Measuring the skill of variance-scaled climate reconstructions and a test for the capture of extremes.

Danny McCarroll*, Giles H.F. Young and Neil J. Loader

Department of Geography, Swansea University, Singleton Park,

Swansea SA2 8PP Wales, UK.

*Corresponding author: d.mccarroll@swansea.ac.uk

Abstract

Climate reconstructions produced using regression, with a proxy as the independent variable, are inevitably biased towards the mean, exhibit reduced variance and underestimate extremes. Scaling the mean and variance to fit those of the target climate data produces a more realistic range of reconstructed values but the cost, in terms of inflated error, is seldom assessed. We provide a simple metric that allows the loss of skill due to scaling to be quantified. It can be calculated retrospectively for published studies, some of which exhibit little or no reconstructive skill. Although scaled reconstructions must have a range that is close to that of the target climate data, there is no guarantee that the correct years are pushed to the extremes. We propose a simple non-parametric test for 'Extreme Value Capture' that gives the statistical significance of a given number of the correct years being 'captured' beyond the thresholds defined by the upper and lower 10% of the measured climate data. The methods are tested using three annually resolved case studies. A tree-growth based summer temperature reconstruction for northern Fennoscandia captures cold summers very well but the capture rate of the warmest summers is no better than might be expected purely due to chance. Such failure to correctly capture the warmest years has important implications for interpretation of the frequency and magnitude of very warm summers in the past.

Key words: Palaeoclimate, tree rings, regression, statistical methods, climate change, medieval warm period.

Introduction

The standard approach to reconstructing the climate of the past, using annually resolved records such as tree rings or historical documents, is to use simple linear regression (National Research Council 2007). The strength of the relationship between the proxy and the climatic target is quantified using correlation statistics; typically the Pearson's correlation coefficient (r) and the squared correlation coefficient (R^2). The latter provides a measure of the amount of variance in the proxy that is common to the variance in the climate record. Where the climate records available for calibration are of sufficient length, it has become standard practice to test the temporal stability of the relationship by using split-period calibration and verification tests. For these tests the data are split into two, often equal, periods and the calibration is carried out on one half of the data (calibration) and tested on the other half (verification). The Reduction of Error (RE) and Coefficient of Efficiency (CE) statistics (Cook and Kairiukstis 1990) provide a measure of how well the reconstruction fits the measured data over the verification period; in contrast to correlation, both the RE and CE tests are sensitive to offsets between the measured and reconstructed values. RE and CE values of zero occur where the average squared difference between the measured and reconstructed values (the Mean Squared Error: MSE) is the same as that obtained when a horizontal line is used as an estimate for the climate of the verification period; in the case of RE this line represents the mean measured climate variable over the calibration period, and for CE it is the mean of the verification period. Often the statistics are calculated twice by switching the calibration and verification periods. Where the RE and CE values are positive, the relationship between proxy and climate target is considered temporally stable and then the proxy data over the full period of available climate data is used to perform the regression

analysis, using the proxy as the independent variable with which to reconstruct past climate (National Research Council 2007).

Climate reconstructions produced in this way, however, suffer from a serious problem in that they inevitably underestimate the variability of past climate (von Storch et al. 2004; Esper et al. 2005). This underestimation occurs because climate proxies are never perfectly correlated with the target climate data, leaving a proportion of the variance in the proxy unexplained. Simple linear regression uses the principle of least squares, so that the regression line that defines the relationship between the proxy and the target is fitted to minimise the squared difference between each point and the line on the vertical (climate) axis. If the correlation were perfect between proxy and climate, the regression line would pass through every point and the MSE would be zero. Any subsequent reconstructed climate would have the same variance as the measured climate over the calibration period, and we presume also for the rest of the reconstruction. If the correlation between the proxy and the target were zero then the line that minimises the MSE would be horizontal and pass through the mean value of the measured climate. In this case all proxy values would predict the same (mean) value for the climate target and the climate reconstruction would be a horizontal line. When, as is always the case, the correlation between the proxy and the climate target falls between zero and (positive or negative) one, the reconstruction captures some, but not all, of the variance in the target climate variable, and the weaker the correlation, the closer the reconstruction comes to being a horizontal line. The loss of variance in this kind of regression-based reconstruction can, in essence, be viewed as a bias towards the mean. Where the purpose of a climate reconstruction is to compare the reconstructed climate of the past with the measured climate of the recent period, or with climate model output, then a bias towards the mean and the resultant underestimation of both positive and negative extremes can be highly problematic.

To overcome the bias towards the mean, and produce climate reconstructions that capture the full range of climate variability in the past, many authors now employ variance scaling, or variance matching (e.g. Briffa et al. 2013 and selected examples in Table 1). Rather than using the least squares regression equation to perform the reconstruction, the proxy data are re-scaled to fit the climate target data by simply giving them the same mean and variance over the calibration period. A simple way to achieve this is to first convert the proxy data into z-scores, by taking the difference between each value and the mean (of the period for which climate data are available) and dividing this by the standard deviation (also of the period with climate data). The z-scored data now have a mean of zero and standard deviation of one over the period with climate data. The z-scoring is then reversed, but rather than using the mean and standard deviation of the proxy, the values from the climate target are used. The result is a climate reconstruction that, over the period of overlap, has exactly the same mean and variance as the climate target data. Variance-scaled reconstructions can capture the full range of measured climate and, we presume, the full range of climate variability in the past. Essentially the same procedure is used in the 'composite plus scaling' method used to collate and calibrate multiple proxies for use in reconstructions over very large geographical regions (e.g. Mann et al. 1999; Moberg et al. 2005; Neukom et al. 2011; Ahmed et al. 2013) and a scaling step is also applied in some spatial field reconstructions (Cook et al. 2007). There is, however, a "cost" involved in that variance-scaling must also increase the mean squared error. Regression-based reconstructions, by definition, produce the minimum MSE that is possible, so any change to the reconstruction must inevitably lead to an increased MSE.

Where variance scaling is applied to climate reconstructions, authors generally acknowledge that there is an increase in error, relative to regression-based reconstructions, but argue that

this is outweighed by the much more realistic range of the reconstructed climate. At present, however, there is no simple method available to quantify the magnitude of the increase in error, or to test whether inflating the range of the reconstructed values results in the correct years being pushed to the extremes. Here we provide a metric, based upon the same logic as the RE and CE statistics, which quantifies the magnitude of increase in error, and therefore relative loss of signal. To compliment this we also provide a simple non-parametric statistical test to determine whether the capture of extremes is significantly better than one would expect to occur simply by chance.

Equivalent variance explained (R_{vs}^2)

A useful way to envisage the strength of a climate reconstruction is to consider how much better it is than simply using the mean climate value for every year (the climatology). A simple measure of strength is the mean squared difference between each pair of measured and estimated values, which is the Mean Squared Error (MSE). If a reconstruction produces a MSE that is no smaller than that obtained when a horizontal line (mean climate) is used, then it clearly possesses little or no skill as a tool for reconstructing past climate. The bigger the difference between the MSE based on the mean (climatology) and the MSE based on the reconstruction the better, and this is effectively what the squared correlation coefficient measures over the period used for calibration, and it can be expressed as:

$$R^2 = 1 - (MSE_{\text{regression estimates}}/MSE_{\text{climatology}}) \quad [1]$$

A climate reconstruction based on least squares regression is automatically scaled so that the MSE is minimised, so any change to the variance of the reconstruction will inevitably increase the MSE. By directly calculating the MSE of the variance-scaled reconstruction, and

comparing it to that obtained using the mean climatology, it is possible to derive a metric that can be viewed in the same way as R^2; as a measure of the amount of variance explained relative to that explained by just using the average for every year:

$$R_{vs}^2 = 1 - (MSE_{\text{scaled estimates}}/MSE_{\text{climatology}}) \quad [2]$$

Applying this logic, one may conclude that a variance-scaled reconstruction that yields an R_{vs}^2 value of zero or less has no skill because just using the mean climate value for every year would give a better estimate (lower MSE).

The approach presented here is exactly the same as that used in calculating the RE and CE statistics, and the equations all have the same form, the only differences being the value that is used to define the climatology and the period over which the comparison is made. The RE and CE metrics are only suitable for a split-period test, where the mean values are taken from the calibration and verification periods respectively and the comparison is always made over the verification period. For R^2 and R_{vs}^2 the mean value is that of the whole period for which climate data are available and the comparison is also made over the whole period. The range for R^2 is one to zero, because least squares regression cannot produce a MSE that is larger than that obtained using the mean value. However, the range for R_{vs}^2, like that for RE and CE, extends below zero because scaling can potentially increase the MSE so much that it becomes larger than that obtained using the climatology.

In fact it is not necessary to calculate R_{vs}^2 directly because the difference between R^2 and R_{vs}^2 is related to the strength of the correlation between the proxy and the climatic target, irrespective of the sign of the correlation, so that it can be defined as a function of the

modulus of pearson's correlation coefficient (|r|), or the square root of the explained variance where:

$$R^2 - R_{vs}^2 = (1 - |r|)^2 \quad [3]$$

Expanding this equation:

$$R^2 - R_{vs}^2 = (|r|)^2 - 2|r| + 1 \quad [4]$$

Which simplifies to:

$$R_{vs}^2 = 2|r| - 1 \quad [5]$$

From this simple equation it is clear that when the (positive or negative) correlation between the proxy and the climate target falls below 0.5, the R_{vs}^2 value falls below zero, and so a variance-scaled reconstruction will have less skill, over the calibration period, than simply using the climatology. The relationship between R^2 and R_{vs}^2 is non-linear, so that for an |r|-value of 0.9, giving an R^2 of 0.81, a variance scaled reconstruction would have an equivalent R_{vs}^2 of 0.8, so the loss of variance explained is very small (1.2%). For an |r|-value of 0.6, however, giving an R^2 of 0.36 the equivalent R_{vs}^2 value for a variance scaled reconstruction is 0.2, so the loss of skill is considerable (44.4%). Given an |r|-value of 0.58 more than half of the skill is lost ($R^2 = 0.34$, $R_{vs}^2 = 0.16$). Since the Pearson's correlation coefficient (r) is the square root of R^2, an equivalent value for a variance-scaled chronology (r_{vs}), taking into account the sense of the relationship , can be derived where r_{vs} is the (positive or negative) square root of Rvs^2.

Capture of extremes

The key aims of variance scaling are to reduce the bias towards the mean inherent in regression-based reconstructions and to ensure that the magnitude of extreme values in the past is not underestimated. A logical measure of how well a reconstruction captures the

extreme values would be to define a threshold for extreme values and calculate how well a reconstruction captures the values beyond this threshold. A simple definition of the extremes is the thresholds beyond which the upper and lower 10% of the measured climate data fall. Since regression-based reconstructions are biased towards the mean, we would expect them to underestimate the number of years where the target climate fell beyond the thresholds for the highest and lowest 10% and a variance scaled reconstruction should perform more successfully.

The 'Extreme Value Capture' (EVC) of a reconstruction can be calculated by first ranking the climate target values to identify the values beyond which the highest and lowest 10% lie and noting which years lie beyond those thresholds. The probability of a value falling within any 10% band of the data, purely by chance, is 1 in 10 (p=0.1). Therefore a reconstruction that has been variance-scaled to have approximately the same range of values as the climate target, but has no skill in capturing extremes beyond pure chance, would still be expected to capture some years. The probability of capturing a given number of extreme years by chance can be calculated using the binomial distribution, providing a simple non-parametric test of statistical significance. The logic and assumptions are the same as those used in applying the Sign Test, the only difference being that the probability of a correct result occurring by chance is 1 in 10 (p=0.1) rather than 1 in 2 (p=0.5). Probabilities can be calculated using a wide variety of software, including Excel (using the BINOM function). One of the assumptions of the Sign Test, and of the EVC test, is that the extremes are independent of each other, and in time series data there can be series autocorrelation which violate this, although it is only likely to be important where autocorrelation is very strong, in which case the data are not really suitable for climate reconstruction using regression or variance scaling.

For example, if a calibration data set is available covering 100 years then there will be 10 years in each of the upper and lower 10% bands. The probability of any year falling into any 10% band is 1 in 10, so even a reconstruction with no skill at all is likely to capture one extreme year in the top 10% and one in the bottom 10%. Using the binomial distribution we can calculate that the probability of capturing 3 years from 10 is p=0.06, and so not statistically significant. However, the probability of capturing 4 years by chance is far more remote (p=0.011) and the probability of capturing 5 is close to one in a thousand (p=0.001). This simple test can be applied to either the two outer 10% bands combined, to give an indication of the overall skill of a reconstruction in capturing extremes, or perhaps more usefully, to the high and low extremes individually, to test for a bias in the ability of a reconstruction to capture either the very high or very low values.

For example, given 150 years of climate data, 10% of which is 15, the critical number of correct captures at p=0.05 and p=0.01 are 4 and 5. In practice however, calculating exact probabilities for the EVC test is more complicated because the number of years with climate data available for calibration will rarely be roundly divisible by ten. In such cases, and as fractional observations are not possible in annually resolved datasets, it is necessary to calculate the significance of the number of captures for the higher and lower integer and combine them using a weighted average. For example, a sample of 154 years requires 15.4 years per 10%, so the significance levels are calculated using the results from both 15 and 16 years using a weighted average where the probability for 16 years is multiplied by 0.4 and the probability for 15 years multiplied by 0.6. To achieve a significance level of p=0.05 in this case requires at least 4 successful extreme value captures. The relative weighting means that they must come from the upper (or lower) 15 years only (indicated in Table 2 as 4/15). The

year ranked 16 cannot be included in the count of 4 captured extremes because 4/16 values is not significant at p=0.05. To achieve a significance level of p=0.01 for 15.4 observations by weighting requires at least 6 correct captures from the upper (or lower) 16 observations, so in this case the year ranked 16 can be included in the count of 6 (indicated in the table as 6/16). The calculations are included for the examples used below, but for convenience and to avoid this procedure we provide a table of critical values (Table 2).

It is important to stress that it is only logical to apply the EVC test to reconstructions that are based either on inverse calibration, where the proxy is the independent variable and underestimation of variance is inevitable, or on matching the variance of the proxy to that of the climate target. In the latter case the total number of years that can possibly fall beyond the upper and lower thresholds is strongly constrained to be very close to 10%, so that the number of 'captures' and the number of erroneous extreme years is effectively a closed set. It is not valid to apply it to reconstructions based on classical correlation, where the proxy is the dependent variable. In this case unexplained variance in the regression model inflates the variance of the reconstruction, so that the number of years that fall beyond the thresholds is not constrained, irrespective of the results of the EVC test.

Examples

Ice-break data and spring temperatures

Amongst the most powerful sources of information on the climate of the past are historical archives, especially those that refer to events that are strongly constrained by climate, such as phenology (e.g. flowering or harvest dates) or the freezing and/or melting of water bodies (Brázdil et al. 2005, 2010). One such record is that of the date of break-up of the winter ice blocking the flow of the Tornio River, close to the Arctic Circle between northern Sweden

and Finland (Kajander 1993; Magnuson et al. 2001; Klingbjer and Moberg 2003). Loader et al. (2011) used this record to reconstruct the mean temperature of April-May back to AD1693, using an overlap with measured climate data of 150 years (AD1860 to 2009). The correlation between the Julian day of ice break and spring temperature over that period is –0.82, so 67% of the variance in break-up date is explained by April-May temperature. Loader et al. (2011) made their reconstruction using regression, but despite the very high correlation, the full range of spring temperatures over the calibration period (-3.9˚C to 5.5˚C = 9.3˚C) is underestimated by about 20% (-2.9˚C to 4.6˚C = 7.4˚C). Using the same data to produce a variance-scaled reconstruction increases the MSE, so that the equivalent variance explained drops from 67% to 64% (R^2 = 0.67, R_{VS}^2 = 0.64), but the range of temperatures is much more realistic (-4.0˚C to 5.1˚C, = 9.1˚C), (Table 3, Fig. 1).

Given 150 years of data the two 10% extremes contain 15 years each. The regression-based reconstruction performs well in capturing the extreme years, with 5 and 6 of the 15 years captured in the highest and lowest 10% bands respectively (Fig. 1, Table 4), both of which are significant (p< 0.01, Table 2), and an overall capture rate of 37% (p< 0.001). Variance scaling raises the capture rate by two years at either extreme (p< 0.001), giving 50% overall, which is a gain of 36%. These results suggest that for a small cost in terms of increased MSE, variance scaling removes the underestimate of the full range of climate variability and also improves the capture rate of extremes. When the ice-break data are used to reconstruct April-May temperatures there appears to be very little bias in skill for reconstructing very warm and very cold years.

Tree growth and summer temperature

McCarroll et al. (2013) present a pine (*Pinus sylvestris* L.) tree growth index for the northern timberline region of Europe based on combining nine tree-growth proxies from four sites. The climatic target used was the June to August mean temperature of the same region, calculated as an average of data from several climate stations. Trees growing in this timberline region are very sensitive to summer temperature, and the mean growth index is based on a very large set of data, so the correlation with summer (June to August) mean temperature is amongst the highest yet reported for any climate reconstruction based on trees ($r = 0.81$).

The tree growth index was used to reconstruct June to August (JJA) temperatures using regression, but as expected there is an underestimate of the variance of summer temperature, with the true range over the calibration period (5.8˚C) being underestimated (4.6˚C) by more than 20% (Table 3). To reduce this bias towards the mean a variance-scaled reconstruction was also produced, where the mean and variance were adjusted to fit the mean and variance of the climatic target data over the whole period for which climate data was available (AD1890 to 2005) and McCarroll et al. (2013) argue that this "gives a more realistic picture of the magnitude of past climate change at the expense of inflating the error". However, the magnitude of error inflation was not quantified and the gains, in terms of the capture of extremes, were not investigated. Using the methods described in this paper we can now say that the reconstruction based on regression explained 66% of the variance ($R^2 = 0.66$) but that variance scaling inflated the MSE, giving an equivalent value ($R_{vs}^2 = 0.62$)) of 62%, which is a signal loss of about 5.5%.

Given 116 years of climate data for calibration, the number of years in the highest and lowest 10% is 11.6. The threshold for the upper 10% of years is 12.6˚C and the regression-based

reconstruction only captures two of the 11 or 12 years by placing them above the threshold (Fig. 1), which is not statistically significant ($p = 0.2$). Using variance scaling captures one more year, but the probability of capturing 3 of the 11 or 12 years is about 8%, so the result is still not statistically significant ($p > 0.05$). The critical threshold for correct captures at $p = 0.05$ for a sample size of 11.6 is 5 from 12 (Table 2).

The regression-based reconstruction performs much better for the coldest 10% of summers, capturing 6 years below the threshold (Fig. 1), which is strongly significant ($p < 0.001$: Table 2), but variance-scaling hardly improves this, with only the 12th year added, which weighted at 0.6 gives a capture rate of 6.6 for 11.6 years (57%). Taking both extremes together, covering 20% of the years, the capture rate is 41% which is strongly significant (n = 23.2, $p < 0.001$, Table 2).

These results demonstrate that even when the correlation between proxy and target is unusually high, at 0.81, the ability of a regression-based reconstruction to capture extreme years is not guaranteed. In this case cold summers are captured much more efficiently than hot summers, to the extent that the capture rate of hot summers is not statistically significant ($p > 0.05$). Variance scaling improves the capture of extremes only very slightly, and the capture of warm summers alone is still not statistically significant ($p > 0.05$). It is not surprising that the growth of pine trees close to the tree line of northern Europe is more sensitive to cold summers than to warm summers, and the 'Extreme Value Capture' method allows this bias to be identified and quantified. In this example the "cost" of variance-scaling, in terms of loss of signal is small, but the gain in terms of the capture of extremes is also small.

Oxygen isotopes in tree rings and summer rainfall

It has been argued, from first principles (Treydte et al. 2014, Barbour 2007; Danis et al. 2006; Saurer et al. 1997), that the oxygen isotope ratios in the latewood cellulose of British oak trees should provide an indication of the amount of summer rainfall (Young et al. In revision). Isotopic data, averaged from several sites across the UK, were calibrated using the total June to August precipitation for the England and Wales region (Wigley et al. 1984) over the period AD1850 to 2012. Given the correlation coefficient of –0.69 a regression-based reconstruction underestimates the true range of precipitation totals (343mm) by more than 28% (245mm), whereas a variance-scaled reconstruction produces a range of values (353mm) within 3% of the target range (Table 3).

The regression-based reconstruction is based on a correlation that explains 48% of the variance, but it performs poorly in capturing the extremes with just 3 and 4 from the upper and lower 10% (Fig. 1), neither of which is statistically significant (Critical threshold for 16.3 years at $p = 0.05$ is 5 from 17: Table 2). Variance-scaling increases the mean squared error, so that the equivalent explained variance ($R_{vs}^2 = 0.38$) drops to 38% (Table 3), however, the skill in capturing the highest and lowest extremes improves considerably (Table 4), with 7 and 9 (43% and 55%) of the upper and lower extremes captured, both of which are strongly significant ($p < 0.001$). In this case variance-scaling improves the capture of extremes by more than 128%. These results suggest that in return for a loss in signal strength (in terms of mean squared error), the range of the reconstruction is much more realistic and the extremes of that range include a significant number of the correct years. There is no evidence of a strong bias and it seems that oxygen isotopes in British oak tree ring cellulose can be used to reconstruct wet and dry summers with approximately equal skill. Stable carbon isotopes from

the same tree rings provide a strong record of summer temperature (Young et al. 2012a, Table 1).

Discussion and Conclusions

One of the central aims of palaeoclimate research is to improve understanding of contemporary climatic changes through comparison of the climate of the past with that of the present. The regression methods that are commonly used, however, inevitably result in an underestimation of the variability of climate in the past and thus a bias towards the mean, with an attendant underestimation of the magnitude and frequency of past extremes. Variance scaling, where the mean and variance of the reconstruction are adjusted to fit the mean and variance of the climate target over the period for which meteorological data are available, overcomes this problem but at the cost of increasing the mean squared error. Authors generally acknowledge the loss of signal, but do not quantify it or demonstrate that there is an improvement in estimating the magnitude and frequency of extreme years.

Just as the squared correlation coefficient (R^2) provides a measure of the strength of a climate reconstruction based on regression, an equivalent measure R_{vs}^2 provides a measure of the strength of a climate reconstruction based on variance scaling. Since a regression-based reconstruction always minimises the MSE, it is inevitable that inflating the variance of the reconstruction to fit the climate target data will increase the error, so R_{vs}^2 must always be lower than R^2 and the difference between them provides a measure of the loss of signal as a result of variance-scaling. Since R_{vs}^2 is a linear function of the modulus of Pearson's correlation coefficient (r as a positive value), the calculation is simple and can be applied retrospectively to estimate the loss of signal in published climate reconstructions based on variance-scaling (Table 1).

When the modulus of the Pearson's correlation coefficient falls below 0.5 ($R^2 < 0.25$), inflating the variance of a reconstruction to meet the climate target results in such a large inflation of the mean squared error that it becomes larger than that obtained simply by using the mean value of the climate data for every year and so, by analogy with the Reduction of Error and Coefficient of Efficiency statistics, the reconstruction can be regarded as having no skill. We conclude therefore that:

- Variance scaling should not be applied where the squared correlation between proxy and target falls below 0.25 ($|r| < 0.5$)

- The amount of variance explained by a scaled reconstruction is less than that explained by a reconstruction based on regression and can be conveniently quantified using R_{vs}^2, which is a linear function of the modulus of Pearson's correlation coefficient (r) where $R_{vs}^2 = 2|r|-1$.

- The loss of skill as a consequence of scaling a reconstruction can be expressed as a percentage $((R^2 - R_{vs}^2)/R^2)*100$

Variance-scaling inevitably expands the range of values in a reconstruction, and this is sometimes quantified by quoting the difference between the total range of measured and estimated values of the climate target value over the period with climate data. Such a comparison is of questionable value, because it is inevitable that variance scaling will perform better than regression in this regard. The critical test is to determine whether the actual years that are pushed to the extremes are the correct years. We propose a simple non-

parametric ‘Extreme Value Capture’ (EVC) test, based on the binomial distribution, which determines whether the number of years correctly placed above and below the thresholds defined by the upper and lower 10% of measured climate data is significantly more than the number that might be expected purely on the basis of chance.

Three examples demonstrate the utility of the EVC test. The most surprising result is that for summer temperature reconstruction in northern Fennoscandia based on a tree-growth index (McCarroll et al. 2013). Although the reconstruction is based on an exceptionally high correlation between proxy and target (June to August temperature) of r = 0.81, the EVC test reveals a clear asymmetry in the ability to capture extreme years. Cold summers are identified very effectively but the capture rate for warm summers is no better than might be expected simply by chance ($p > 0.05$). Such a strong asymmetry in the capture of extremes has important implications for the interpretation of palaeoclimate reconstructions, particularly with regard to comparing the magnitude and frequency of warm extremes in the past with those over the period for which meteorological measurements are available. Missing the warmest years will also result in low-frequency temperature reconstructions of past warm periods being underestimated, since low frequency curves in calibrated reconstructions should simply represent a smoothing of the high frequency signal.

Acknowledgements

This work was supported by C3W and the EU project Millennium (017008) and we thank our many friends in the project for helpful discussion about the perils and intricacies of regression and scaling.

References

Ahmed M, Anchukaitis KJ, Asrat A, et al. (2013) Continental-scale temperature variability during the past two millennia. *Nature Geoscience* 6: 339–346.

Barbour MM. (2007). Stable oxygen isotope composition of plant tissue: a review. *Functional Plant Biology* 34: 83–94.

Brázdil R, Dobrovolný P, Luterbacher J, et al. (2010) European climate of the past 500 years: new challenges for historical climatology. *Climatic Change* 101: 7–40.

Brázdil R, Pfister C, Wanner H, et al. (2005) Historical climatology in Europe - The state of the art. *Climatic Change* 70: 363–430.

Briffa KR, Melvin TM, Osborn TJ, et al. (2013) Reassessing the evidence for tree-growth and inferred temperature change during the Common Era in Yamalia, northwest Siberia. *Quaternary Science Reviews* 72: 83–107.

Büntgen U, Frank DC, Grudd H, et al. (2008) Long-term summer temperature variations in the Pyrenees. *Climate Dynamics* 31: 615–631.

Büntgen U, Frank DC, Nievergelt D, et al. (2006) Summer temperature variations in the European Alps, AD 755-2004. *Journal of Climate* 19: 5606–5623.

Büntgen U, Trouet V, Frank D, et al. (2010) Tree-ring indicators of German summer drought over the last millennium. *Quaternary Science Reviews* 29: 1005–1016.

Cook ER and Kairiukstis LA. (1990) Methods of dendrochronology applications in the environmental sciences. *Cook, E. R. and L. A. Kairiukstis (Ed.). Methods of Dendrochronology: Applications in the Environmental Sciences. Xii+394p. Kluwer*

Academic Publishers: Dordrecht, Netherlands; Boston, Massachusetts, USA. Illus. Maps: XII+394P-XII+394P.

Cook ER, Seager R, Cane MA, et al. (2007) North American drought: Reconstructions, causes, and consequences. *Earth-Science Reviews* 81: 93-134.

Danis PA, Masson-Delmotte V, Stievenard M, et al. (2006). Reconstruction of past precipitation $\delta^{18}O$ using tree-ring cellulose $\delta^{18}O$ and $\delta^{13}C$: a calibration study near Lac d'Annecy, France. *Earth and Planetary Science Letters* 243: 439–448.

D'Arrigo R, Wilson R, Palmer J, et al. (2006) The reconstructed Indonesian warm pool sea surface temperatures from tree rings and corals: Linkages to Asian monsoon drought and El Niño-Southern Oscillation. *Paleoceanography* 21, PA3005.

Esper J, Wilson RJS, Frank DC, et al. (2005) Climate: past ranges and future changes. *Quaternary Science Reviews* 24: 2164–2166.

Kiss A, Wilson R and Bariska I. (2011) An experimental 392-year documentary-based multi-proxy (vine and grain) reconstruction of May-July temperatures for KAszeg, West-Hungary. *International Journal of Biometeorology* 55: 595–611.

Kajander J. (1993) Methodological aspects on river cryophenology exemplified by a tricentennial break-up time series from Tornio. *Geophysica* 29: 73–95.

Klingbjer P and Moberg A. (2003) A composite monthly temperature record from Tornedalen in northern Sweden, 1802-2002. *International Journal of Climatology* 23: 1465–1494.

Linan ID, Buentgen U, Gonzalez-Rouco F, et al. (2012) Estimating 750 years of temperature variations and uncertainties in the Pyrenees by tree-ring reconstructions and climate simulations. *Climate of the Past* 8: 919–933.

Liu JJ, Yang B, Huang K, et al. (2012) Annual regional precipitation variations from a 700 year tree-ring record in south Tibet, western China. *Climate Research* 53: 25–41.

Loader NJ, Jalkanen R, McCarroll D, et al. (2011) Spring temperature variability in northern Fennoscandia AD 1693-2011. *Journal of Quaternary Science* 26: 566–570.

Loader NJ, Young GHF, Grudd H, et al. (2013) Stable carbon isotopes from Torneträsk, northern Sweden provide a millennial length reconstruction of summer sunshine and its relationship to Arctic circulation. *Quaternary Science Reviews* 62: 97–113.

Magnuson JJ. (2001) Historical trends in lake and river ice cover in the Northern Hemisphere (vol 290, pg 1743, 2000). *Science* 291: 254–254.

Mann ME, Bradley RS and Hughes MK. (1999) Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophysical Research Letters* 26: 759–762.

McCarroll D, Loader NJ, Jalkanen R, et al. (2013) A 1200-year multiproxy record of tree growth and summer temperature at the northern pine forest limit of Europe. *Holocene* 23: 471–484.

Moberg A, Sonechkin DM, Holmgren K, et al. (2005) Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data. *Nature* 433: 613–617.

National Research Council (2007) *Surface temperature reconstructions for the last 2,000 years*. The National Academies Press: Washington DC.

Neukom R, Luterbacher J, Villalba R, et al. (2011) Multiproxy summer and winter surface air temperature field reconstructions for southern South America covering the past centuries. *Climate Dynamics* 37: 35–51.

Opala M and Mendecki MJ. (2014) An attempt to dendroclimatic reconstruction of winter temperature based on multispecies tree-ring widths and extreme years chronologies (example of Upper Silesia, Southern Poland). *Theoretical and Applied Climatology* 115: 73–89.

Poljanšek S, Ceglar A and Levanič T. (2013) Long-term summer sunshine/moisture stress reconstruction from tree-ring widths from Bosnia and Herzegovina. *Climate of the Past* 9: 27–40.

Popa I and Kern Z. (2009) Long-term summer temperature reconstruction inferred from tree-ring records from the Eastern Carpathians. *Climate Dynamics* 32: 1107–1117.

Rinne KT, Loader NJ, Switsur VR, et al. (2013) 400-year May-August precipitation reconstruction for Southern England using oxygen isotopes in tree rings. *Quaternary Science Reviews* 60: 13–25.

Sano M, Ramesh R, Sheshshayee MS, et al. (2012) Increasing aridity over the past 223 years in the Nepal Himalaya inferred from a tree-ring delta O-18 chronology. *Holocene* 22: 809–817.

Saurer M, Borella S, Leuenberger M. (1997) $\delta^{18}O$ of tree rings of beech (*Fagus sylvatica*) as a record of $\delta^{18}O$ of the growing season precipitation. *Tellus Series B–Chemical and Physical Meteorology* 49: 80–92.

Treydte, K, Boda, S, Graf Pannatier E, et al. (2014) Seasonal transfer of oxygen isotopes from precipitation and soil to the tree ring: source water versus needle water enrichment. *New Phytologist* 202: 772–783.

Trouet V, Panayotov MP, Ivanova A, et al. (2012) A pan-European summer teleconnection mode recorded by a new temperature reconstruction from the northeastern Mediterranean (AD1768-2008). *Holocene* 22: 887–898.

von Storch H, Zorita E, Jones JM, et al. (2004) Reconstructing past climate from noisy data. *Science* 306: 679–682.

Wigley TML, Briffa KR and Jones PD. (1984) On the average value of correlated time-series, with applications in dendroclimatology and hydrometeorology. *Journal of Climate and Applied Meteorology* 23: 201–213.

Yang B, Kang X, Liu J, et al. (2010) Annual temperature history in Southwest Tibet during the last 400 years recorded by tree rings. *International Journal of Climatology* 30: 962–971.

Young GHF, Bale RJ, Loader NJ, et al. (2012a) Central England temperature since AD 1850: the potential of stable carbon isotopes in British oak trees to reconstruct past summer temperatures. *Journal of Quaternary Science* 27: 606–614.

Young GHF, McCarroll D, Loader NJ, et al. (2012b) Changes in atmospheric circulation and the Arctic Oscillation preserved within a millennial length reconstruction of summer cloud cover from northern Fennoscandia. *Climate Dynamics* 39: 495–507.

Figure captions

Figure 1. Comparison of measured (smooth black lines, presented in rank order) and reconstructed climate parameters using regression and scaling. The red lines and dashed boxes show the extent and target range for the lowest and highest 10% of measured values. A: Spring temperature reconstructed using ice break dates on the Tornio River, B: Summer temperature reconstructed using tree growth in northern Fennoscandia and C: Summer precipitation reconstructed using stable oxygen isotopes in British oak tree rings.

Table 1. A selection of publications that have used variance-scaling to reconstruct past climate with Pearson's correlation coefficient (r) and squared correlation coefficient (R^2), the equivalents for a variance-scaled reconstruction (r_{vs} and R_{vs}^2) and the percentage loss of signal due to variance-scaling.

Reference	Proxy	Target	r	R^2	R_{vs}^2	r_{vs}	% loss
McCarroll et al. (2013)	Tree growth	June-August Temperature	0.81	0.66	0.62	0.79	5.5
Kiss et al. (2011)	Documentary (vine & grain)	May-July temperature	0.75	0.57	0.51	0.71	10.5
Rinne et al. (2013)	Tree ring $\delta^{18}O$	May-August Precipitation	–0.73	0.53	0.46	–0.68	13.7
Young et al. (2012a)	Tree ring $\delta^{13}C$	June-August Temperature	0.69	0.480	0.39	0.62	20.0
Büntgen et al. (2006)	Tree ring density	June-September Temperature	0.69	0.48	0.38	0.62	20.0
Young et al. (in Revision)	Tree ring $\delta^{18}O$	June-August Precipitation	–0.69	0.48	0.39	–0.62	20.0
Young et al. (2012b)	Tree ring $\delta^{13}C$	June-July % cloud cover	–0.67	0.42	0.34	–0.58	19.0
Trouet et al. (2012)	Tree ring MXD	June-August Temperature	0.65	0.42	0.30	0.55	28.6
Dorado-Liñán et al. (2012)	Tree ring width & density	May-September Temperature	0.62	0.38	0.24	0.49	37.6
Yang et al. (2010)	Tree ring width	12 month Precipitation	0.59	0.35	0.18	0.42	48.3
Sano et al. (2011)	Tree ring $\delta^{18}O$	June-September PDSI	–0.58	0.34	0.16	–0.40	52.4
Liu et al. (2012)	Tree ring width	12 month Precipitation	0.56	0.31	0.12	0.35	61.7
Poljanšek et al. (2013)	Tree ring width	June-July Sunshine	0.54	0.29	0.08	0.28	72.6
Büntgen et al. (2008)	Tree ring width & density	May-September Max.Temperature	0.53	0.28	0.06	0.24	78.6
Opala & Mendecki(2014)	Tree ring width	Winter Temperature	0.43 - 0.47	0.18 - 0.22	<0	<0	>100
Popa & Kern (2009)	Tree ring width	June-August Temperature	0.43	0.18	0.14	<0	175.7
Büntgen et al. (2010)	Tree ring width	June-September scPDSI	-0.42	0.18	-0.16	<0	190.7

Table 2. Critical values for the 'Extreme Value Capture' test where *p* is the probability of capturing at least the tabulated number of correct years. N represents the highest or lowest 10% of years or their combination.

N	0			.1			.2			.3			.4		
p<	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
3	2/3	3/3	3/3	2/3	3/3	3/3	2/3	3/3	--	2/3	3/3	--	2/3	3/3	--
4	2/4	3/4	4/4	3/5	3/4	4/4	3/5	3/4	4/4	3/5	3/4	4/5	3/5	3/4	4/5
5	3/5	3/5	4/5	3/5	3/5	4/5	3/5	3/5	4/5	3/5	3/5	4/5	3/6	4/6	4/5
6	3/6	4/6	4/6	3/6	4/6	4/6	3/6	4/6	4/6	3/6	4/6	5/7	3/6	4/7	5/7
7	3/7	4/7	5/7	3/7	4/7	5/7	3/7	4/7	5/7	3/7	4/7	5/7	3/7	4/7	5/7
8	3/8	4/8	5/8	3/8	4/8	5/8	3/8	4/8	5/8	3/8	4/8	5/8	3/8	4/8	5/8
9	3/9	4/9	5/9	3/9	4/9	5/9	3/9	4/9	5/9	3/9	4/9	5/9	3/9	4/9	5/9
10	4/10	5/10	5/10	4/10	5/11	6/11	4/11	5/11	6/11	4/11	5/11	6/11	4/11	5/11	6/11
11	4/11	5/11	6/11	4/11	5/11	6/11	4/11	5/11	6/11	4/11	5/11	6/11	4/11	5/11	6/11
12	4/12	5/12	6/12	4/12	5/12	6/12	4/12	5/12	6/12	4/12	5/12	6/12	4/12	5/12	6/12
13	4/13	5/13	6/13	4/13	5/13	6/13	4/13	5/13	6/13	4/13	5/13	6/13	4/13	5/13	6/13
14	4/14	5/14	6/14	4/14	5/14	6/14	4/14	5/14	6/14	4/14	5/14	6/14	4/14	5/14	7/15
15	4/15	5/15	7/15	4/15	6/16	7/15	4/15	6/16	7/15	4/15	6/16	7/16	4/15	6/16	7/16
16	5/16	6/16	7/16	5/17	6/16	7/16	5/17	6/16	7/16	5/17	6/16	7/16	5/17	6/17	7/16
17	5/17	6/17	7/17	5/17	6/17	7/17	5/17	6/17	7/17	5/18	6/17	7/17	5/18	6/17	7/17
18	5/18	6/18	7/18	5/18	6/18	7/18	5/18	6/18	7/18	5/18	6/18	7/18	5/18	6/18	7/18
19	5/19	6/19	7/19	5/19	6/19	7/19	5/19	6/19	8/20	5/19	6/19	8/20	5/19	6/19	8/20
20	5/20	6/20	8/20	5/20	6/20	8/20	5/20	6/20	8/20	5/20	6/20	8/20	5/20	6/20	8/21
21	5/21	7/21	8/21	5/21	7/22	8/21	5/21	7/22	8/21	5/21	7/22	8/21	5/21	7/22	8/21
22	5/22	7/22	8/22	5/22	7/22	8/22	5/22	7/22	8/22	5/22	7/22	8/22	5/22	7/23	8/22
23	6/23	7/23	8/23	6/24	8/23	8/23	6/24	7/23	8/23	6/24	7/23	8/23	6/24	7/23	8/23
24	6/24	7/24	8/24	6/24	7/24	8/24	6/24	7/24	8/24	6/25	7/24	8/24	6/25	7/24	9/25
25	6/25	7/25	9/25	6/25	7/25	9/25	6/25	7/25	9/25	6/25	7/25	9/26	6/26	7/25	9/26
26	6/26	7/26	9/26	6/26	7/26	9/26	6/26	7/26	9/26	6/26	7/26	9/26	6/26	7/26	9/26
27	6/27	8/27	9/27	6/27	8/28	9/27	6/27	8/28	9/27	6/27	8/28	9/27	6/27	8/28	9/27
28	6/28	8/28	9/28	6/28	8/28	9/28	6/28	8/28	9/28	6/28	8/29	9/28	6/28	8/29	9/28
29	6/29	8/29	9/29	6/29	8/29	9/29	6/29	8/29	9/29	6/29	8/29	9/29	6/29	8/29	9/29
30	6/30	8/30	10/30	6/30	8/30	10/31	6/30	8/30	10/31	6/30	8/30	10/31	6/30	8/30	10/31
31	7/31	8/31	10/31	7/32	8/31	10/31	7/32	8/31	10/31	7/32	8/31	10/31	7/32	8/31	10/32
32	7/32	8/32	10/32	7/32	8/32	10/32	7/32	8/32	10/32	7/33	8/32	10/32	7/33	8/32	10/32
33	7/33	8/33	10/33	7/33	8/33	10/33	7/33	8/33	10/33	7/33	8/33	10/33	7/33	9/34	10/33
34	7/34	9/34	10/34	7/34	9/34	10/34	7/34	9/34	10/34	7/34	9/35	10/34	7/34	9/35	10/34
35	7/35	9/35	10/35	7/35	9/35	10/35	7/35	9/35	10/35	7/35	9/35	10/35	7/35	9/36	10/35
36	7/36	9/36	11/36	7/36	9/36	11/36	7/36	9/36	11/37	7/36	9/36	11/37	7/36	9/36	11/37
37	7/37	9/37	11/37	7/37	9/37	11/37	7/37	9/37	11/37	7/37	9/37	11/37	7/37	9/37	11/38
38	7/38	9/38	11/38	7/38	9/38	11/38	7/38	9/38	11/38	7/38	9/38	11/38	7/38	9/38	11/38
39	8/39	9/39	11/39	8/40	9/39	11/39	8/40	9/39	11/39	8/40	9/39	11/39	8/40	9/39	11/39
40	8/40	9/40	11/40	8/40	10/41	11/40	8/40	10/41	11/40	8/41	10/41	11/40	8/41	10/41	11/40

N	.5			.6			.7			.8			.9		
p<	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
3	2/3	3/3	--	2/3	3/3	--	2/3	3/3	--	2/3	3/4	--	2/3	3/4	--
4	3/5	3/4	4/4	3/5	3/4	4/4	3/5	3/4	4/5	3/5	3/4	4/5	3/5	3/4	4/5
5	3/6	4/6	4/5	3/6	4/6	4/5	3/6	4/6	4/5	3/6	4/6	4/5	3/6	4/6	4/5
6	(3)	4/7	5/7	(3)	4/7	5/7	3/7	4/7	5/7	3/7	4/7	5/7	3/7	4/7	5/7
7	3/7	4/7	5/8	3/7	4/7	5/8	3/7	4/8	5/8	3/7	4/8	5/8	3/8	4/8	5/8
8	3/8	4/8	5/8	3/8	4/8	5/8	3/8	4/8	5/8	3/8	4/8	5/8	3/8	4/9	5/9
9	4/10	4/9	5/9	4/10	4/9	5/9	4/10	4/9	5/9	4/10	4/9	5/9	4/10	5/10	5/9
10	4/11	5/11	6/11	4/11	5/11	6/11	4/11	5/11	6/11	4/11	5/11	6/11	4/11	5/11	6/11
11	4/12	5/12	6/12	4/12	5/12	6/12	4/12	5/12	6/12	4/12	5/12	6/12	4/12	5/12	6/12
12	4/12	5/12	6/12	4/12	5/12	6/12	4/13	5/13	6/13	4/13	5/13	6/13	4/13	5/13	6/13
13	4/13	5/13	6/13	4/13	5/13	6/13	4/13	5/13	6/13	4/14	5/13	6/13	4/14	5/14	6/13
14	4/14	5/14	7/15	4/14	5/14	7/15	4/14	5/14	7/15	4/14	5/14	7/15	4/15	5/14	7/15
15	4/15	6/16	7/16	4/15	6/16	7/16	4/15	6/16	7/16	4/15	6/16	7/16	5/16	6/16	7/16
16	5/17	6/17	7/16	5/17	6/17	7/16	5/17	6/17	7/17	5/17	6/17	7/17	5/17	6/17	7/17
17	5/18	6/17	7/17	5/18	6/18	7/17	5/18	6/18	7/17	5/18	6/18	7/17	5/18	6/18	7/18
18	5/19	6/18	7/18	5/19	6/18	7/18	5/19	6/18	7/18	5/19	6/19	7/18	5/19	6/19	7/18
19	5/19	6/19	8/20	5/19	6/19	8/20	5/20	6/19	8/20	5/20	6/19	8/20	5/20	6/19	8/20
20	5/20	6/20	8/21	5/20	6/20	8/21	5/20	6/20	8/21	5/21	7/21	8/21	5/21	7/21	8/21
21	5/21	7/22	8/21	5/21	7/22	8/21	5/21	7/22	8/22	5/21	7/22	8/22	5/22	7/22	8/22
22	5/22	7/23	8/22	5/22	7/23	8/22	5/22	7/23	8/22	5/22	7/23	8/22	5/22	7/23	8/23
23	6/24	7/23	8/23	6/24	7/24	8/23	6/24	7/24	8/23	6/24	7/24	8/23	6/24	7/24	8/23
24	6/25	7/24	9/25	6/25	7/24	9/25	6/25	7/24	9/25	6/25	7/25	9/25	6/25	7/25	9/25
25	6/26	7/25	9/26	6/26	7/25	9/26	6/26	7/25	9/26	6/26	7/25	9/26	6/26	7/26	9/26
26	6/26	7/26	9/26	6/27	7/26	9/27	6/27	7/26	9/27	6/27	7/26	9/27	6/27	8/27	9/27
27	6/27	8/28	9/27	6/27	8/28	9/27	6/28	8/28	9/27	6/28	8/28	9/28	6/28	8/28	9/28
28	6/28	8/29	9/28	6/28	8/29	9/28	6/28	8/29	9/28	6/28	8/29	9/28	6/29	8/29	9/29
29	6/29	8/29	9/29	6/29	8/30	9/29	6/29	8/30	9/29	6/29	8/30	9/29	6/29	8/30	10/30
30	6/30	8/30	10/31	7/31	8/30	10/31	7/31	8/31	10/31	7/31	8/31	10/31	7/31	8/31	10/31
31	7/32	8/31	10/32	7/32	8/31	10/32	7/32	8/31	10/32	7/32	8/31	10/32	7/32	8/32	10/32
32	7/33	8/32	10/32	7/33	8/32	10/33	7/33	8/32	10/33	7/33	8/32	10/33	7/33	8/32	10/33
33	7/34	9/34	10/33	7/34	9/34	10/33	7/34	9/34	10/33	7/34	9/34	10/34	7/34	9/34	10/34
34	7/34	9/35	10/34	7/35	9/35	10/34	7/35	9/35	10/34	7/35	9/35	10/34	7/35	9/35	10/34
35	7/35	9/36	10/35	7/35	9/36	11/36	7/36	9/36	11/36	7/36	9/36	11/36	7/36	9/36	11/36
36	7/36	9/36	11/37	7/36	9/37	11/37	7/36	9/37	11/37	7/36	9/37	11/37	7/37	9/37	11/37
37	7/37	9/37	11/38	7/37	9/37	11/38	7/37	9/37	11/38	7/37	9/38	11/38	7/37	9/38	11/38
38	7/38	9/38	11/38	8/39	9/38	11/38	8/39	9/38	11/39	8/39	9/38	11/39	8/39	9/39	11/39
39	8/40	9/40	11/39	8/40	9/40	11/39	8/40	9/40	11/39	8/40	9/40	11/40	8/40	9/40	11/40
40	8/41	10/41	11/40	8/41	10/41	11/40	8/41	10/41	11/40	8/41	10/41	11/40	8/41	10/41	11/40