature of the pseudoproxy analysis debates over the *hockey stick study* can initially be quite hard-to-follow and/or overwhelming to a non-specialist. This is especially so, since both critics and supporters of the *hockey stick study* have claimed that their position is backed by the data.

For this reason, it may be helpful to briefly summarise the main *current* arguments of the two camps:

- Critics of the *hockey stick study* argue that the original MBH method seems to have been highly flawed, and to yield unreliable results[A96].
- Supporters of the *hockey stick study* argue that, even if the original MBH method was flawed, the newer "RegEM TTLS" method gives similar results[A164].

The latter argument seems to us reminiscent of Babbage's quandary,

> *'On two occasions I have been asked, – "Pray, Mr. Babbage, if you put into the ma-*

chine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question. - p67, Babbage, 1864[B97].

Nonetheless, von Storch et al., 2004[A43]'s study has been very useful in that it has led to a recognition of the value of pseudoproxy studies and the development of more robust reconstruction methods. This should be of benefit if applied to more reliable proxy networks, as we discussed in Section 2.4.

4.3 Principal Component Analysis ("PCA") problems

One problem with the *hockey stick study's* proxy network was that most of the proxies were from similar areas. In particular, 70 of the 95 series used for the 1400-1450 step[3] were U.S. tree ring series. Therefore, if the *hockey stick study* had given all the series a similar weighting then their entire "Northern Hemisphere" estimate would be strongly dominated by those proxies, and would be little more than a "U.S." temperature estimate.

In an attempt to overcome this weighting problem, Mann et al., 1998 tried to estimate the main "climatic signals" of these high density networks through principal component analysis[4] ("PCA"). They then treated the top few principal components for those networks as replacement "proxies". For the 1400-1450 step, 3 out of the 22 series used were principal components, while for Mann et al., 1999's 1000-1400 step, they comprised 3 out of 12 series ("PC1-3" in Figure 11).

McIntyre & McKitrick, 2005a noticed that the Mann et al., 1998 algorithm for calculating principal components for those high density networks was non-standard. Mann et al., 1998 normalised all of the individual proxies to their 1902-1980 means instead of the standard approach of normalising the data over the means of the entire period being considered, e.g., 1400-1980 in the 1400 step[A44, A45, A116, A166–B168, B98].

This was significant because it gave very high weights to proxies whose 1902-1980 mean was substantially different from the mean over the entire period. This meant that those tree ring series which did *not* show unusual 20th century growth (i.e., proxies without "hockey stick" shapes) received negligible weighting, while those series with the greatest "hockey stick" shape received the greatest weighting.

For the North American network, Sheep Mountain (the proxy with the strongest "hockey stick" - and also one of the bristlecone pine proxies mentioned in Section 3.4.1) received 390 times the weight of Mayberry Slough (the proxy with the weakest "hockey stick")[A44, A45] in the 1400-1450 step - see Figure 12.

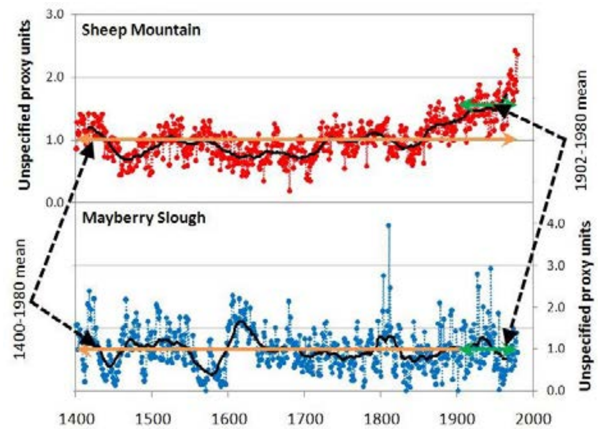

Figure 12: *The highest weighted (top panel) and lowest weighted (bottom panel) series in Mann et al., 1998's 1400-1450 1st principal component of the North American ITRDB tree ring proxies. Black curves correspond to 31-year running means. Proxy data taken from the Supplementary Information to the 2004 corrigendum to Mann et al., 1998[A135]. Proxy names taken from ClimateAudit.*

McIntyre & McKitrick demonstrated the problem this introduced, by generating a large number of random, red noise simulations with no overall trend. When they applied the standard principal component analysis to these simulations, the 1st principal components (PC1) showed no trend. This was the correct result, since the simulations had no overall trend. But, when they applied the Mann et al., 1998 version, the 1st principal components tended to have "hockey stick" shapes, even though they had no intrinsic trend.

As von Storch & Zorita, 2005 noted[A166], the magnitude of McIntyre & McKitrick's red noise "hockey sticks" was small compared to the Mann et al., 1998 global temperature estimate. But, McIntyre & McKitrick were not suggesting that this artefact

[3] 70 out of the 110 series they considered[A135].

[4] Not to be confused with the separate principal component analysis of the calibration data which they used for their reconstruction, i.e., the "MBH method" discussed in Section 4.2.

in itself led to the hockey stick shape of the *hockey stick study*[A167] (although Mann mistakenly seems to have thought they were[B99]). Rather, the significance was that it showed that the Mann et al., 1998 version effectively "mined" the high density networks for "hockey sticks". As a result, the 1st principal component for the North American network ended up excessively dominated by the problematic Graybill bristlecone/foxtail strip-bark pines discussed in Section 3.4.1.

Huybers, 2005[A168] agreed that the Mann et al., 1998 version was flawed, and underestimated pre-20th century temperatures. However, he argued that McIntyre & McKitrick should have scaled their proxies to unit variance before their analysis, since some of the proxies showed less variability than others. When Huybers did this, he obtained an intermediate result between Mann et al., 1998 and McIntyre & McKitrick, 2005a.

McIntyre & McKitrick responded that this was only really of relevance to accommodate two of the 70 tree rings in the 1400-1450 North American network[A116]. They argued that it also underestimated the variance of those proxies which showed strong trends, i.e., the strip-bark pines. Moreover, they noted that when the three different 1st principal components were plotted to the 1400-1980 mean, instead of the 1902-1980 mean as Huybers had done, both Huybers' and McIntyre & McKitrick's versions were actually quite similar, while the Mann et al., 1998 version was a clear outlier.

The effect of the *hockey stick study's* non-standard principal component analysis was most pronounced in the earliest step (1400-1450). In Mann et al., 1999's 1000-1400 step, the North American 1st principal component ("PC1" in Figure 11) was also dominated by the Graybill strip-bark pines, but this was mainly due to the fact that these were the trees with the longest chronologies in that network.

McIntyre & McKitrick noted that, using the standard approach, the strong "hockey stick" shape of the bristlecones were instead relegated to the 4th principal component (PC4). If they then carried out the rest of the Mann et al., 1998 algorithm (i.e., including the top two principal components), this made the 15th century appear comparable to the 20th century, i.e., the "hockey stick" disappeared[A45].

Mann and his colleagues attempted to counter this criticism in a few ways, although their arguments seem to have been based on a misunderstanding of the criticism and/or the reasons for using principal component analysis. For instance, Mann claimed that the Mann et al., 1998 approach was a well-established form of principal component analysis, which had been recommended by Jolliffe for certain applications[B99]. Jolliffe denied this and strongly criticised its use in Mann et al., 1998 once he became aware of it[B100].

Mann also claimed on his Real Climate blog that, if they had used the standard approach which McIntyre & McKitrick favoured, then the top *five* principal components should be used, rather than the top two used with the *hockey stick study* approach, stating that Mann et al., 1998 had used "Preisendorfer's Rule N"[B44, B101, B102]. Hence, he argued they could still include the hockey stick shape of the Graybill pines. This argument was later repeated by Ammann & Wahl[A169, A170]. However, McIntyre noted that:

- There was no evidence that Mann et al., 1998 had actually used Preisendorfer's Rule N[B103].
- There were many other selection rules which could have been used[B104]
- It was unclear if Preisendorfer's Rule N was appropriate[B105]

Wahl & Ammann argued that it was important to include the bristlecone/foxtails, otherwise the *hockey stick study* failed its verification tests[A170]. But, this had been McIntyre & McKitrick's essential criticism - if the *hockey stick study* was supposed to be genuinely representative of northern hemispheric temperatures, then it should not have to rely on a small subset of trees in western U.S.[A44, A45]. This point had also been made earlier by Soon et al., 2003b[A38].

Mann et al. argued that Mann et al., 1998's hockey stick shape could also be obtained without using any principal component analysis[A89, A169, A170][B44, B101, B102]. However, that was merely because the entire proxy network was then dominated by the U.S. tree ring network - the problem the principal component analysis was supposed to reduce. In that case, the *hockey stick study* was again biased by the problematic Graybill pines, due to them comprising 20 of the 95 series[B106]. This was easily confirmed by removing the Graybill pines from the network, since the 15th century temperatures then appeared comparable to those of the 20th century[A45].

Finally, Rutherford et al., 2005[A147] had repeated the Mann et al., 1998 estimate using a slightly different approach (the "RegEM" method described in

Section 4.2), and obtained a similar result. Mann et al. claimed that this vindicated the approach of the original *hockey stick study*[B44, B101, B102]. However, Rutherford et al., 2005 had used the same proxy network and principal component analysis as Mann et al., 1998[5], so the criticisms of the original *hockey stick study* still held[B107].

4.4 Lack of statistical robustness

It is often assumed that the temperature proxies used for proxy-based temperature estimates are at least moderately correlated to actual local temperature measurements[A16]. Indeed, most readers would probably consider this an essential requirement. However, McIntyre & McKitrick noted that many of the proxies used by Mann et al., 1998 were very poorly correlated to local temperatures[A167]. Most of the U.S. tree ring proxies they used appeared to be better correlated to other factors, such as precipitation or CO_2 concentrations[A167].

Mann et al., 1998 were not overly concerned with how well individual proxies were correlated to local temperatures, and in fact several of the Mann et al., 1998 proxy series were actually precipitation weather records[A10][6]. Instead, they believed that their climate field reconstruction method ("MBH" in Section 4.2) would be able to detect *global* changes in climate patterns from their proxies. They pointed out that changes in local climate could sometimes also reflect more widespread climate change, via climate teleconnections, e.g., El Niño-Southern Oscillation (ENSO) variations[A10, A169, A170]. However, they did not offer a mechanism by which a proxy would be affected by global climate signals, but not by local climate signals, and this assumption seems to be at best unrealistic[B108, B109].

With this in mind, McIntyre & McKitrick decided to investigate Mann et al., 1998's claim that their hemispheric reconstruction had a "high level of skill' back to their earliest step (1400-1450). First, they considered standard statistical variables, such as R^2, the correlation coefficient of determination (also known as r^2), which we mentioned in Section 2.5. R^2 varies from 0 (non-correlated) to 1 (perfectly correlated). They found that the reconstructed temperatures showed a negligible correlation ($R^2 = 0.02$) to instrumental temperatures in the verification period[A44, A45] for that step.

It is true that a high R^2 value would not in itself indicate robustness. For instance, the verification data Mann et al., 1998 used (thermometer-based data for 1854-1901) consisted of an almost continuous global warming trend from start to finish. As a result the data was highly "autocorrelated". That is, values for early sections of the data appear to be "correlated to" later sections of the data, because the trend was similar (i.e., warming) over the entire period. In such cases, high R^2 values often occur spuriously[A108]. Hence, it is important to also consider other statistics. However, a negligible (or even low) R^2 value should have been a serious concern[B110].

Wahl & Ammann[A169, A170] claimed that the *hockey stick study* was more concerned with the long-term trends of the estimates being similar to the verification data, than in ensuring the annual temperatures were themselves accurate. They reckoned the most important issue was how the averages over the 18541901 verification period compared to the averages over the 1902-1980 calibration period. For this reason, they argued that the *hockey stick study* favoured a different statistic[A89, A147] - the *"reduction of error"* (RE, called "β" in Mann et al., 1998).

Nonetheless, McIntyre & McKitrick were also concerned with the RE results of the *hockey stick study*. Mann et al., 1998[A10] had arbitrarily decided that a non-zero value of RE indicated statistical significance. Hence, they believed that the $RE = 0.51$ value of the 1400-1450 step was statistically significant. However, McIntyre & McKitrick, 2005a[A44] found that red noise series (the ones they used in their PC1 simulations - see Section 4.3) which had no intrinsic climatic signal actually yielded higher RE values.

By assuming that the RE of a genuinely climatic series would have to be higher than 99% of the red noise series, they obtained a benchmark value of statistical significance of $RE = 0.59$. In other words, an apparently climatic series with an RE value of less than 0.59 would actually be no better than random noise. On that basis, the *hockey stick study's* 1400-1450 step was *not* statistically significant. It also failed other cross-validation statistical tests.

Huybers, 2005 criticised this benchmarking process, by pointing out that McIntyre & McKitrick had not scaled their red noise simulations to have the same variance as the calibration/verification data.

[5] They also considered the case without any principal component analysis as described above.

[6] Some of the precipitation records were seriously mislocated, e.g., Mann et al., 1998's "Boston" (U.S.A.) precipitation series actually appeared to be the series for Paris (France)[A42].

When Huybers did this, he calculated a benchmark of 0.0, i.e., the same as Mann et al., 1998 had assumed[A168].

McIntyre & McKitrick accepted this criticism, but noted that they also should have carried out a more complete emulation of the Mann et al., 1998 reconstruction in their benchmarking. In their initial study, they had only simulated *one* of the proxy series Mann et al., 1998 had used, i.e., the "PC1" series, while the actual 1400-1450 step used 22 series. Hence, they increased the variance of their red noise series by combining them with another 21 white noise pseudoproxies[B50]. This yielded a benchmark of $RE = 0.54$[A116], lower than their original analysis, but still higher than the 1400-1450 step.

Wahl & Ammann, 2007[A169, A170] initially claimed that they had obtained a benchmark of 0.0[A169, A170]. However, when they published their Supplementary Information, it transpired that they had actually calculated a benchmark of $RE = 0.52$[B111, B112] - only slightly lower than McIntyre & McKitrick's $RE = 0.54$. Moreover, there were also statistical problems with Wahl Ammann's lower value[B112, B113]. In any case, the more serious issue was still that it had a negligible R^2 statistic, while a robust estimate should have passed both tests[B112, B114].

4.5 Summary of the criticisms of the *hockey stick study*

To summarise, although the *hockey stick study* and its conclusions had a powerful impact on the scientific community[A40] and general public[B2], a number of serious flaws have since been found with it. In particular:

- Its characteristic "hockey stick" description of millennial temperature changes was highly dependent on the inclusion of a small set of problematic bristlecone/foxtail pine proxies (Section 3.4.1).
- By using a flawed approach to principal component analysis, the influence of these problematic proxies was dramatically increased (Section 4.3).
- Attempts to adjust these proxies to account for their non-climatic trends were themselves problematic (Section 4.1).
- The dataset used by Mann et al., 1998 was error-ridden and badly organised, reducing the study's reliability (Section 4.1).
- Its reconstruction method substantially underestimated actual temperature variability, making the "handle" of the "hockey stick" seem unrealistically flat (Section 4.2).
- It failed to pass basic tests of statistical significance (Section 4.4).

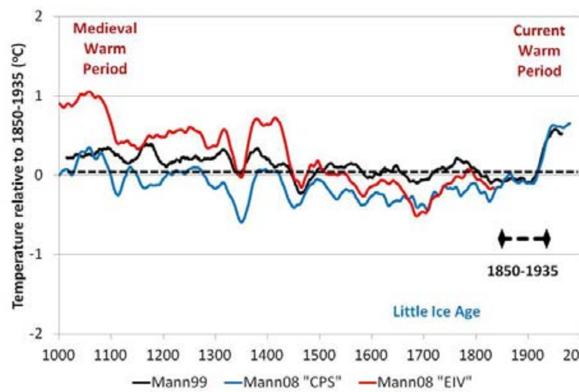

Figure 13: *Millennial temperature estimates of the* hockey stick study*[A11] compared to Mann et al.'s more recent 2008 "CPS" and "EIV" estimates[A22]. The Mann et al., 1999 and CPS estimates were rescaled and smoothed as described in Section 2.6, but the EIV estimate was only archived up to 1850, so was rescaled by assuming the estimate would have the same 1850-1935 mean and standard deviation as its calibration/verification data.*

Following the controversy over the *hockey stick study*, two independent reviews were carried out in the U.S. - one for the National Academy of Sciences ("the NAS Report")[A152] and one headed by a team of statisticians ("the Wegman Report")[A153]. The NAS Report partially agreed with some of the conclusions of the *hockey stick study*[A152][B115], i.e., that the Current Warm Period was warmer than the Little Ice Age. It also noted several studies which agreed with the *hockey stick study's* conclusion that the Current Warm Period is warmer than the Medieval Warm Period. However, both of the reports agreed with much of the criticism of the *hockey stick study*[A152, A153][B116].

The authors of the *hockey stick study* have been quite vocal in their insistence that the criticisms of Mann et al., 1998 and Mann et al., 1999 have all been countered or shown to be irrelevant[A15, A16, A82, A87, A89, A91, A94, A124, A147, A159, A161, A163][B44, B101, B102, B117]. However, their most recent millennial reconstructions[A22] actually show considerably more variability and uncertainty over the millennium than their *hockey stick study* - see Figure 13. This suggests that even Mann et al. probably now agree that the original *hockey stick study* was unreliable. Hence, in the next section, we will discuss the other millennial reconstructions.

5 Comparison between current reconstructions

5.1 Problems with the overlap period

A common mistake made when considering the current proxy-based temperature studies is to directly compare the *proxy-based* temperature estimates in the pre-instrumental period to the *thermometer-based* temperature estimates for the instrumental period. This can lead an unwary researcher into drawing conclusions which are unwarranted by the proxy-based estimates.

There are at least three major reasons why such comparisons should be avoided:

1. If the thermometer-based estimates are used for calibrating the proxies to temperature, then the calibration data can only be used for "training" the proxy-based estimates. If there are features of the calibration data that the estimates are unable to reproduce, then there is no reason to assume it would perform better outside of the training period. A serious danger with this training process which is often overlooked is that the proxy-based temperature estimates may then be affected by the "over-fitting" problem[A68]. This can be particularly problematic for estimates which rely on the apparent fit of a proxy to the training data for determining its "reliability"[B118–B120]. For example, the Mann et al., 2008[A22] approach yields different estimates depending on which calibration period is used - see Figure S10 of the supplementary information on Mann's website[B86].

2. Thermometer measurements are typically both physically and statistically different from the proxy measurements (e.g., tree ring widths, isotopic variations of an ice core), so a direct comparison between them should not be treated with any more seriousness than Sandford, 1995[B121]'s satirical "comparison" between apples and oranges. One approach that is sometimes taken to improve the statistical similarity of the proxies and the thermometer readings is to scale the proxy measurements so that both data sets have a similar variance (i.e., the inter-annual fluctuations are of a similar magnitude). However, it should be remembered that this does not actually improve the signal-to-noise ratio of the proxy measurements[7].

3. In general, the various proxy-based temperature estimates are only partially able to reproduce the trends and/or inter-annual variability of the thermometer-based estimates. In particular, depending on both the proxies used and the period to which the estimates are normalised, two fitting "problems" have been identified as causes for concern. In recent decades, many of the proxies (particularly tree ring widths and densities for high northern latitudes) suggest a *"divergence problem"*, in that they fail to show the strong warming suggested by the thermometer-based estimates, in some cases suggesting a cooling. For the early part of the calibration period (19th century), many of the estimates have a *"convergence problem"*, whereby they imply warming trends immediately following a very cold "Little Ice Age", while the thermometer-based estimates show a cooling trend. These two problems will be discussed separately in Sections 5.2 and 5.3.

The thermometer-based estimates of the Current Warm Period usually suggest warmer temperatures than the proxy-based estimates. However, the thermometer-based estimates only begin in the 19th century, and therefore do not provide estimates for the Medieval Warm Period. As a result, superimposing the thermometer-based estimates on top of the proxy-based estimates typically creates the impression that the Current Warm Period was much warmer than the Medieval Warm Period, even if the proxy-based estimates imply the two periods were just as warm. This is misleading, and therefore directly comparing the two types of estimates can lead to inaccurate conclusions.

Remarkably, this appears to be the sole basis on which the IPCC Working Group 1 made the following statement in their 2007 Summary for Policymakers:

[7] A somewhat outdated, but still useful, analogy is that of increasing the volume on a gramophone when listening to an old scratchy record. The volume of the signal may increase, but so does the volume of the noise.

> *"Palaeoclimatic information supports the interpretation that the warmth of the last half century is unusual in at least the previous 1,300 years."* - IPCC, 2007[A171]

And a similar statement in their 2013 Summary for Policymakers:

> *"In the Northern Hemisphere, 1983-2012 was likely the warmest 30-year period of the last 1400 years (medium confidence)."* - IPCC, 2013[A172]

In this article, such comparisons will not be made, and as will be seen in Section 5.4, this leads to more equivocal interpretations. But, it is nonetheless instructive to compare the proxy-based estimates to the thermometer-based estimates.

5.2 The "divergence problem"

Instead of showing the almost continuous warming trends of the thermometer-based global temperature estimates[A1], many tree ring proxies suggest there was cooling in the second half of the 20th century (at least until the 1980s, when most of the tree rings were collected). This divergence between the thermometer-based estimates and the tree ring-based estimates has come to be known as the *"divergence problem"*[A56, A173, A174][B122].

Jacoby & D'Arrigo, 1995[A175] first noted this for several samples of trees in Alaska. Briffa et al., 1998a & b[A176, A177] found the same phenomenon across much of the Northern Hemisphere, but they suggested that it was mostly confined to the more northerly regions.

From Figure 14, it can be seen that this divergence also exists between many of the proxy-based estimates and the thermometer-based estimates, although only 11 of the 19 proxy-based estimates actually consider temperatures after 1980 (see Table 1), and the Loehle, 2007 estimate finishes in 1935[A21, A51]. Many of the proxy-based estimates[8] reach a 20th century peak in the 1940s or 50s, and then show cooling until they finish. Indeed, the Moberg et al., 2005[A17] estimate actually has its peak 20th century temperatures in the 1920s, although it does imply that the 1940s and 1950s were still relatively warm. In contrast, although the CRUTEM3 thermometer-based estimates imply a *slight* cooling in the 1950s and 1960s, they suggest an almost continuous warming from the start of the 20th century.

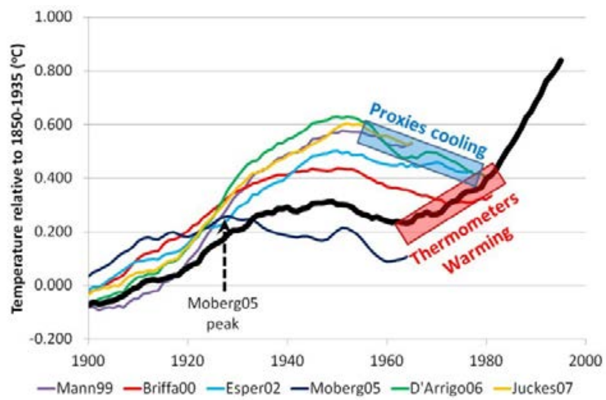

Figure 14: *Comparison of several proxy-based millennial temperature estimates to the thermometer-based CRUTEM3 estimate for the 20th century[A1, A2]. All plots are the 31 year running means of the archived data, rescaled following the procedure described in Section 2.6.*

As it is the post-1950s warming suggested by the thermometer-based estimates which is alleged to be due to "man-made global warming"[A172], the fact that it is not replicated by the proxy-based estimates is significant[9]. It raises the possibility that either (i) there are problems with the thermometer-based estimates, or (ii) proxy-based estimates are somehow unable to detect the recent warming. If the latter applies, then it is possible that the same could have happened during previous warming periods, e.g., during the Medieval Warm Period.

We argue elsewhere[B18–B21] that the apparent strong warming trends in the thermometer-based estimates of recent decades were mistakenly biased warm. Instead, we suggest that, since the late 19th century, there have been two relatively cool periods and two relatively warm periods, with "global warming" *and* "global cooling" between them, i.e., global cooling from the 1950s-1970s has been underestimated, while global warming since the 1980s has been overestimated.

If this is accurate, then the so-called divergence problem is not necessarily a proxy "problem". However, as mentioned in Section 2.3, most researchers constructing global or regional temperature proxy constructions have assumed (either implicitly or ex-

[8] See Supplementary Information for a graph of all 19 (smoothed) estimates in the period 1750-2000.

[9] Although, a few of the estimates do show some late-20th century warming - see Supplementary Information.

plicitly) that the thermometer-based temperature estimates are completely reliable. Therefore, they *assume* that the apparent divergence is a problem exclusively with the proxies.

On this basis, several researchers have even removed the "diverging" data or replaced it with thermometer-based estimates[B123]. One popular justification for doing that is that the divergence *might* be just a recent phenomenon, due to some kind of human activity[A176, A177] - see D'Arrigo et al, 2008[A173] for a brief summary of such theories. We find it difficult to see why such a speculative, untested (possibly untestable) hypothesis should be used as the sole basis for discarding a critical portion of the proxy-based temperature estimates[B124–B126].

Regardless, later research has suggested that the divergence problem is not as well-defined as originally proposed. A number of studies have found that even in regions where some trees show divergence, others can be found which do not show divergence[A173, A174, A178–A180][B126, B127]. These findings have been quite divisive in dendroclimatological circles, as can be seen by reading the review comments of Wilmking et al., 2008[B127].

From a reanalysis of their earlier work in Esper et al., 2002[A14], Cook et al., 2004[A112] agreed with Briffa et al.'s suggestion[A176] that the problem was real but limited to the more northerly regions. They came to this conclusion by dividing the sites used in the Esper reconstruction into two halves - north and south. The northern half showed a divergence, while the southern half did not. However, McIntyre pointed out[B125] that the southern half consisted of only 5 sites, and two of those sites were foxtail sites (a problem discussed in Section 3.4.1). Moreover, the chronologies constructed from the two subsets disagreed radically over the strength of the Medieval Warm Period, and the two subsets were poorly correlated to each other, although it seems that they did give reasonable correlation statistics over the specific 1200-1950 period Cook et al. chose.

More recently, Esper et al., 2010[A181] found that temperature-sensitive trees in Siberia were following local weather station temperature trends after all. They suggested that the divergence problem may simply have arisen from inappropriate tree ring standardisation approaches, and by not considering the possibility of errors/biases in the weather station records. Interestingly, the local weather station temperature trends in Siberia appear to have been quite modest compared to the warming trends of the global weather station-based temperature estimates.

5.3 The "convergence problem"

Another noteworthy discrepancy between the thermometer-based and proxy-based estimates occurs near the start of the thermometer-based estimates. From Figure 15, it can be seen that most of the proxy-based temperature estimates suggest strong global warming following "the Little Ice Age" during the 19th century. However, the first few years of the global thermometer-based estimates (solid black line) suggest global cooling.

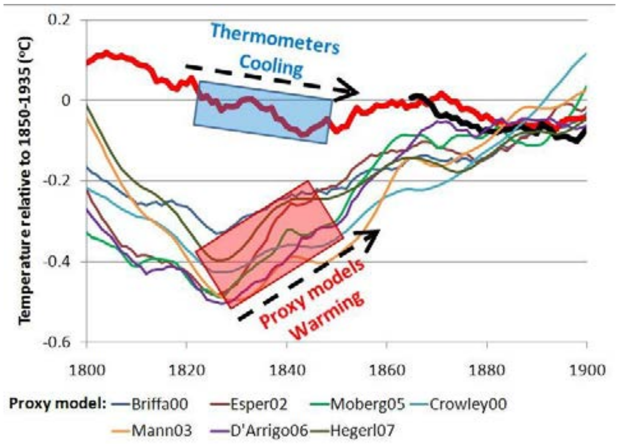

Figure 15: *Comparison of several proxy-based millennial temperature estimates to the thermometer-based global CRUTEM3[A1, A2] (thick black line) and Dobrovolný et al., 2010[A74, A115] "Central Europe" (thick red line) estimates for the 19th century. All plots are the 31 year running means of the archived data, rescaled following the procedure described in Section 2.6.*

While the early portion of the global thermometer-based estimate is too short a period to assess if this is significant or not, there are a few long instrumental records which cover a longer period, e.g., England[A182, A183], Sweden[A184, A185], Iceland[A35]. A few groups have combined some of these long records together to construct long European temperature estimates, e.g., Dobrovolný et al., 2010[A74, A115]. These longer records (thick red line in Figure 15) suggest the apparent *"convergence problem"* is significant.

For some of the proxy-based estimates, the convergence is less pronounced (see Supplementary Information), e.g., the *hockey stick study*[A10, A11] or Loehle, 2007[A21, A51]. However, it should be noted that some of the proxy-based estimates actually included long European thermometer records as "temperature proxies"[A10, A157], so this may have partially hidden the convergence problem for some of the studies.

Some researchers have suggested that inadequate thermometer exposure of the weather stations in the early part of their records led to the recorded measurements of the long records being too warm[A115, A186–A188]. For instance, early measurements were often recorded indoors in well-ventilated rooms, while later measurements were recorded outdoors in specially-designed instrument shelters[A182]. However, Dobrovolný et al., 2010[A115] believed that the station records from which they constructed their Central Europe estimates (the thick red line in Figure 15) had been adequately homogenised to account for the early exposure bias[A188][10].

We argue elsewhere[B18–B21] that there are a number of serious biases in the thermometer-based estimates for recent decades, including issues with station exposure[B19]. So, it is plausible that there are also biases for the earlier periods. However, as for the recent biases, it is a challenging problem to resolve. Hence, we should also consider the possibility that the proxy-based estimates may have exaggerated the apparent coldness of the Little Ice Age.

As an aside, it seems odd that researchers considering the divergence problem (Section 5.2) seem inclined to assume the proxies are at fault, while researchers considering the convergence problem seem inclined to assume the thermometer records are at fault. Perhaps we should be more prepared to recognise that there may be problems with either or even both of the data sets[A74] (a possibility Frank et al., 2007a[A187] do admittedly acknowledge).

5.4 Comparing and contrasting the 19 different estimates

In Figures 16, 17 and 18, all of the millennial proxy-based temperature estimates discussed in this article are plotted - rescaled and smoothed following the description in Section 2.6. One noteworthy difference between the plots in Figures 16-18 and other presentations of the data, e.g., that in the 2007 IPCC report[A49], is that thermometer-based estimates are

[10] They also believed the stations had been homogenised to account for urbanisation, but the *step* bias homogenisation that was used[A189] is often inadequate for dealing with *trend* biases, such as urbanisation[B21], so it is likely urbanisation bias still remains.

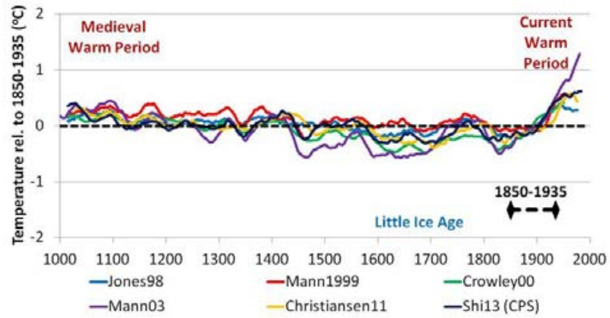

Figure 16: *The "hockey stick-like" proxy-based temperature estimates for the period 1000-2000, which suggest the Current Warm Period is unusually warm: Jones et al., 1998[A9]; "MBH99", i.e., the original* hockey stick study*[A11]; Crowley, 2000[A12, A111]; Mann & Jones, 2003[A15, A16]; Christiansen & Ljungqvist, 2011[A26]; and Shi et al., 2013 ("CPS")[A28]. All plots are rescaled and smoothed as described in Section 2.6.*

not superimposed over the plots. This is for the reasons discussed in Section 5.1.

Also, in most presentations until now, all estimates are usually shown superimposed on top of each other in hard-to-interpret, *"spaghetti graphs"*, e.g., the NAS 2006 report[A152], Figure 6.10b of the 2007 IPCC report[A49] or Figure 5.7 of the 2013 IPCC report[A190]. Spaghetti graphs have their name because the multiple overlapping curves on the graph resemble a tangled collection of spaghetti noodles. This makes it visually hard to follow each curve from start to finish, making it hard to compare and contrast individual estimates[B128]. For this reason, we have grouped the 19 different estimates into separate groups for better visual clarity. There appear to be three main groups of estimates - Figures 16-18.

5.4.1 "Hockey stick" estimates

The first group comprise those implying a "hockey stick"-type description of the last millennium[A9, A11, A12, A15, A16, A26, A28, A111] (Figure 16). These estimates suggest that the recent global warming of the Current Warm Period is highly unusual in the context of the last millennium.

None of the other estimates are quite as dramatic, however. They all suggest that there was a substantial Medieval Warm Period about a thousand years ago.

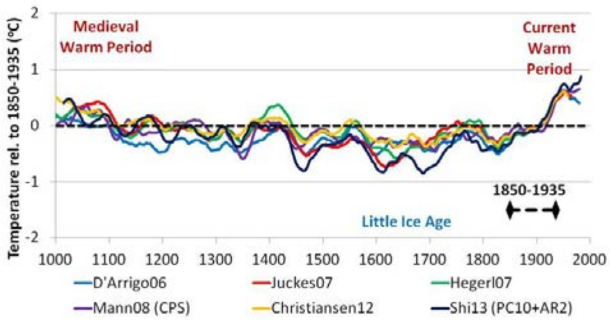

Figure 17: *The proxy-based temperature estimates for the period 1000-2000, which suggest the Current Warm Period is warmer than the Medieval Warm Period: D'Arrigo et al., 2006[A18]; Juckes et al., 2007[A20]; Hegerl et al., 2007[A19]; Mann et al., 2008 ("CPS" northern hemisphere estimate)[A22]; Christiansen & Ljungqvist, 2012[A27]; and Shi et al., 2013 ("PC10+AR2")[A28]. All plots are rescaled and smoothed as described in Section 2.6.*

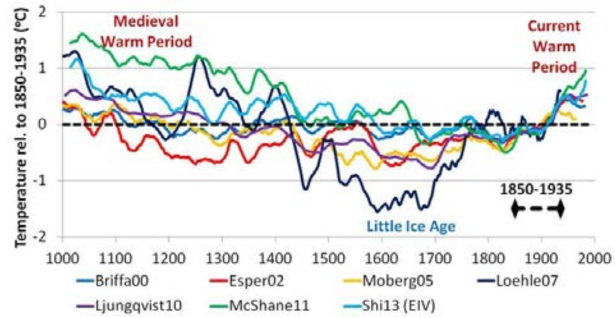

Figure 18: *The proxy-based temperature estimates for the period 1000-2000, which suggest the Medieval Warm Period was similar to or warmer than the Current Warm Period: Briffa, 2000[A13]; Esper et al., 2002[A14]; Moberg et al., 2005[A17]; Loehle, 2007[A21, A51]; Ljungqvist, 2010a[A24]; McShane & Wyner, 2011[A25]; and Shi et al., 2013 ("EIV")[A28]. All plots are rescaled and smoothed as described in Section 2.6.*

5.4.2 "Current Warm Period is warmer than Medieval Warm Period" estimates

Several of the estimates suggest that the Current Warm Period is warmer[A18–A20, A22, A27, A28] (Figure 17).

You might argue that Figure 17 supports the argument that at least some of the recent warming is "unusual". However, we do not see how you can claim that one of two warm periods is "unusual" and the other is "usual", merely because one is warmer than the other. Indeed, some of the estimates which cover more than the thousand years shown in Figures 16, 17 and 18, also suggest earlier warm periods, such as the so-called "Roman Warm Period" around two thousand years ago[A24]. Mayewski et al., 2004 suggest that there have been *several* pronounced global climatic changes over the last 12,000 years each of which could have lasted a few centuries[A191].

The Mann et al., 2008[A22] estimates came under particular criticism, as they had specifically claimed their estimates were robust to the exclusion of problematic proxy series or the use of different reconstruction methods, but this claim later transpired to be wholly inaccurate - see Section 3.4.3. In addition, their estimates appeared to be strongly affected by the over-fitting problem[A68][B118, B119] - see Figure S10 of Ref. [B86].

In 2011, two statisticians with no prior experience in palaeoclimate, McShane & Wyner, constructed their own estimates using Mann et al., 2008's dataset[A25]. Their analysis suggested that the wide variability of the proxy data (Section 3) meant that the necessary error bars were too great to definitively resolve the question of whether the Current Warm Period was warmer, colder or similar to the Medieval Warm Period. However, both periods did appear to be warmer than the Little Ice Age, and the mean values of their estimate *suggested* that the Medieval Warm Period was the warmer of the two.

The McShane & Wyner, 2011 study was published in a statistical journal as a discussion essay, and responses were sought from both statisticians and climate scientists (see links in Ref. [A25]). It also generated considerable discussion on various blogs (e.g., see Refs. [B129–B133] for some of the more considered discussion). There was a general impression that their analysis was weakened by a number of mistakes, misunderstandings and errors which could have been averted if they had collaborated with palaeoclimatologists. But, it was still considered useful, and there was considerable agreement with McShane & Wyner's recommendation that palaeoclimatologists should seek more advice from statisticians for future studies.

5.4.3 "Medieval Warm Period was comparable to Current Warm Period" estimates

We included McShane & Wyner, 2011's estimate in the third group of estimates[A13, A14, A17, A21, A24, A25, A28, A51, A114] which suggest that the Medieval Warm Period was comparable to, if not warmer than, the Current Warm Period (Figure 18).

Surprisingly, these estimates are often taken to imply the opposite conclusion[A49, A190]. This seems to happen when researchers incorrectly compare the proxy-based Medieval Warm Period estimates to the thermometer-based estimates for the Current Warm Period, rather than to the proxy-based Current Warm Period. As discussed in Section 5.1, this is inappropriate, and Ljungqvist, 2010[A24] correctly urged caution over such comparisons.

Mann & Hughes were critical of the Esper et al., 2002 estimate, as it disagreed with their *hockey stick study* and the other "hockey stick-like" estimates of Figure 16, leading to some debate[A192]. Esper et al. have been concerned about the robustness of the early part of their estimate, since it was only based on a small sample of trees. So, they have since revisited the study twice[A112, A113]. Each time, their reanalysis has slightly lowered their estimates of the warmth during the Medieval Warm Period. Hence, Frank et al., 2007[A113] now suggests that the Current Warm Period is a bit warmer than the Medieval Warm Period.

From pseudoproxy analysis, Mann et al., 2005[A82] suggested that Moberg et al., 2005[A17] was less reliable than the *hockey stick study*[A11]. However, other pseudoproxy analyses have suggested the opposite[A84, A97].

5.4.4 Differences between the various estimates

It is worth noting that (as discussed in Section 2.4) each of the three Shi et al., 2013 estimates fits into a different one of our three groups, i.e., the "CPS" estimate is in Figure 16; the "PC10+AR2" estimate is in Figure 17; and the "EIV" estimate is in Figure 18. All three of these estimates used the same proxy dataset, but different reconstruction methods. This suggests that at least some of the differences between the various estimates are purely statistical in nature.

Even though we grouped the 19 estimates into three separate figures (Figures 16-18) to avoid creating cluttered and confusing "spaghetti graphs"[B128], we note that there is *still* some "spaghetti"-nature to all three of the figures. This indicates that, while all 19 estimates agree on the approximate timing (and to a lesser extent, the magnitude) of the two warm periods, there is less agreement on the intervening periods.

For instance, some estimates suggest there was a (possibly brief) warm period around 1400 A.D., e.g., Briffa, 2000[A13]; Hegerl et al., 2007[A19]; Mann et al., 2008[A22]; Moberg et al., 2005[A17]. Indeed, when McIntyre & McKitrick, 2005b made some minor (yet plausible) substitutions to the *hockey stick study*, it implied that the 1400s were warmer than the Current Warm Period[A45]. However, in other estimates the 1400s were a relatively cold period.

Some estimates place the Little Ice Age at its coldest around 1600 A.D. In particular, the Loehle, 2007 estimate implies that global temperatures were more than $1°C$ colder in the 1600s than the 1850-1935 average[A21, A51]. However, other estimates, suggest minimum temperatures occurred during the 1800s, e.g., the *"hockey stick study"*[A10, A11]. This has significance for those arguing the apparent recovery from the Little Ice Age was due to increases in atmospheric CO_2 since the Industrial Revolution. If the "recovery"[A36] started in the 1600s, then that would have pre-dated the Industrial Revolution by a few centuries.

When we consider the lack of consistency between proxies (Section 3), as well as the fact that many researchers are pre-disposed to "finding" a Medieval Warm Period[A119], Little Ice Age[A117] and Current Warm Period[A172] in their data, all of us working with temperature proxy data should be acutely conscious of the possibility that we may be affected by "confirmation bias"[A118]. That is, we should be wary of prematurely accepting a particular peak or trough in our data as "accurate" because we were expecting something similar, and discarding another peak or trough as "unreliable" because we were expecting something different.

6 Conclusions and recommendations

In recent decades, there has been considerable interest[A5–A31] in statistically combining different temperature proxies (e.g., tree rings, ice cores, lake sediments) together to construct large-scale estimates of global (or at least hemispheric) temperature changes over the last millennium or so.

All 19 of the millennial proxy-based temperature estimates discussed in this review (Table 1) have identified at least three climatically distinct periods: two relatively warm periods - the "Current Warm Period" (c. 1900 AD on) and the "Medieval Warm Period" (c. 800-1200 AD), and a relatively cool period - the "Little Ice Age" (c.1500-1850 AD). Disagreement between estimates appears to be mainly limited to establishing exactly how much temperatures have differed between each of the periods (Section 5.4).

This might offer cause for optimism that we are close to reaching a reasonable understanding of temperature changes of the last millennium. However, unfortunately, much of the apparent agreement between estimates may be due to the substantial overlap in the proxy series used by the estimates (Section 3.4).

More worryingly, there seem to be a number of paradigms already accepted by many in the palaeoclimate community. Bradley & Jones, 1992[A117] and Hughes & Diaz, 1994[A119] warned of two such paradigms and their danger - the common belief that palaeoclimatologists should *expect* to find a "Little Ice Age"[A117] and "Medieval Warm Period"[A119] in their data. A third paradigm seems to have arisen in recent decades - that researchers should *expect* to find unusual recent warming due to man-made global warming.

This is not to imply that any of these paradigms are necessarily wrong - they may well be valid. However, if a researcher is expecting to find a particular result, it is quite possible that they will (in good faith) eventually "find" it, regardless of whether it actually occurred or not. This is why Konrad Lorenz (1903-1989) humorously suggested that: *"It is a good morning exercise for a research scientist to discard a pet hypothesis every day before breakfast. It keeps him young."*[B134]

Simmons et al., 2011[A70] have illustrated, by presenting the results of an intentionally nonsensical study, how confirmation bias can easily lead unwary researchers to reach false conclusions - see Nickerson, 1998 for a good review of the confirmation bias problem[A118]. As funding is rarely prioritised for attempting to reproduce earlier studies, these conclusions may then become embedded in the scientific literature.

We see enough contradictions in the current palaeoclimate data to suggest that the current paradigms should be treated cautiously, at the very least.

On the Little Ice Age paradigm: Since Bradley & Jones, 1993[A7], there seems to have been a general consensus that there was a period of several centuries before the Current Warm Period that was particularly cold. It has even been suggested that current estimates are underestimating this coldness[A24].

However, the existence of the "convergence problem" (Section 5.3) suggests that if there was a Little Ice Age, it might not have been that long or cold after all. In other words, the uncertainties over exactly how long and cold it was[A35] remain.

On the Medieval Warm Period paradigm: There is considerable inconsistency in the estimates of the "Medieval Warm Period" (in terms of time and extent) between different proxy series, even for the same area (Section 3). Unless the reasons for these differences can be satisfactorily resolved, and it can be objectively established which series (if any) are reliable, considerable uncertainties will remain.

On the unusual recent global warming paradigm: Much of the "unusual" 20th century temperatures implied by several proxy-based estimates seems to depend on the inclusion of particularly controversial proxy series, i.e., the Yamal chronology or bristlecone/foxtail series (Section 3.4). If such trends are genuinely climatic then they should not be dependent on the inclusion of particular series.

In addition, most proxy-based estimates do not show the strong global warming of recent decades suggested by the thermometer-based estimates (Section 5.2). This suggests that either there are problems with the thermometer-based estimates (something we discuss elsewhere[B18–B21]), or the proxy-based estimates are unable to detect recent warming, in which case it is plausible that they might have also missed earlier warm periods.

However, there are also other significant contradictions between estimates, which need to be investigated. In Section 5.4, we saw that some estimates suggest temperatures in the 15th century may have been relatively warm, or at least mild. But this is not shown in other estimates. This suggests an ambiguity. Indeed, McIntyre & McKitrick noted[A45] that the *hockey stick study's* conclusion that 15th century temperatures were colder than the 20th century could be reversed with relatively minor and reasonable alterations to the study (Sections 4.1 and 4.3).

We realise that this review has been highly critical of many aspects of paleoclimate research, as currently practised. Therefore, it is important to stress that our aim is *not* to discredit this important field, but rather to suggest how it can be improved, so that future studies will be more meaningful. We are of the opinion that the first stage in dealing with problems is often to recognise the existence of those problems. These problems might initially seem intimidating, and lead researchers to take premature short-cuts and assumptions, in the hopes of getting a quick answer. However, we think it is better to aim for more reliable answers, even if it requires more effort.

We should recognise that estimating climatic conditions of the last millennium or so, is a very challenging research problem. We believe the following ten recommendations could help:

1. Rigorous research into testing and validating the theoretical basis behind individual temperature proxies should be a top priority (Sections 2.1-2.3).

2. Pseudoproxy analysis offers a useful approach to assessing and devising the various reconstruction methods, and further research along these lines should be encouraged. However, we should remember that this type of analysis only allows a negative check. That is, if a reconstruction method fails a pseudoproxy test, this shows it is unreliable, but "passing" the test does *not* prove that it is reliable (Section 2.4).

3. Many researchers have mistakenly assumed that it is a good idea to "screen" or "weight" the proxies in a proxy dataset on the basis of how well they correlate to the thermometer-based estimates. However, counter-intuitively, such "data-mining" actually makes the estimates *less* reliable. This practice should be abandoned (Section 2.5).

4. Serious inconsistencies exist between many of the individual proxy series used (Section 3). Research into understanding and quantifying these inconsistencies should be a high priority.

5. Rigorous proxy substitution and sensitivity experiments should be a routine requirement for all future proxy-based temperature reconstructions (Section 3.1). In particular, if a reconstruction is heavily reliant on the inclusion of one or two series, e.g., Yamal, bristlecone/foxtail series or the Tiljander proxies, then it is *not* a reliable reconstruction (Section 3.4).

6. There is a shortage of available proxies with long records covering at least the last millennium. However, the common practice of relying on short proxy records instead, is misleading. A proxy whose record begins in 1600 might give us some information about the temperatures of the last four hundred years, but does not by itself tells us anything about temperatures in 1000. So, the development of new, long proxy records should be encouraged (Section 3.2).

7. Many of the proxy records which have been used by the proxy-based global temperature trend estimates have been poorly documented, and in a number of cases the data has not been publicly archived (Section 3.3). In addition, research into the "file-drawer problem"[A130–A132] suggests that in many fields researchers are reluctant to publish "null result" findings[A132]. The developers of proxy records should be actively encouraged to publicly document and archive the results of their research, even if they believe their findings are "null results" or they might not "get a paper out of it".

8. The original *"hockey stick study"* by Mann et al., 1998[A10] and Mann et al., 1999[A11] has been shown to have had numerous serious flaws (Section 4). We appreciate that the debate over this high profile study has become highly politicised. However, it is 15 years later, and even its authors (Mann, Bradley and Hughes) have switched to using a different estimate (i.e., Mann et al., 2008[A22] - see Figure 13). So, it is probably time to acknowledge these flaws and move on.

9. There are significant discrepancies between thermometer-based temperature estimates and the proxy-based estimates, e.g., the "divergence problem" (Section 5.2) and the "convergence problem" (Section 5.3). Therefore, we should stop treating the two types of estimates as directly comparable, e.g., the common practice of superimposing the two types of estimate on the same plot and treating one as an extension of the other should be discontinued (Section 5.1).

10. The insidious problem of "confirmation bias" plagues most fields of scientific research[A118], and paleoclimate is no exception. In particular, there is evidence that many researchers *expect* to "find" a "Medieval Warm Period"[A119], a "Little Ice Age"[A117] and/or a "Current Warm Period"[A172] in their data. But, if we already

knew with confidence what the global temperature trends of the last millennium have been, then the 19 different proxy-based estimates reviewed in this paper would not have been carried out. So, let us actively try to avoid letting these expectations influence our analysis:

"If a man will begin with certainties, he shall end in doubts: but if he will be content to begin with doubts, he shall end in certainties" - Francis Bacon, Sr. (1561-1626)[B135]

Acknowledgements

No funding was received for this study.

We thank Don Zieman and Dr. Anton O'Connor for some useful comments and suggestions on an early draft. We also thank Dr. Craig Loehle and Willis Eschenbach for their constructive reviews of version 0.1. In addition, we would like to thank the other readers of version 0.1 who offered us helpful feedback.

References

Those references which have gone through a "peer review process", e.g., journal articles, are prefixed by "A", while those which have not, e.g., blog posts, are prefixed by "B".

Peer reviewed

[A1] P. D. Jones and A. Moberg. "Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001". *J. Clim.* 16 (2003), pp. 206–223. DOI: 10.1175/1520-0442(2003)016<0206:HALSSA>2.0.CO;2.

[A2] P. Brohan et al. "Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850". *J. Geophys. Res.* 111 (2006), p. D12106. DOI: 10.1029/2005JD006548.

[A3] Y. Aono and S. Saito. "Clarifying springtime temperature reconstructions of the medieval period by gap-filling the cherry blossom phenological data series at Kyoto, Japan". *Int. J. Biometeorol.* 54 (2010), pp. 211–219. DOI: 10.1007/s00484-009-0272-x.

[A4] B. A. S. Davis et al. "The temperature of Europe during the Holocene reconstructed from pollen data". *Quat. Sci. Rev.* 22 (2003), pp. 1701–1716. DOI: 10.1016/S0277-3791(03)00173-2.

[A5] H. H. Lamb. "The early medieval warm epoch and its sequel". *Palaeogeo. Palaeoclim. Palaeoecol.* 1 (1965), pp. 13–19. DOI: 10.1016/0031-0182(65)90004-0.

[A6] B. S. Groveman and H. E. Landsberg. "Simulated Northern Hemisphere temperature departures 1579-1880". *Geophys. Res. Lett.* 6 (1979), pp. 767–769. DOI: 10.1029/GL006i010p00767.

[A7] R. S. Bradley and P. D. Jones. ""Little Ice Age" summer temperature variations: their nature and relevance to recent global warming trends". *Holocene* 3 (1993), pp. 367–376. DOI: 10.1177/095968369300300409.

[A8] J. Overpeck et al. "Arctic environmental change of the last four centuries". *Science* 278 (1997), pp. 1251–1256. DOI: 10.1126/science.278.5341.1251.

[A9] P. D. Jones et al. "High-resolution palaeoclimatic records for the last millennium: Interpretation, integration and comparison with general circulation model control-run temperatures". *Holocene* 8 (1998), pp. 455–471. DOI: 10.1191/095968398667194956.

[A10] M. E. Mann, R. S. Bradley, and M. K. Hughes. "Global-scale temperature patterns and climate forcing over the past six centuries". *Nature* 392 (1998), pp. 778–787. DOI: 10.1038/33859.

[A11] M. E. Mann, R. S. Bradley, and M. K. Hughes. "Northern hemisphere temperatures during the past millenium: Inferences, uncertainties, and limitations". *Geophys. Res. Lett.* 26 (1999), pp. 759–762. DOI: 10.1029/1999GL900070.

[A12] T. J. Crowley and T. S. Lowery. "How warm was the Medieval Warm Period?" *Ambio* 29 (2000), pp. 51–54. DOI: 10.1579/0044-7447-29.1.51.

[A13] K. R. Briffa. "Annual climate variability in the Holocene: interpreting the message of ancient trees". *Quat. Sci. Rev.* 19 (2000), pp. 87–105. DOI: 10.1016/S0277-2791(99)00056-6.

[A14] J. Esper, E. R. Cook, and F. H. Schweingruber. "Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability". *Science* 295 (2002), pp. 2250–2253. DOI: 10.1126/science.1066208.

[A15] M. E. Mann and P. D. Jones. "Global surface temperatures over the past two millennia". *Geophys. Res. Lett.* 30 (2003), p. 1820. DOI: 10.1029/2003GL017814.

[A16] P. D. Jones and M. E. Mann. "Climate over past millennia". *Rev. Geophys.* 42 (2004), RG2002. DOI: 10.1029/2003RG000143.

[A17] A. Moberg et al. "Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data". *Nature* 433 (2005), pp. 613–617. DOI: 10.1038/nature03265.

[A18] R. D'Arrigo, R. Wilson, and G. Jacoby. "On the long-term context for late 20th century warming". *J. Geophys. Res.* 111 (2006), p. D03103. DOI: 10.1029/2005JD006352.

[A19] G. C. Hegerl et al. "Detection of human influence on a new, validated 1500-year temperature reconstruction". *J. Clim.* 20 (2007), pp. 650–666. DOI: 10.1175/JCLI4011.1.

[A20] M. N. Juckes et al. "Millennial temperature reconstruction intercomparison and evaluation". *Clim. Past* 3 (2007), pp. 591–609. DOI: 10.5194/cp-3-591-2007.

[A21] C. Loehle. "A 2000-year global temperature reconstruction based on non-treering proxies". *Energy Environ.* 18 (2007), pp. 1049–1058. DOI: 10.1260/095830507782616797.

[A22] M. E. Mann et al. "Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millenia". *Proc. Nat. Acad. Sci.* 105 (2008), pp. 13252–13257. DOI: 10.1073/pnas.0805721105.

[A23] D. S. Kaufman et al. "Recent warming reverses long-term Arctic cooling". *Science* 325 (2009), pp. 1236–1239. DOI: 10.1126/science.1173983.

[A24] F. C. Ljungqvist. "A new reconstruction of temperature variability in the extra-tropical northern hemisphere during the last two millennia". *Geograf. Ann. A* 92 (2010), pp. 339–351. DOI: 10.1111/j.1468-0459.2010.00399.

[A25] B. B. McShane and A. J. Wyner. "A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable?" *Ann. Appl. Stat.* 5 (2011). Discussion paper which led to 15 comments and/or discussion papers for Issue 1 of Vol. 5 of Ann. Appl. Stat., pp. 5–44. DOI: 10.1214/10-AOAS398. URL: http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&page=toc&handle=euclid.aoas/1300715166.

[A26] B. Christiansen and F. C. Ljungqvist. "Reconstruction of the extratropical NH mean temperature over the last millennium with a method that preserves low-frequency variability". *J. Clim.* 24 (2011), pp. 6013–6034. DOI: 10.1175/2011JCLI4145.1.

[A27] B. Christiansen and F. C. Ljungqvist. "The extra-tropical NH temperature in the last two millennia: reconstructions of low-frequency variability". *Clim. Past* 8 (2012), pp. 765–786. DOI: 10.5194/cp-8-765-2012.

[A28] F. Shi et al. "Northern Hemisphere temperature reconstruction during the last millennium using multiple annual proxies". *Clim. Res.* 56 (2013), pp. 231–244. DOI: 10.3354/cr01156.

[A29] PAGES 2k Consortium. "Continental-scale temperature variability during the past two millennia". *Nature Geosci.* 6 (2013), pp. 339–346. DOI: 10.1038/NGEO1797.

[A30] S. Hanhijärvi, M. P. Tingley, and A. Korhola. "Pairwise comparisons to reconstruct mean temperature in the Arctic Atlantic region over the last 2,000 years". *Clim. Dyn.* 41 (2013), pp. 2039–2060. DOI: 10.1007/s00382-013-1701-4.

[A31] F. C. Ljungqvist et al. "Northern Hemisphere temperature patterns in the last 12 centuries". *Clim. Past* 8 (2012), pp. 227–249. DOI: 10.5194/cp-8-227-2012.

[A32] F. C. Ljungqvist. "Temperature proxy records covering the last two millennia: A tabular and visual overview". *Geograf. Ann. A* 91 (2009), pp. 11–29. DOI: 10.1111/j.1468-0459.2009.00350.x.

[A33] F. C. Ljungvist. "A regional approach to the Medieval Warm Period and the Little Ice Age". In: *Climate change and variability*. Ed. by S. Simard. ISBN 978-953-307-144-2. InTech, 2010. URL: http://www.intechopen.com/articles/show/title/a-regional-approach-to-the-medieval-warm-period-and-the-little-ice-age.

[A34] D. Frank et al. "A noodle, hockey stick, and spaghetti plate: a perspective on high-resolution paleoclimatology". *WIRES Clim. Change* 1 (2010), pp. 507–516. DOI: 10.1002/wcc.53.

[A35] A. E. J. Ogilvie and T. Jónsson. ""Little Ice Age" research: A perspective from Iceland". *Clim. Change* 48 (2001), pp. 9–52. DOI: 10.1023/A:1005625729889.

[A36] S.-I. Akasofu. "On the recovery from the Little Ice Age". *Nat. Sci.* 2 (2010), pp. 1211–1224. DOI: 10.4236/ns.2010.211149.

[A37] W. Soon and S. Baliunas. "Proxy climatic and environmental changes of the past 1000 years". *Clim. Res.* 23 (2003), pp. 89–110. DOI: 10.3354/cr023089.

[A38] W. Soon et al. "Reconstructing Climatic and Environmental Changes of the Past 1000 Years: A Reappraisal". *Energy Environ.* 14 (2003), pp. 233–296. DOI: 10.1260/095830503765184619.

[A39] W. Broecker. "Was the Medieval Warm Period global?" *Science* 291 (2001), pp. 1497–1499. DOI: 10.1126/science.291/5508.1497.

[A40] C. K. Folland et al. "2. Observed climate variability and change". In: *Climate change 2001: The scientific basis. Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change.* Ed. by J. T. Houghton et al. Cambridge University Press. Cambridge. New York., 2001, 881pp. URL: http://www.ipcc.ch/.

[A41] K. R. Briffa and T. J. Osborn. "Seeing the wood from the trees". *Science* 284 (1999), pp. 926–927. DOI: 10.1126/science.284.5416.926.

[A42] S. McIntyre and R. McKitrick. "Corrections to the Mann et. al. (1998) proxy data base and northern hemispheric average temperature series". *Energy Environ.* 14 (2003), pp. 751–771. DOI: 10.1260/095830503322793632.

[A43] H. von Storch et al. "Reconstructing past climate from noisy data". *Science* 306 (2004), pp. 679–682. DOI: 10.1126/science.1096109.

[A44] S. McIntyre and R. McKitrick. "Hockey sticks, principal components, and spurious significance". *Geophys. Res. Lett.* 32 (2005), p. L03710. DOI: 10.1029/2004GL021750.

[A45] S. McIntyre and R. McKitrick. "The M&M critique of the MBH98 northern hemisphere climate index: update and implications". *Energy Environ.* 16 (2005), pp. 69–100. DOI: 10.1260/0958305053516226.

[A46] G. Bürger and U. Cubasch. "Are multiproxy climate reconstructions robust?" *Geophys. Res. Lett.* 32 (2005), p. L23711. DOI: 10.1029/2005GL024155.

[A47] G. Bürger, I. Fast, and U. Cubasch. "Climate reconstruction by regression - 32 variations on a theme". *Tellus A* 58 (2006), pp. 227–235. DOI: 10.1111/j.1600-0870.2006.00164.x.

[A48] G. Bürger. "On the verification of climate reconstructions". *Clim. Past* 3 (2007), pp. 397–409. DOI: 10.5194/cp-3-397-2007.

[A49] E. Jansen et al. "6. Palaeoclimate". In: *Climate change 2007: The physical science basis. Contribution of Working Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* Ed. by S. Solomon et al. Cambridge University Press. Cambridge. New York., 2007, 1056pp. URL: http://www.ipcc.ch/.

[A50] P. D. Jones et al. "High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects". *Holocene* 19 (2009), pp. 3–49. DOI: 10.1177/0959683608098952.

[A51] C. Loehle and J. H. McCulloch. "Correction to: A 2000-year global temperature reconstruction based on non-treering proxies". *Energy Environ.* 19 (2008), pp. 93–100. URL: http://www.econ.ohio-state.edu/jhm/AGW/Loehle/.

[A52] D. R. Anderson et al. "Concerns about finding effects that are actually spurious". *Wildlife Soc. Bull.* 29 (2001), pp. 311–316. URL: http://www.jstor.org/stable/3784014.

[A53] W. E. Ferson, S. Sarkissian, and T. T. Simin. "Spurious regressions in financial economics?" *J. Finance* 58 (2003), pp. 1393–1414. DOI: 10.1111/1540-6261.00571.

[A54] D. Ventosa-Santaulària. "Spurious regression". *J. Prob. Stat.* 2009 (2009), p. ID 802975. DOI: 10.1155/2009/802975.

[A55] D. Frank et al. "A challenge for spatially explicit reconstructions: the climate response of trees is a function of climate". In: *TRACE - Tree Rings in Archaeology, Climatology and Ecology.* Ed. by D. Elferts et al. Vol. 6. GFZ Potsdam, Scientific Technical Report STR 08/05, Potsdam, pp. 31–36.

[A56] A. H. Lloyd and A. G. Bunn. "Responses of the circumpolar boreal forest to 20th century climate variability". *Environ. Res. Lett.* 2 (2007), p. 045013. DOI: 10.1088/1748-9326/2/4/045013.

[A57] R. S. Bradley. "Dendroclimatology". In: *Paleoclimatology: Reconstructing climates of the Quaternary.* 2nd ed. Harcourt Academic Press. San Diego. London, 1999, 613pp.

[A58] S. Helama et al. "Detection of climate signal in dendrochronological data analysis: a comparison of tree-ring standardization methods". *Theor. Appl. Clim.* 79 (2004), pp. 239–254. DOI: 10.1007/s00704-004-0077-0.

[A59] K. R. Briffa and T. M. Melvin. "A closer look at Regional Curve Standardization of tree-ring records: Justification of the need, a warning of some pitfalls, and suggested improvements in its application". *Dendroclim. Devel. Paleoenviron. Res.* 11 (2011), pp. 112–145. DOI: 10.1007/978-1-4020-5725-0_5.

[A60] J. Esper et al. "Climate signal age effects - Evidence from young and old trees in the Swiss Engadin". *Forest Ecol. Manag.* 255 (2008), pp. 3783–3789. DOI: 10.1016/j.foreco.2008.03.015.

[A61] H. Visser et al. "Detecting instabilities in tree-ring proxy calibrations". *Clim. Past* 6 (2010), pp. 367–377. DOI: 10.5194/cp-6-367-2010.

[A62] M. Gagen et al. "Exorcising the 'segment length curse': summer temperature reconstruction since AD 1640 using non-detrended stable carbon isotope ratios from pine trees in northern Finland". *Holocene* 17 (2007), pp. 435–446. DOI: 10.1177/0959683607077012.

[A63] J. Esper et al. "Tests of the RCS method for preserving low-frequency variability in long tree-ring chronologies". *Tree-Ring Res.* 59 (2003), pp. 81–98. URL: http://www.treeringsociety.org/TRBTRR/TRRvol59_2_81-98.pdf.

[A64] T. M. Melvin, H. Grudd, and K. R. Briffa. "Potential bias in 'updating' tree-ring chronologies using regional curve standardisation: Re-processing 1500 years of Torneträsk density and ring-width data". *Holocene* 23 (2012), pp. 364–373. DOI: 10.1177/0959683612460791.

[A65] T. M. Melvin and K. R. Briffa. "CRUST: Software for the implementation of Regional Chronology Standardisation: Part 1. Signal-Free RCS". *Dendrochronologia* 32 (2014), pp. 7–20. DOI: 10.1016/j.dendro.2013.06.002.

[A66] B. Yang et al. "CRUST: Software for the implementation of Regional Chronology Standardisation: Part 1. Signal-Free RCS".

Theor. Appl. Climatol. 106 (2011), pp. 489–497. DOI: 10.1007/s00704-011-0451-7.

[A67] C. Loehle. "A mathematical analysis of the divergence problem in dendroclimatology". *Clim. Change* 94 (2009), pp. 233–245. DOI: 10.1007/s10584-008-9488-8.

[A68] M. A. Babyak. "What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models". *Psychosom. Med.* 66 (2004), pp. 411–421. DOI: 10.1097/01.psy.0000127692.23278.a9.

[A69] D. M. Hawkins. "The problem of overfitting". *J. Chem. Inf. Comput. Sci.* 44 (2004), pp. 1–12. DOI: 10.1021/ci0342472.

[A70] J. P. Simmons, L. D. Nelson, and U. Simonsohn. "False-positive psychology: Undisclosed felxibility in data collection and analysis allows presenting anything as significant". *Psych. Sci* 20 (2011), pp. 1–8. DOI: 10.1177/0956797611417632.

[A71] S.-E. Lauritzen and J. Lundberg. "Calibration of the speleothem delta function: an absolute temperature record for the Holocene in northern Norway". *Holocene* 9 (1999), pp. 659–669. DOI: 10.1191/095968399667823929.

[A72] F. McDermott. "Palaeo-climate reconstruction from stable isotope variations in speleothems: a review". *Quat. Sci. Rev.* 23 (2004), pp. 901–918. DOI: 10.1016/j.quascirev.2003.06.021.

[A73] J. Weckström et al. "Temperature patterns over the past eight centuries in Northern Fennoscandia inferred from sedimentary diatoms". *Quat. Res.* 66 (2006), pp. 78–86. DOI: 10.1016/j.yqres.2006.01.005.

[A74] A. Moberg and G. Brattström. "Prediction intervals for climate reconstructions with autocorrelated noise - An analysis of ordinary least squares and measurement error methods". *Palaeogeog. Palaeoclim. Palaeoecol.* 306 (2011), pp. 313–329. DOI: 10.1016/j.palaeo.2011.05.035.

[A75] T. J. Osborn and K. R. Briffa. "The real color of climate change?" *Science* 306 (2004), pp. 621–622. DOI: 10.1126/science.1104416.

[A76] T. C. K. Lee, F. W. Zwiers, and M. Tsao. "Evaluation of proxy-based millennial reconstruction methods". *Clim. Dyn.* 31 (2008), pp. 263–281. DOI: 10.1007/s00382-007-0351-9.

[A77] E. A. Rexstad et al. "Questionable multivariate statistical inference in wildlife habitat and community studies". *J. Wildlife Management* 52 (1988), pp. 794–798. URL: http://www.jstor.org/stable/3800948.

[A78] J. Taylor. "Questionable multivariate statistical inference in wildlife habitat and community studies: A comment". *J. Wildlife Management* 54 (1990), pp. 186–189. URL: http://www.jstor.org/stable/3808920.

[A79] E. A. Rexstad et al. "Questionable multivariate statistical inference in wildlife habitat and community studies: A reply". *J. Wildlife Management* 54 (1990), pp. 189–193. URL: http://www.jstor.org/stable/3808921.

[A80] F. C. James and C. E. McCulloch. "Multivariate analysis in ecology and systematics: Panacea or Pandora's box?" *Ann. Rev. Ecol. Systematics* 21 (1990), pp. 129–166. URL: http://www.jstor.org/stable/2097021.

[A81] E. Zorita, F. González-Rouco, and S. Legutke. "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 (2003), pp. 1378–1390. DOI: 10.1175/1520-0442(2003)16<1378:TTMEAA>2.0.CO;2.

[A82] M. E. Mann et al. "Testing the fidelity of methods used in proxy-based reconstructions of past climate". *J. Clim.* 18 (2005), pp. 4097–4107. DOI: 10.1175/JCLI3564.1.

[A83] E. R. Wahl, D. M. Ritson, and C. M. Ammann. "Comment on "Reconstructing past climate from noisy data"". *Science* 312 (2006), p. 529. DOI: 10.1126/science.1120866.

[A84] E. Zorita and H. von Storch. "Methodical aspects of reconstructing non-local historical temperatures". *Mem. Soc. Astron. Ital.* 76 (2005), pp. 794–801. URL: http://sait.oat.ts.astro.it/MSAIt760405/index.html.

[A85] H. von Storch et al. "Response to comment on "Reconstructing past climate from noisy data"". *Science* 312 (2006), p. 529. DOI: 10.1126/science.1121571.

[A86] E. Zorita, F. Gonzalez-Rouco, and H. von Storch. "Comments on "Testing the fidelity of methods used in proxy-based reconstructions of past climate"". *J. Clim.* 20 (2007), pp. 3693–3698. DOI: 10.1175/JCLI4171.1.

[A87] M. E. Mann et al. "Reply". *J. Clim.* 20 (2007), pp. 3699–3703. DOI: 10.1175/JCLI4172.1.

[A88] J. E. Smerdon and A. Kaplan. "Comments on "Testing the fidelity of methods used in proxy-based reconstructions of past climate": The role of the standardization interval". *J. Clim.* 20 (2007), pp. 5666–5670. DOI: 10.1175/JCLI1794.1.

[A89] M. E. Mann et al. "Robustness of proxy-based climate field reconstruction methods". *J. Geophys. Res.* 112 (2007), p. D12109. DOI: 10.1029/2006JD008272.

[A90] J. E. Smerdon, González-Rouco, and E. Zorita. "Comment on "Robustness of proxy-based climate field reconstruction methods"". *J. Geophys. Res.* 113 (2008), p. D18106. DOI: 10.1029/2007JD009542.

[A91] S. Rutherford et al. "Reply to comment by Jason E. Smerdon et al. on "Robustness of proxy-based climate field reconstruction methods"". *J. Geophys. Res.* 113 (2008), p. D18107. DOI: 10.1029/2008JD009964.

[A92] J. E. Smerdon, A. Kaplan, and D. Chang. "On the origin of the standardization sensitivity in RegEM climate field reconstructions". *J. Clim.* 21 (2008), pp. 6710–6723. DOI: 10.1175/2008JCLI2182.1.

[A93] B. Christiansen, T. Schmith, and P. Thejll. "A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness". *J. Clim.* 22 (2009), pp. 951–976. DOI: 10.1175/2008JCLI2301.1.

[A94] S. D. Rutherford et al. "Comments on "A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness"". *J. Clim.* 23 (2010), pp. 2832–2838. DOI: 10.1175/2009JCLI3146.1.

[A95] B. Christiansen, T. Schmith, and P. Thejll. "Reply to comment on "A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness"". *J. Clim.* 23 (2010), pp. 2839–2844. DOI: 10.1175/2010JCLI3281.1.

[A96] H. von Storch, E. Zorita, and J. F. González-Rouco. "Assessment of three temperature reconstruction methods in the virtual reality of a climate simulation". *Int. J. Earth Sci. (Geol. Rundsch.)* 98 (2009), pp. 67–82. DOI: 10.1007/s00531-008-0349-5.

[A97] A. Moberg, R. Mohammad, and T. Mauritsen. "Analysis of the Moberg et al. (2005) hemispheric temperature reconstruction". *Clim. Dyn.* 31 (2008), pp. 957–971. DOI: 10.1007/s00382-008-0392-8.

[A98] N. Riedwyl et al. "Comparison of climate field reconstruction techniques: application to Europe". *Clim. Dyn.* 32 (2009), pp. 381–395. DOI: 10.1007/s00382-008-0395-5.

[A99] B. Li, D. W. Nychka, and C. M. Ammann. "The value of multiproxy reconstruction of past climate". *J. Am. Stat. Assoc.* 105 (2010), pp. 883–911. DOI: 10.1198/jasa.2010.ap09379.

[A100] M. P. Tingley and P. Huybers. "A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: Development and applications to paleoclimate reconstruction problems". *J. Clim.* 23 (2010), pp. 2759–2781. DOI: 10.1175/2009JCLI3015.1.

[A101] M. P. Tingley and P. Huybers. "A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: Development and applications to paleoclimate reconstruction problems". *J. Clim.* 23 (2010), pp. 2782–2800. DOI: 10.1175/2009JCLI3016.1.

[A102] C. M. Ammann, M. G. Genton, and B. Li. "Technical note: Correcting for signal attenuation from noisy proxy data in climate reconstructions". *Clim. Past.* 6 (2010), pp. 273–279. DOI: 10.5194/cp-6-273-2010.

[A103] B. Christiansen. "Reconstructing the NH mean temperature: Can underestimation of trends and variability be avoided?" *J. Clim.* 24 (2011), pp. 674–692. DOI: 10.1175/2010JCLI3646.1.

[A104] J. D. Annan and J. C. Hargreaves. "Identification of climatic state with limited proxy data". *Clim. Past* 8 (2012), pp. 1141–1151. DOI: 10.5194/cp-8-1141-2012.

[A105] J. E. Smerdon. "Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments". *WIREs Clim. Change* 3 (2012), pp. 63–77. DOI: 10.1002/wcc.149.

[A106] N. J. Steiger et al. "Assimiliation of time-averaged pseudoproxies for climate reconstruction". *J. Clim.* 27 (2014), pp. 426–441. DOI: 10.1175/JCLI-D-12-00693.1.

[A107] G. U. Yule. "Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series". *J. Roy. Stat. Soc.* 89 (1926), pp. 1–63. URL: http://www.jstor.org/stable/2341482.

[A108] C. W. J. Granger and P. Newbold. "Spurious regressions in econometrics". *J. Econometrics* 2 (1974), pp. 111–120. DOI: 10.1016/0304-4076(74)90034-7.

[A109] P. D. Jones, T. J. Osborn, and K. R. Briffa. "The evolution of climate over the last millennium". *Science* 292 (2001), pp. 662–667. DOI: 10.1126/science.1059126.

[A110] K. R. Briffa et al. "Low-frequency temperature variations from a northern tree ring density network". *J. Geophys. Res.* 106 (2001), pp. 2929–2941. DOI: 10.1029/2000JD900617.

[A111] T. J. Crowley. "Causes of climate change over the past 1000 years". *Science* 289 (2000), pp. 270–277. DOI: 10.1126/science.289.5477.270.

[A112] E. R. Cook, J. Esper, and R. D. D'Arrigo. "Extra-tropical Northern Hemisphere land temperature variability over the past 1000 years". *Quat. Sci. Rev.* 23 (2004), pp. 2063–2074. DOI: 10.1016/j.quascirev.2004.08.013.

[A113] D. Frank, J. Esper, and E. R. Cook. "Adjustment for proxy number and coherence in a large-scale temperature reconstruction". *Geophys. Res. Lett.* 34 (2007), p. L16709. DOI: 10.1029/2007GL030571.

[A114] A. Moberg et al. "Corrigendum: Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data". *Nature* 439 (2006), p. 1014. DOI: 10.1038/nature04575.

[A115] P. Dobrovolný et al. "Monthly, seasonal and annual temperature reconstructions for Central Europe derived from documentary evidence and instrumental records since AD 1500". *Clim. Change* 101 (2010), pp. 69–107. DOI: 10.1007/s10584-009-9724-x.

[A116] S. McIntyre and R. McKitrick. "Reply to comment by Huybers on "Hockey sticks, principal components, and spurious significance" by S. McIntyre and R. McKitrick". *Geophys. Res. Lett.* 32 (2005), p. L20713. DOI: 10.1029/2005GL023586.

[A117] R. S. Bradley and P. D. Jones. "When was the "Little Ice Age"?" Proceedings, International Symposium on the Little Ice Age Climate. Tokyo Metropolitan University, Tokyo, Japan. 1992.

[A118] R. S. Nickerson. "Confirmation bias: A ubiquitous phenomenon in many guises". *Rev. Gen. Psych.* 2 (1998), pp. 175–220. DOI: 10.1037/1089-2680.2.2.175.

[A119] M. K. Hughes and H. F. Diaz. "Was there a "Medieval Warm Period", and if so, where and when?" *Clim. Change* 26 (1994), pp. 109–142. DOI: 10.1007/BF01092410.

[A120] R. S. Bradley, M. K. Hughes, and H. F. Diaz. "Climate in medieval time". *Science* 302 (2003), pp. 404–405. DOI: 10.1126/science.1090372.

[A121] T. J. Osborn and K. R. Briffa. "The spatial extent of 20th-century warmth in the context of the past 1200 years". *Science* 311 (2006), pp. 841–844. DOI: 10.1126/science.1120514.

[A122] T. J. Osborn and K. R. Briffa. "Response to comment on "The spatial extent of 20th-century warmth in the context of the past 1200 years"". *Science* 316 (2007), p. 1844. DOI: 10.1126/science.1141446.

[A123] M. E. Mann et al. "Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly". *Science* 326 (2009), pp. 1256–1260. DOI: 10.1126/science.1177303.

[A124] R. S. Bradley et al. "The scope of medieval warming". *Science* 292 (2001), pp. 2011–2012. DOI: 10.1126/science.292.5524.2011b.

[A125] J. Esper and D. Frank. "The IPCC on a heterogeneous Medieval Warm Period". *Clim. Change* 94 (2009), pp. 267–273. DOI: 10.1007/s10584-008-9492-z.

[A126] K. R. Briffa et al. "Unusual twentieth-century summer warmth in a 1,000-year temperature record from Siberia". *Nature* 376 (1995), pp. 156–159. DOI: 10.1038/376156a0.

[A127] J. Esper, D. C. Frank, and R. J. S. Wilson. "Climate reconstructions: Low-frequency ambition and high-frequency ratification". *Eos* 85 (2004), pp. 113–115. DOI: 10.1029/2004EO120002.

[A128] H. Grudd. "Torneträsk tree-ring width and density AD 500-2003: a test of climatic sensitivity and a new 1500-year reconstruction of north Fennoscandian summers". *Clim. Dyn.* 31 (2008), pp. 843–857. DOI: 10.1007/s00382-007-0358-2.

[A129] G. Bürger. "Comment on "The spatial extent of 20th-century warmth in the context of the past 1200 years"". *Science* 316 (2007), p. 1844. DOI: 10.1126/science.1140982.

[A130] R. Rosenthal. "The file drawer problem and tolerance for null results". *Psych. Bull.* 86 (1979), pp. 638–641. DOI: 10.1037/0033-2909.86.3.638.

[A131] J. D. Scargle. "Publication bias: The "file-drawer" problem in scientific inference". *J. Sci. Exploration* 14 (2000), pp. 91–106. URL: http://www.scientificexploration.org/journal/jse_14_1_scargle.pdf.

[A132] A. Franco. "Publication bias in the social sciences: Unlocking the file drawer". *Science* In press (2014). DOI: 10.1126/science.1255484.

[A133] S. McIntyre and R. McKitrick. "Proxy inconsistency and other problems in millennial paleoclimate reconstructions". *Proc. Nat. Acad. Sci.* 106 (2009), E11. DOI: 10.1073/pnas.0812509106.

[A134] M. E. Mann, R. S. Bradley, and M. K. Hughes. "Reply to McIntyre and McKitrick: Proxy-based temperature reconstructions are robust". *Proc. Nat. Acad. Sci.* 106 (2009), E11. DOI: 10.1073/pnas.0812936106.

[A135] M. E. Mann, R. S. Bradley, and M. K. Hughes. "Supplementary Information: Corrigendum to Global-scale temperature patterns and climate forcing over the past six centuries". *Nature* (2004). URL: http://www.nature.com/nature/journal/v430/n6995/suppinfo/nature02478.html.

[A136] A. H. Lloyd and L. J. Graumlich. "Holocene dynamics of treeline forests in the Sierra Nevada". *Ecology* 78 (1997), pp. 1199–1210. DOI: 10.2307/2265869.

[A137] D. A. Graybill and S. B. Idso. "Detecting the aerial fertilization effect of atmospheric CO2 enrichment in tree-ring chronologies". *Glob. Biogeochem. Cycles* 7 (1993), pp. 81–95. DOI: 10.1029/92GB02533.

[A138] V. C. Jr. LaMarche et al. "Increasing atmospheric carbon dioxide: Tree ring evidence for growth enhancement in natural vegetation". *Science* 225 (1984), pp. 1019–1021. DOI: 10.1126/science.225.4666.1019.

[A139] C. F. Cooper et al. "Carbon Dioxide Enhancement of Tree Growth at High Elevations". *Science* 231 (1986), pp. 859–860. DOI: 10.1126/science.231.4740.860.

[A140] G. C. Jacoby and R. D. D'Arrigo. "Tree rings, carbon dioxide, and climatic change". *Proc. Nat. Acad. Sci.* 94 (1997), pp. 8350–8353.

[A141] L. A. Scuderi. "A 2000-year tree ring record of annual temperatures in the Sierra Nevada mountains". *Science* 259 (1993), pp. 1433–1436. DOI: 10.1126/science.259.5100.1433.

[A142] C. I. Millar et al. "Late Holocene forest dynamics, volcanism, and climate change at Whitewing Mountain and San Joaquin Ridge, Mono County, Sierra Nevada, CA, USA". *Quat. Res.* 66 (2006), pp. 273–287. DOI: 10.1016/j.yqres.2006.05.001.

[A143] L. N. Ababneh. “Analysis of radial growth patterns of strip-bark and wholebark bristlecone pine trees in the White Mountains of California”. Ph.D. Thesis. University of Arizona, 2006. URL: http://www.geo.arizona.edu/Antevs/Theses/.

[A144] R. J. Bale et al. “Temporal stability in bristlecone pine tree-ring stable oxygen isotope chronologies over the last two centuries”. *Holocene* 20 (2010), pp. 3–6. DOI: 10.1177/0959683609348867.

[A145] A. G. Bunn, L. J. Graumlich, and D. L. Urban. “Trends in twentieth-century tree growth at high elevations in the Sierra Nevada and White Mountains, USA”. *Holocene* 15 (2005), pp. 481–488. DOI: 10.1191/0959683605hl827rp.

[A146] M. W. Salzer et al. “Recent unprecedented tree-ring growth in bristlecone pine at the highest elevations and possible causes”. *Proc. Nat. Acad. Sci.* 106 (2009), pp. 20348–20353. DOI: 10.1073/pnas.0903029106.

[A147] S. Rutherford et al. “Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to method, predictor network, target season, and target domain”. *J. Clim.* 18 (2005), pp. 2308–2329. DOI: 10.1175/JCLI3351.1.

[A148] M. W. Salzer et al. “Five millennia of paleotemperature from tree-rings in the Great Basin, USA”. *Clim. Dyn.* In press (2013). DOI: 10.1007/s00382-013-1911-9.

[A149] K. R. Briffa et al. “Trends in recent temperature and radial tree growth spanning 2000 years across northwest Eurasia”. *Phil. Trans. R. Soc.* 363 (2008), pp. 2269–2282. DOI: 10.1098/rstb.2007.2199.

[A150] K. R. Briffa et al. “Reassessing the evidence for tree-growth and inferred temperature change during the Common Era in Yamalia, northwest Siberia”. *Quat. Sci. Rev.* 72 (2013), pp. 83–107. DOI: 10.1016/j.quascirev.2013.04.008.

[A151] M. M. Naurzbaev, M. K. Hughes, and E. A. Vaganov. “Tree-ring growth curves as sources of climatic information”. *Quat. Res.* 62 (2004), pp. 126–133. DOI: 10.1016/j.qres.2004.06.005.

[A152] NAS. “Surface temperature reconstructions for the last 2,000 years”. In: *Report for the National Academy of Sciences.* Ed. by G. R. North et al. The National Academies Home. Washington, D.C., U.S.A., 2006, 160pp. URL: http://books.nap.edu/openbook.php?isbn=0309102251.

[A153] E. J. Wegman, D. W. Scott, and Y. H. Said. “Ad hoc committee report on the ‘hockey stick’ global climate reconstruction”. In: Report to the USA’s Committee on Energy and Commerce, 2006, 95pp. URL: http://scienceandpublicpolicy.org/reprint/ad_hoc_report.html.

[A154] M. E. Mann, R. S. Bradley, and M. K. Hughes. “Corrigendum to Global-scale temperature patterns and climate forcing over the past six centuries”. *Nature* 430 (2004), p. 105. DOI: 10.1038/nature02478.

[A155] M. Tiljander et al. “A 3000-year palaeoenvironmental record from annually laminated sediment of Lake Korttajärvi, central Finland”. *Boreas* 32 (2003), pp. 566–577. DOI: 10.1080/03009480310004152.

[A156] D. S. Kaufman et al. “Corrections and clarifications”. *Science* 327 (2010), p. 644. DOI: 10.1126/science.327.5966.644-d.

[A157] P. D. Jones. “It was the best of times, it was the worst of times”. *Science* 280 (1998), pp. 544–545. DOI: 10.1126/science.280.5363.544.

[A158] M. E. Mann et al. “Global temperature patterns”. *Science* 280 (1998), pp. 2029–2030. DOI: 10.1126/science.280.5372.2027e.

[A159] M. Mann et al. “On past temperatures and anomalous late-20th century warmth”. *Eos Trans.* 84 (2003), pp. 256–257. DOI: 10.1029/2003EO270039.

[A160] W. Soon, S. Baliunas, and D. Legates. “Comment on “On past temperatures and anomalous late-20th century warmth””. *Eos Trans.* 84 (2003), pp. 473–476. DOI: 10.1029/2003EO440007.

[A161] M. Mann et al. “Reply to comment on “On past temperatures and anomalous late-20th century warmth””. *Eos Trans.* 84 (2003), pp. 473–476. DOI: 10.1029/2003EO440008.

[A162] T. J. Osborn, S. C. B. Raper, and K. R. Briffa. “Simulated climate change during the last 1,000 years: comparing the ECHO-G general circulation model with the MAGICC simple climate model”. *Clim. Dyn.* 27 (2006), pp. 185–197. DOI: 10.1007/s00382-006-0129-5.

[A163] M. E. Mann et al. “Reply”. *J. Clim.* 20 (2007), pp. 5671–5674. DOI: 10.1175/JCLI1894.1.

[A164] J. E. Smerdon et al. “A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium.” *J. Clim.* 23 (2010), pp. 4856–4880. DOI: 10.1175/2010JCLI3328.1.

[A165] J. E. Smerdon et al. “Spatial performance of four climate field reconstruction methods targeting the Common Era”. *Geophys. Res. Lett.* 38 (2011), p. L11705. DOI: 10.1029/2011GL047372.

[A166] H. von Storch and E. Zorita. “Comment on “Hockey sticks, principal components, and spurious significance” by S. McIntyre and R. McKitrick”. *Geophys. Res. Lett.* 32 (2005), p. L20701. DOI: 10.1029/2005GL022753.

[A167] S. McIntyre and R. McKitrick. “Reply to comment by von Storch and Zorita on “Hockey sticks, principal components, and spurious significance” by S. McIntyre and R. McKitrick”. *Geophys. Res. Lett.* 32 (2005), p. L20714. DOI: 10.1029/2005GL023089.

[A168] P. Huybers. “Comment on “Hockey sticks, principal components, and spurious significance” by S. McIntyre and R. McKitrick”. *Geophys. Res. Lett.* 32 (2005), p. L20705. DOI: 10.1029/2005GL023395.

[A169] C. M. Ammann and E. R. Wahl. “The importance of the geophysical context in statistical evaluations of climate reconstruction procedures”. *Clim. Change* 85 (2007), pp. 71–88. DOI: 10.1007/s10584-007-9276-x.

[A170] E. R. Wahl and C. M. Ammann. “Robustness of the Mann, Bradley, Hughes reconstruction of Northern Hemisphere surface temperatures: Examination of criticisms based on the nature and processing of proxy climate evidence”. *Clim. Change* 85 (2007), pp. 33–69. DOI: 10.1007/s10584-006-9105-7.

[A171] IPCC. “Summary for Policymakers”. In: *Climate change 2007: The physical science basis. Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change.* Ed. by S. Solomon et al. Cambridge University Press. Cambridge. New York., 2007, 1056pp. URL: http://www.ipcc.ch/.

[A172] IPCC. “Summary for Policymakers”. In: *Climate change 2013: The physical science basis. Contribution of Working Group 1 to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Ed. by T. F. Stocker et al. Cambridge University Press. Cambridge. New York., 2013, 1535pp. URL: http://www.ipcc.ch/.

[A173] R. D’Arrigo et al. “On the “divergence problem” in northern forests: A review of the tree-ring evidence and possible causes”. *Glob. Planet. Change* 60 (2008), pp. 289–305. DOI: 10.1016/j.gloplacha.2007.03.004.

[A174] R. Wilson et al. “A matter of divergence: Tracking recent warming at hemispheric scales using tree ring data”. *J. Geophys. Res.* 112 (2007), p. D17103. DOI: 10.1029/2006JD008318.

[A175] G. C. Jacoby and R. D’Arrigo. “Tree-ring width and density evidence of climatic and potential forest change in Alaska”. *Glob. Biogeochem. Cycles* 9 (1995), pp. 227–234. DOI: 10.1029/95GB00321.

[A176] K. R. Briffa et al. “Reduced sensitivity of recent tree-growth to temperature at high northern latitudes”. *Nature* 391 (1998), pp. 678–682. DOI: 10.1038/35596.

[A177] K. R. Briffa et al. “Trees tell of past climates: but are they speaking less clearly today?” *Phil. Trans. R. Soc. Lond. B* 353 (1998), pp. 65–73. DOI: 10.1098/rstb.1998.0191.

[A178] M. Wilmking et al. “Recent climate warming forces contrasting growth responses of white spruce at treeline in Alaska through temperature thresholds”. *Glob. Change Biol.* 10 (2004), pp. 1724–1736. DOI: 10.1111/j.1365-2486.2004.00826.x.

[A179] R. D’Arrigo, G. C. Jacoby, and G. P. Judayi. “Increased temperature sensitivity and divergent growth trends in circumpolar boreal forests”. *Geophys. Res. Lett.* 32 (2005), p. L15715. DOI: 10.1029/2005GL023331.

[A180] U. Büntgen et al. “The ‘Divergence Problem’ in tree-ring research”. In: *TRACE - Tree Rings in Archaeology, Climatology* *and Ecology*. Ed. by R. Kaczka et al. Vol. 7. GFZ Potsdam, Scientific Technical Report STR 09/03, Potsdam, 226pp.

[A181] J. Esper et al. "Trends and uncertainties in Siberian indicators of 20th century warming". *Glob. Change Biol.* 16 (2010), pp. 386–398. DOI: 10.1111/j.1365-2486.2009.01913.x.

[A182] G. Manley. "Central England temperatures: monthly means 1659-1973". *Quart. J. R. Met. Soc.* 100 (1974), pp. 389–405. DOI: 10.1256/smsqj.42510.

[A183] D. Parker and B. Horton. "Uncertainties in central England temperature 1878-2003 and some improvements to the maximum and minimum series". *Int. J. Clim.* 25 (2003), pp. 1173–1188. DOI: 10.1002/joc.1190.

[A184] A. Moberg and H. Bergström. "Homogenization of Swedish temperature data. Part III: the long temperature records from Uppsala". *Int. J. Clim.* 17 (1997), pp. 667–699. DOI: 10.1002/(SICI)1097-0088(19970615)17:7<667::AID-JOC115>3.0.CO;2-J.

[A185] A. Moberg et al. "Daily air temperature and pressure series for Stockholm (1756-1998)". *Clim. Change* 53 (2002), pp. 171–212. DOI: 10.1023/A:1014966724670.

[A186] A. Moberg et al. "Were southern Swedish summer temperatures before 1860 as warm as measured?" *Int. J. Clim.* 23 (2003), pp. 1495–1521. DOI: 10.1002/joc.945.

[A187] D. Frank et al. "Warmer early instrumental measurements versus colder reconstructed temperatures: shooting at a moving target". *Quat. Sci. Rev.* 26 (2007), pp. 3298–3310. DOI: 10.1016/j.quascirev.2007.08.002.

[A188] R. Böhm et al. "The early instrumental warm-bias: a solution for long central European temperature series 1760-2007". *Clim. Change* 101 (2010), pp. 41–67. DOI: 10.1007/s10584-009-9649-4.

[A189] I. Auer et al. "HISTALP - historical instrumental climatological surface time series of the Greater Alpine Region". *Int. J. Clim.* 27 (2007), pp. 17–46. DOI: 10.1002/joc.1377.

[A190] V. Masson-Delmotte et al. "2013: Information from paleoclimate archives". In: *Climate change 2013: The physical science basis. Contribution of Working Group 1 to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T. F. Stocker et al. Cambridge University Press. Cambridge. New York., 2013, 1535pp. URL: http://www.ipcc.ch/.

[A191] P. A. Mayewski et al. "Holocene climate variability". *Quat. Res.* 62 (2004), pp. 243–255. DOI: 10.1016/j.yqres.2004.07.001.

[A192] M. E. Mann et al. "Tree-ring chronologies and climate variability". *Science* 296 (2002), pp. 848–849. DOI: 10.1126/science.296.5569.848.

Non-peer reviewed

[B1] RealClimate. *Hockey stick*. 2004. URL: http://www.realclimate.org/index.php/archives/2004/11/hockey-stick/.

[B2] A. Gore. *An inconvenient truth*. 1st Ed. Bloomsbury Publishing. London., 2006.

[B3] A. W Montford. *The Hockey Stick Illusion*. 1st Ed. Stacey International. London. UK., 2010.

[B4] C. Booker. *The real global warming disaster*. 1st Ed. Continuum International Publishing Group. London . New York., 2009.

[B5] M. E. Mann. *The hockey stick and the climate wars: Dispatches from the front lines*. 1st Ed. Columbia University Press. New York. USA., 2012, p. 448.

[B6] J. Hoggan and R. Littlemore. *Climate cover-up: The crusade to deny global warming*. 1st Ed. Greystone Books. Vancouver. Canada, 2009.

[B7] ClimateAudit. *MBH99 and proxy calibration*. 2008. URL: http://climateaudit.org/2008/04/23/univariate-calibration/\#comment-145122.

[B8] K. Briffa and E. Cook. "What are the sources of uncertainty in the tree-ring data: How can they be quantified and represented?" White paper submitted to Trieste, 2008 Paleoclimate Uncertainties Workshop. 2008. URL: http://www.ncdc.noaa.gov/paleo/reports.

[B9] ClimateAudit. *Wilson on Yamal Substitution. Comment #44184*. 2006. URL: http://climateaudit.org/2006/02/22/wilson-on-yamal-substitution/#comment-44184.

[B10] ClimateAudit. *Yamal: A "divergence" problem*. 2009. URL: http://climateaudit.org/2009/09/27/yamal-a-divergence-problem/.

[B11] ClimateAudit. *Almagre strip bark*. 2007. URL: http://climateaudit.org/2007/10/17/almagre-strip-bark/.

[B12] J. Bouldin. *Severe analytical problems in dendroclimatology, parts one to fourteen*. 2012. URL: https://ecologicallyoriented.wordpress.com/?s="Severe+analytical+problems+in+dendroclimatology"&submit=Search.

[B13] J. Bouldin. *Tree rings, signal extraction and experiments*. 2013. URL: https://ecologicallyoriented.wordpress.com/2013/03/10/tree-rings-signal-extraction-and-experiments/.

[B14] J. Bouldin. *Briffa et al., 2013, parts one to four*. 2013. URL: https://ecologicallyoriented.wordpress.com/?s=Briffa+et+al.+2013&submit=Search.

[B15] J. Cecile, C. Pagnutti, and M. Anand. "A likelihood perspective on tree-ring standardization: eliminating modern sample bias". *Clim. Past Discuss.* 9 (2013), pp. 4499–4551. DOI: 10.5194/cpd-9-4499-2013.

[B16] LLC. Statisticat. *Bayesian-Inference: Overfitting*. URL: http://www.bayesian-inference.com/overfitting.

[B17] T. H. Huxley. *Presidential Address to the British Association for the Advancement of Science, "Biogenesis and abiogenesis"*. 1870. URL: http://aleph0.clarku.edu/huxley/CE8/B-Ab.html.

[B18] R. Connolly and M. Connolly. "Urbanization bias I. Is it a negligible problem for global temperature estimates?" 28 (Clim. Sci.). Ver. 0.1 (non peer reviewed draft). 2014. URL: http://oprj.net/articles/climate-science/28.

[B19] R. Connolly and M. Connolly. "Has poor station quality biased U.S. temperature trend estimates?" 11 (Clim. Sci.). Ver. 0.1 (non peer reviewed draft). 2014. URL: http://oprj.net/articles/climate-science/11.

[B20] R. Connolly and M. Connolly. "Urbanization bias II. An assessment of the NASA GISS urbanization adjustment method". 31 (Clim. Sci.). Ver. 0.1 (non peer reviewed draft). 2014. URL: http://oprj.net/articles/climate-science/31.

[B21] R. Connolly and M. Connolly. "Urbanization bias III. Estimating the extent of bias in the Historical Climatology Network datasets". 34 (Clim. Sci.). Ver. 0.1 (non peer reviewed draft). 2014. URL: http://oprj.net/articles/climate-science/34.

[B22] O. Bothe and D. Zanchettin. "Ambiguity of large scale temperature reconstructions from artificial tree growth in millennial climate simulations". *arXiv* 1207.2279 (2012). URL: http://arxiv.org/abs/1207.2279.

[B23] G. O'Toole. *Everything should be made as simple as possible, but not simpler*. 2011. URL: http://quoteinvestigator.com/2011/05/13/einstein-simple/.

[B24] ClimateAudit. *D'Arrigo: Making cherry pie*. 2006. URL: http://climateaudit.org/2006/03/07/darrigo-making-cherry-pie/.

[B25] Statpad. *EIV/TLS Regression - Why use it?* 2010. URL: http://statpad.wordpress.com/2010/12/19/eivtls-regression-why-use-it/.

[B26] William M. Briggs. *Do not smooth times series, you hockey puck!* 2008. URL: http://wmbriggs.com/blog/?p=195.

[B27] CO2Science. *Medieval Warm Period Project*. 2011. URL: http://www.co2science.org/data/mwp/mwpp.php.

[B28] ClimateAudit. *Polar Urals #1*. 2005. URL: http://climateaudit.org/2005/04/04/polar-urals-1.

[B29] ClimateAudit. *Polar Urals: Briffa versus Esper*. 2006. URL: http://climateaudit.org/2006/02/24/polar-urals-briffa-versus-esper/.

[B30] ClimateAudit. *Polar Urals #2: Broken core*. 2005. URL: http://climateaudit.org/2005/04/04/polar-urals-2-broken-core.

[B31] ClimateAudit. *Polar Urals #3: Crossdating*. 2005. URL: http://climateaudit.org/2005/04/05/polar-urals-3-crossdating.

[B32] ClimateAudit. *Wilson on Yamal substitution*. 2006. URL: http://climateaudit.org/2006/02/22/wilson-on-yamal-substitution/.

[B33] ClimateAudit. *Briffa's Tornetrask reconstruction*. 2005. URL: http://climateaudit.org/2005/03/27/briffas-tornetrask-reconstruction/.

[B34] ClimateAudit. *Tornetrask digital version - Hooray!* 2008. URL: http://climateaudit.org/2008/07/25/tornetrask-digital-version-hooray/.

[B35] ClimateAudit. *Making apple pie instead of cherry pie*. 2006. URL: http://climateaudit.org/2006/03/11/making-apple-pie/.

[B36] ClimateAudit. *Jones et al 1998: Impact of new versions*. 2008. URL: http://climateaudit.org/2008/08/04/jones-et-al-1998-impact-of-new-versions/.

[B37] Climate Audit. *Juckes and the Esper and Jones CVMs*. 2006. URL: http://climateaudit.org/2006/11/01/team-euro-code-2/.

[B38] Willis Eschenbach. *Can't see the signal for the trees*. 2008. URL: http://climateaudit.org/2008/11/23/cant-see-the-signal-for-the-trees/.

[B39] TheAirVent. *Robust?* 2010. URL: http://noconsensus.wordpress.com/2010/10/16/robust/.

[B40] Willis Eschenbach. *Kill it with fire*. 2011. URL: http://wattsupwiththat.com/2011/05/30/kill-it-with-fire/.

[B41] BishopHill. *A good trick to create a decline*. 2010. URL: http://bishophill.squarespace.com/blog/2010/4/26/a-good-trick-to-create-a-decline.html.

[B42] ClimateAudit. *Making hockey sticks the Jones way*. 2006. URL: http://climateaudit.org/2006/06/04/making-hockey-sticks-the-jones-way/.

[B43] ClimateAudit. *Bring the proxies up to date!!* 2005. URL: http://climateaudit.org/2005/02/20/bring-the-proxies-up-to-date/.

[B44] RealClimate. *Myth vs. fact regarding the "Hockey stick" - comment #345*. 2004. URL: http://www.realclimate.org/index.php/archives/2004/12/myths-vs-fact-regarding-the-hockey-stick/#comment-345.

[B45] ClimateAudit. *A little secret*. 2007. URL: http://climateaudit.org/2007/10/12/a-little-secret/.

[B46] ClimateAudit. *The MBH98 corrigendum*. 2008. URL: http://climateaudit.org/2008/09/29/the-mbh98-corrigendum/.

[B47] ClimateAudit. *Silence of the lambs*. 2008. URL: http://climateaudit.org/2008/10/15/the-silence-of-the-lambs/.

[B48] ClimateAudit. *Top fifteen reasons for withholding data or code*. 2005. URL: http://climateaudit.org/2005/03/05/top-eleven-reasons-for-withholding-data-or-code/.

[B49] M. Rundkvist. *Open source dendrochronology*. 2009. URL: http://scienceblogs.com/aardvarchaeology/2009/06/open_source_dendrochronology.php.

[B50] ClimateAudit. *Huybers 2: Re-scaling*. 2005. URL: http://climateaudit.org/2005/09/17/huybers-2-re-scaling/.

[B51] ClimateAudit. *In the mail today*. 2005. URL: http://climateaudit.org/2005/09/08/in-the-mail-today/.

[B52] ClimateAudit. *Cicerone of NAS acquiesces in data obstruction*. 2007. URL: http://climateaudit.org/2007/07/30/cicerone-of-nas-refuses-to-require-thompson-to-archive-data/.

[B53] F. Pearce. *Secrecy in science - An argument for open access*. 2011. URL: http://www.indexoncensorship.org/2011/11/climate-change-secrecy-freedom-information/.

[B54] M. Nielsen. *The new Einsteins will be scientists who share*. 2011. URL: http://online.wsj.com/article/SB10001424052970204644504576653573191370088.html.

[B55] ClimateAudit. *Thompson et al [1993] on Dunde*. 2005. URL: http://climateaudit.org/2005/09/18/thompson-et-al-1993-on-dunde/.

[B56] ClimateAudit. *Behind closed doors: "Perpetuating rubbish"*. 2011. URL: http://climateaudit.org/2011/11/25/behind-closed-doors-perpetuating-rubbish/.

[B57] ClimateAudit. *What was "first" about MBH98?* 2006. URL: http://climateaudit.org/2006/07/02/what-was-first-about-mbh98/.

[B58] ClimateAudit. *Bristlecone/Foxtail #1: Cirque Park - comment 32888*. 2005. URL: http://climateaudit.org/2005/03/29/bristleconefoxtail-site-1-cirque-peak/#comment-32888.

[B59] M. Hughes, T. Crowley, and K. Briffa. *Leaked e-mail correspondence*. 2000. URL: http://assassinationscience.com/climategate/1/FOIA/mail/0969618170.txt.

[B60] ClimateAudit. *Bristlecones, foxtails and temperature*. 2006. URL: http://climateaudit.org/2006/06/05/bristlecones-foxtails-and-temperature/.

[B61] ClimateAudit. *The Sheep Mountain update*. 2007. URL: http://climateaudit.org/2007/11/14/the-sheep-mountain-update/.

[B62] ClimateAudit. *Almagre: the Graybill photos*. 2007. URL: http://climateaudit.org/2007/10/16/almagre-the-graybill-photos/.

[B63] ClimateAudit. *Strip bark at Upper Wright Lakes foxtails*. 2007. URL: http://climateaudit.org/2007/10/24/strip-bark-at-upper-wright-lakes-foxtails/.

[B64] ClimateAudit. *Bunn et al. [2005] - 20th century tree growth in the Sierra Nevadas*. 2005. URL: http://climateaudit.org/2005/06/07/bunn-et-al-2005-20th-century-tree-growth-in-the-sierra-nevadas/.

[B65] ClimateAudit. *Salzer et al 2009 - A first look*. 2009. URL: http://climateaudit.org/2009/11/17/salzer-et-al-2009-a-first-look/.

[B66] ClimateAudit. *Rob Wilson on bristlecones*. 2006. URL: http://climateaudit.org/2006/03/30/rob-wilson-on-bristlecones/#comment-47556/.

[B67] BishopHill. *The Yamal implosion*. 2009. URL: http://bishophill.squarespace.com/blog/2009/9/29/the-yamal-implosion.html.

[B68] ClimateAudit. *Fresh data on Briffa's Yamal #1*. 2009. URL: http://climateaudit.org/2009/09/26/briffas-yamal-crack-cocaine-for-paleoclimatologists/.

[B69] ClimateAudit. *YAD06 - the most influential tree in the world*. 2009. URL: http://climateaudit.org/2009/09/30/yamal-the-forest-and-the-trees/.

[B70] K. R. Briffa. *The Yamal ring-width chronology of Briffa (2000)*. 2009. URL: http://www.cru.uea.ac.uk/cru/people/briffa/yamal2000/.

[B71] K. R. Briffa and T. M. Melvin. *Examining the validity of the published RCS Yamal tree-ring chronology*. 2009. URL: http://www.cru.uea.ac.uk/cru/people/briffa/yamal2009/.

[B72] ClimateAudit. *Keith Briffa responds*. 2009. URL: http://climateaudit.org/2009/10/01/keith-briffa-responds.

[B73] ClimateAudit. *Is Yamal homogeneous? An Esper-style answer*. 2009. URL: http://climateaudit.org/2009/11/11/is-yamal-homogeneous-an-esper-style-answer/.

[B74] TheAirVent. *The unstoppable Dirty Dozen*. 2009. URL: http://noconsensus.wordpress.com/2009/11/08/the-unstoppable-dirty-dozen/.

[B75] ClimateAudit. *CRU abandons Yamal superstick*. 2013. URL: http://climateaudit.org/2013/06/28/cru-abandons-yamal-superstick/.

[B76] ClimateAudit. *Re-visiting the "Yamal Substitution"*. 2009. URL: http://climateaudit.org/2009/10/19/re-visiting-the-yamal-substitution/.

[B77] AMac. *Lightsum and darksum are* calculated*, not* measured. 2011. URL: http://amac1.blogspot.com/2011/08/lightsum-and-darksum-are-calculated-not.html.

[B78] ClimateAudit. *It's Saturday Night Live*. 2008. URL: http://climateaudit.org/2008/10/02/its-saturday-night-live/.

[B79] AMac. *Two views of Tiljander*. 2010. URL: http://amac1.blogspot.com/2010/08/two-views-of-tiljander.html.

[B80] ClimateAudit. *Upside down Mann and the "peerreviewedliterature"*. 2009. URL: http://climateaudit.org/2009/10/14/upside-side-down-mann-and-the-peerreviewedliterature/.

[B81] AMac. *Blog links: Mann '08 and the Korttajarvi (Tiljander) varve proxies*. 2009. URL: http://amac1.blogspot.com/2009/11/blog-links-mann-08-and-korttajarvi.html.

[B82] M. E. Mann. *Supplementary information for: Global signatures of the Little Ice Age and Medieval Climate Anomaly and plausible dynamical origins*. 2009. URL: http://www.meteo.psu.edu/~mann/supplements/MultiproxySpatial09/.

[B83] ClimateAudit. *The no-dendro illusion*. 2010. URL: http://climateaudit.org/2010/08/01/the-no-dendro-illusion/.

[B84] ClimateAudit. *Mann and his bristlecones*. 2010. URL: http://climateaudit.org/2010/08/07/mann-and-his-bristlecones/.

[B85] ClimateAudit. *Make a stick, make a stick*. 2010. URL: http://climateaudit.org/2010/07/30/make-a-stick-make-a-stick/.

[B86] M. E. Mann. *Supplementary information for: PNAS article "Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia"*. 2009. URL: http://www.meteo.psu.edu/~mann/supplements/MultiproxyMeans07/.

[B87] ClimateAudit. *Al Gore and "Dr Thompson's thermometer" # 2*. 2007. URL: http://climateaudit.org/2007/11/10/al-gore-and-dr-thompsons-thermometer-2/.

[B88] M. E. Mann. *A look into our climate: Past to present to future*. 2011. URL: http://tedxtalks.ted.com/video/TEDxPSU-Michael-Mann-A-Look-Int.

[B89] F. C. Ljungqvist et al. "Northern Hemisphere temperature patterns in the last 12 centuries". *Clim. Past Discuss.* 7 (2011), pp. 3349–3397. DOI: 10.5194/cpd-7-3349-2011.

[B90] ClimateAudit. *The Hockey Stick and the Milankovitch Theory*. 2010. URL: http://climateaudit.org/2010/02/03/the-hockey-stick-and-milankovitch-theory/.

[B91] ClimateAudit. *The MBH99 CO2 "Adjustment"*. 2007. URL: http://climateaudit.org/2007/11/13/the-mbh99-co2-adjustment-part-1/.

[B92] ClimateAudit. *The gift that keeps on giving*. 2007. URL: http://climateaudit.org/2007/11/13/the-gift-that-keeps-on-giving/.

[B93] ClimateAudit. *Bristlecone "Adjustment" # 2*. 2005. URL: http://climateaudit.org/2005/02/09/bristlecone-adjustment-2/.

[B94] H. Von Storch. *The CR problem*. 2003. URL: http://coast.hzg.de/staff/storch/CR-problem/cr.2003.htm.

[B95] ClimateAudit. *More on requests for data*. 2005. URL: http://climateaudit.org/2005/07/02/more-on-requests-for-data/.

[B96] ClimateAudit. *Rutherford, Mann et al [2005]*. 2006. URL: http://climateaudit.org/2006/01/28/rutherford-mann-et-al-2005/.

[B97] C. Babbage. *Passages from the life of a philosopher*. Longman, Green, Longman, Roberts & Green. London, 1864, 541pp. URL: https://archive.org/details/passagesfromlif01babbgoog.

[B98] ClimateAudit. *Some principal components illustrations*. 2006. URL: http://climateaudit.org/2006/03/29/some-principal-components-illustrations/.

[B99] RealClimate. *On yet another false claim by McIntyre and McKitrick*. 2005. URL: http://www.realclimate.org/index.php/archives/2005/01/on-yet-another-false-claim-by-mcintyre-and-mckitrick/.

[B100] ClimateAudit. *Ian Jolliffe comments at Tamino*. 2008. URL: http://climateaudit.org/2008/09/08/ian-jolliffe-comments-at-tamino/.

[B101] RealClimate. *Dummies guide to the latest "Hockey Stick" controversy*. 2005. URL: http://www.realclimate.org/index.php/archives/2005/02/dummies-guide-to-the-latest-hockey-stick-controversy/.

[B102] RealClimate. *Peer review: A necessary but not sufficient condition II*. 2005. URL: http://www.realclimate.org/index.php/archives/2005/01/peer-review-ii/.

[B103] ClimateAudit. *Source code: Preisendorfer's Rule N*. 2005. URL: http://climateaudit.org/2005/08/04/source-code-preisendorfers-rule-n/.

[B104] ClimateAudit. *Errors matter #3: Preisendorfer's Rule N*. 2005. URL: http://climateaudit.org/2005/02/13/errors-matter-3-preisendorfers-rule-n/.

[B105] ClimateAudit. *Roll over, Preisendorfer*. 2005. URL: http://climateaudit.org/2005/08/05/roll-over-preisendorfer/.

[B106] ClimateAudit. *Errors matter #1: the no-PC alternative*. 2005. URL: http://climateaudit.org/2005/02/11/errors-matter-1-the-no-pc-alternative/.

[B107] ClimateAudit. *Errors matter #2: the "different" method of Rutherford et al [2005]*. 2005. URL: http://climateaudit.org/2005/02/12/errors-matter-2-the-different-method-of-rutherford-et-al-2005/.

[B108] ClimateAudit. *The Team defends paleo-phenology*. 2010. URL: http://climateaudit.org/2010/07/25/the-team-defends-paleo-phrenology/.

[B109] ClimateAudit. *Re-post of "Tamino and the Magic Flute"*. 2010. URL: http://climateaudit.org/2010/07/25/repost-of-tamino-and-the-magic-flute/.

[B110] ClimateAudit. *Answers to the House Committee on cross-validation statistics*. 2005. URL: http://climateaudit.org/2005/07/22/answers-to-the-house-committee-on-cross-validation-statistics/.

[B111] ClimateAudit. *Well, well. Look what the cat dragged in*. 2008. URL: http://climateaudit.org/2008/08/06/well-well-look-what-the-cat-dragged-in/.

[B112] ClimateAudit. *Caspar Ammann, Texas Sharpshooter*. 2008. URL: http://climateaudit.org/2008/08/08/caspar-ammann-texas-sharpshooter/.

[B113] BishopHill. *Caspar and the Jesus paper*. 2008. URL: http://bishophill.squarespace.com/blog/2008/8/11/caspar-and-the-jesus-paper.html.

[B114] ClimateAudit. *The RE benchmark of 0*. 2008. URL: http://climateaudit.org/2008/04/17/a-reanalysis-of-manns-proxies/.

[B115] RealClimate. *National Academies Synthesis Report*. 2006. URL: http://www.realclimate.org/index.php/archives/2006/06/national-academies-synthesis-report/.

[B116] ClimateAudit. *The Wegman and North reports for newbies*. 2007. URL: http://climateaudit.org/2007/11/06/the-wegman-and-north-reports-for-newbies/.

[B117] RealClimate. *The missing piece at the Wegman hearing*. 2006. URL: http://www.realclimate.org/index.php/archives/2006/07/the-missing-piece-at-the-wegman-hearing/.

[B118] TheAirVent. *Hockey stick CPS revisited - Part 1*. 2009. URL: http://noconsensus.wordpress.com/2009/06/20/hockey-stick-cps-revisited-part-1/.

[B119] TheAirVent. *Historic hockey stick - Pr 2 Shock and recovery*. 2009. URL: http://noconsensus.wordpress.com/2009/06/23/histori-hockey-stick-pt-2/.

[B120] D. R. B. Stockwell. "Reconstruction of past climate using series with a red noise". Archived on David Stockwell's website. 2006. URL: http://landshape.org/images/script.pdf.

[B121] S. A. Sandford. "Apples and oranges - a comparison". *Ann. Improbable Res.* 1 (1995), p. 1i3. URL: http://improbable.com/airchives/paperair/volume1/v1i3/air-1-3-apples.html.

[B122] SkepticalScience. "Tree-ring proxies and the divergence problem" (2010). URL: http://www.skepticalscience.com/Tree-ring-proxies-divergence-problem.htm.

[B123] ClimateEtc. *Hiding the decline*. 2011. URL: http://judithcurry.com/2011/02/22/hiding-the-decline/.

[B124] ClimateAudit. *Briffa: Large-scale decline in ring widths*. 2006. URL: http://climateaudit.org/2006/02/13/briffa-large-scale-decline-in-ring-widths/.

[B125] ClimateAudit. *Cook et al[2004]: More cargo cult?* 2006. URL: http://climateaudit.org/2006/03/14/cook-et-al2004-more-cargo-cult/.

[B126] ClimateAudit. *Divergence: the Young Dendros rebel*. 2007. URL: http://climateaudit.org/2007/12/15/divergence-the-young-dendros-rebel/.

[B127] M. Wilmking and J. Singh. "Eliminating the "divergence problem" at Alaska's northern treeline". *Clim. Past Discuss.* 4 (2008), pp. 741–759. DOI: 10.5194/cpd-4-741-2008.

[B128] Storytelling with data. *Strategies for avoiding the spaghetti graph*. 2013. URL: http://www.storytellingwithdata.com/2013/03/avoiding-spaghetti-graph.html.

[B129] E. Zorita. *McShane and Wyner on climate reconstruction methods*. 2010. URL: http://klimazwiebel.blogspot.com/2010/08/mcshane-and-wyner-on-climate.html.

[B130] ClimateAudit. *McShane and Wyner discussion*. 2010. URL: http://climateaudit.org/2010/12/14/mcshane-and-wyner-discussion-2/.

[B131] RealClimate. *Responses to McShane and Wyner*. 2010. URL: http://www.realclimate.org/index.php/archives/2010/12/responses-to-mcshane-and-wyner/.

[B132] StrangeWeather. *McShane and Wyner*. 2010. URL: http://strangeweather.wordpress.com/2010/11/12/mcshane-and-wyner/.

[B133] ClimateAudit. *McShane and Wyner weights on Mann 2008 proxies*. 2011. URL: http://climateaudit.org/2011/06/09/mcshane-and-wyner-weights/.

[B134] K. Z. Lorenz. *On aggression*. Routledge, 1966, p12.

[B135] F. Bacon. *The advancement of learning*. 1605, Book 1, v, 8.