Testing the hockey-stick hypothesis by statistical analyses of a large dataset of proxy records.

Travaglini, Guido

Università di Roma 1

23 April 2014

Pattern Recognition in Physics

Pattern Recogn. Phys., 2(2), 36-63.

http://www.pattern-recognition-in-physics.com

Testing the hockey-stick hypothesis by statistical analyses of a large dataset of proxy records

[1]Guido Travaglini

Abstract

This paper is a statistical time-series investigation addressed at testing the anthropogenic climate change hypothesis known as the "hockey-stick". The time-series components of a select batch of 258 long-term yearly Climate Change Proxies (CCP) included in 19 paleoclimate datasets, all of which running back as far as the year 2192 B.C., are reconstructed by means of univariate Bayesian Calibration. The instrumental temperature record utilized is the Global Best Estimated Anomaly (BEA) of the HADCRUT4 time series readings available yearly for the period 1850-2010. After performing appropriate data transformations, Ordinary Least Squares parameter estimates are obtained, and subsequently simulated by means of multi-draw Gibbs sampling for each year of the pre-1850 period. The ensuing Time-Varying Parameter sequence is utilized to produce high-resolution calibrated estimates of the CCP series, merged with BEA to yield Millennial-scale Time Series (MTS). Finally, the MTS are individually tested for temperature single break date and multiple peak dates. As a result, the estimated temperature breaks and peaks suggest widespread rejection of the hockey-stick hypothesis since they are mostly centered in the Medieval Warm Period.

Research Areas: Earth Science — Climate Change
Received: 10/Aug/2013 - **Revised:** 13/Dec/2013 - **Accepted:** 09/Jan/2014 - **Published:** 23/Apr/2014
[1]Dipartimento di Scienze Economiche e Giuridiche, Università di Roma "La Sapienza", 00181, Italy.
Correspondence to: G. Travaglini (jay_of_may@yahoo.com).

1. Introduction

More than a decade ago some climatologists led by Michael Mann, after performing past temperature reconstructions on a millennial scale, have come up with the conclusion that the Recent Warming Period (RWP) is an unprecedented phenomenon in the climatic history of the Earth (Mann *et al.*, 1998, 1999). The unusual behavior of recorded temperatures in the late 20th century was attributed by the authors to anthropogenic influences, and chiefly to substantial hikes in the recorded greenhouse gas concentrations caused by the worldwide expansion of industrial activities and to the sharp world population increase.

The authors produce statistical evidence graphically shaped as a hockey-stick that has been prominently featured in the Nobel-prized Intergovernmental Panel on Climate Change (IPCC) activity since the Third Assessment Report (IPCC, 2001). This evidence spurred a worldwide dispute on both the validity of the empirical evidence and on its causes. By consequence, the purported dramatic rise of recent temperatures and the associated anthropogenic origin have found advocates and skeptics still to date igniting the "hockey-stick curve" controversy (Montford, 2010).

Criticism of the anthropogenic origins of global warming includes studies questioning the methods utilized for temperature reconstructions (Baliunas and Soon, 2003; McKitrick, 2006; McIntyre and McKitrik, 2006, 2009), and other studies pointing to the prevalence of long-run natural causes such as solar activity (Abdussamatov, 2004; Alanko-Huotari, 2006; Fouka *et al.*, 2006), cosmic rays (Shaviv, 2005; Svensmark and Frijs-Christensen, 2007; Bard and Frank, 2006; Usoskin *et al.*, 2004a, 2004b, 2006), ocean currents (Gray *et al.*, 1997; Trouet *et al.*, 2009), and volcanic activity (Shindell *et al.*, 2004).

In this context, probably the only consensus among the opposing sides is couched in terms of the available evidence of climate changes on a millennial scale that may inform on the role of anthropogenic forcing in the RWP. (*e.g.* Folland *et al.* 2001). In fact, the lack of widespread instrumental surface temperature estimates prior to the mid-19th century has placed particular emphasis on the need to accurately track the history of climate changes, which can only be achieved by utilizing carefully reconstructed long-term empirical evidence (Jones *et al.*, 2001; Von Storch *et al.*, 2004; Mann *et al.*, 2008, 2009). Such evidence consists of long-term Climate Change Proxies (CCP) which may shed some light on the difference between natural and anthropogenic influences on the climate system and enable statistical inferences on millennial-scale anomalies, such as the Medieval Warming Period (MWP) and the RWP.

Many regional or global sea and/or surface temperature reconstructions are now available customarily utilizing proxies of climate variability derived from the environment; and from documentary evidence (Crowley, 1991; Bradley, 1999; Jones *et al.*, 2001; Guiot *et al.*, 2010). Particularly useful are the high-resolution proxies such as tree rings (Fritts *et al.*, 1971; Fritts, 1991), corals (Evans *et al.*, 2002; Hendy *et al.*, 2002), ice cores (Appenzeller *et al.*, 1998), lake sediments (Loehle, 2007; Loehle and McCulloch, 2008), oceanic oscillations (Li *et al.*, 2011; Wilson *et al.*, 2006), and many more reported in the Reference Section.

All of the reconstruction methods contain a sizable element of uncertainty, usually determined by red-noise behavior. This is the reason why very different results can be obtained by using the same or similar methodological approaches (Christiansen *et al.*, 2009; Christiansen and Ljungqvist, 2011; Christiansen, 2011; McShane and Wyner, 2011; Smerdon *et al.*, 2008). The method based on Gibbs sampling that is utilized in this paper significantly reduces this kind of risk and avoids all too common calibration pitfalls (Mann *et al.*, 1999).

The ample taxonomical and geographical diversity of the CCP series contained in the 19 datasets supplied offers enough information for testing the hockey-stick hypothesis on a global scale, in spite of the official statements of the IPCC (2007) according to which the reconstructed estimates of the MWP are significantly heterogenous because regionally confined to Northern Europe, Northern America and Greenland (Crowley and Lowery, 2000; Bradley *et al.*, 2001; Folland *et al.*, 2001; Esper *et al.*, 2002; D'Arrigo *et al.*, 2006; Juckes et al., 2007; Mann *et al.*, 2009; Graham *et al.*, 2010).

However, several more or less recent studies reverse this conclusion by proving that the MWP was a global phenomenon (Broecker, 2001; Cook *et al.*, 2002; Zunli *et al.*, 2012; Scafetta, 2013). In addition to this evidence, a vast amount of papers - more than 1,000 - reporting results obtained from applied research in the six continents demonstrate that the MWP was a global phenomenon that has indeed been warmer than the RWP. References are available at: http://www.co2science.org/data/mwp.

Most of the CCP series scrutinized in the paper are longer than one millennium and thus meet the MWP requirement (~900–1350 A.D.). Also, quite a few series include previous likely warmings, like the Roman Warm Period (Ljungqvist, 200) to be compared with the RWP (Crowley and Lowery, 2000; Bradley *et al.*, 2001; Baliunas and Soon, 2003; Loso, 2008; Esper and Frank, 2009; Graham *et al.*, 2010; Tingley and Li, 2012). The series also include the Maunder Minimum, *i.e.*, the European Little Ice Age (~1645–1715 A.D.) as a relevant event of the climatic cyclical pattern recognized by several authors (*e.g.* Baliunas and Soon, 2003; Büntgen *et al.*, 2011).

The instrumental temperatures utilized for calibration are the medians of the global Best Estimated Anomaly (BEA) of the HADCRUT4 gridded dataset, which includes the CRUTEM4 land-surface air temperature and the HadSST3 sea-surface temperature datasets (Morice at al., 2012). The recorded readings of ensemble medians cover the period 1850–2010.The authors supply four different geographical areas of the readings: Northern Hemisphere (NH), Southern Hemisphere (SH), Tropics and Global (GL), which is the mean of NH+SH. In the present context, although the vast majority of the CC series has been recorded in the NH, the GL series is utilized with no loss of generality, especially because the correlation coefficients between NH and SH exceeds 0.97.

Sect. II tackles the classical calibration method and its associated likelihood of obtaining hockey-stick behavior (HSB) in the presence of a statistically inconsistent Millennial-scale Time Series (MTS) made of calibrated and nonstationary actual time series (Noriega and Ventosa-Santaulària, 2007; Ventosa-Santaulària, 2009). Sect. III discusses the superiority of the Bayesian calibration technique over its classical counterpart and illustrates the principles of Gibbs sampling of the two-way Time Varying Parameter (TVP) Kalman Filter (KF) model for state prediction in a State-Space (SS) context (Kalman, 1960; Carter and Kohn, 1994).Sect. IV illustrates the stepwise procedure necessary to produce correct calibration estimators and statistically consistent pre- and post-1850 series that enter the MTS. Sect. V introduces both the BEA and CCP series by displaying their timelines and their major descriptive characteristics and also produces the results of the TVP parameter estimation and the median single break dates as well as the temperature multiple peak dates of the MTS. Sect. VI concludes.

2. Classical Calibration and Hockey-Stick Behavior

Classical Calibration (CC) and Bayesian Calibration (BC) are methods utilized in computer engineering, climate modeling, economic forecasting, and other disciplines such as chemistry and biology (Kennedy and O'Hagan, 2001; O'Keefe and Kueter, 2004; Sansó and Forest, 2009; Cooley, 1997).

The CC univariate problem (*e.g.* Juckes *et al.*, 2007) consists of finding some Time-Fixed Parameter (TFP) set B that "best" reproduces an actual time series y_τ, observed over a given timespan $\tau \in \mathrm{T}$, into a virtual time series $\hat{Y}_t$ defined over a different but adjacent timespan $t \in T$. In SS jargon, the two variables are respectively defined as the "observable" and as the "state". We have $\tau \neq t$, $\mathrm{T} \neq T$ and either $t = \mathrm{T}+1, \mathrm{T}+2, ..., \mathrm{T}+T$ for the forward or leading indicator (Stock and Watson, 1989; Banerjee *et al.*, 2005), or $t = \tau - 1$, $\tau - 2, ..., 1-\mathrm{T}$ for adaptive backward reconstruction (*e.g.* Mann *et al.*, 1998). In both cases, an observable dataset Y_s is produced, where $s \in [1, S]$ and $S = T + \mathrm{T}$ that spans the timeline of both adjacent components in either direction.

For the specific case of adaptive backward reconstruction, we have the following parameter and data sizes

$$B : (n \times 1) \tag{1}$$

$$y_\tau : (\mathrm{T} \times 1) \tag{2}$$

$$Y_\tau : (\mathrm{T} \times n) \tag{3}$$

$$\hat{B} = (Y_\tau ' Y_\tau)^{-1} Y_\tau ' y_\tau \tag{4}$$

where $n \geq 2$ includes a constant term and $\hat{B}$ is the direct Ordinary Least Squares (OLS) estimator of the size of (1). In addition, we have

$$Y_t : (T \times n) \tag{5}$$

$$\hat{Y}_t : (T \times 1) \tag{6}$$

where (5) and (6) are bound together by means of the following conventional fit

$$\hat{Y}_t = Y_t \hat{B} \tag{7}$$

where $\hat{Y}_t$ is the virtual or calibrated time series. Concatenating (7) and (2) produces the following MTS

$$\hat{Y}_s = \left\{ \hat{Y}_t, y_\tau \right\}' \tag{8}$$

where $\hat{Y}_s : (S \times 1)$.

This calibration procedure produces statistically consistent results if the correlation among y_τ and Y_τ is significant. However, the risk of obtaining fitted values in (7) that do not consider data uncertainties (Von Storch *et al.*, 2004; Scafetta, 2013) and the likelihood of obtaining HSB in (8) are very high. In fact, after defining the standard deviations of y_τ and $\hat{Y}_t$ respectively as $\sigma(y_\tau)$ and $\sigma(\hat{Y}_t)$, we have

$$(\hat{Y}_t) = f \cdot \sigma(y_\tau) \tag{9}$$

where $f \propto |\hat{B}|$ and $f \approx 0$ if the slope parameter of $|\hat{B}| \approx 0$ as shown in Appendix A. Since $\infty > f > 0$, f may be comfortably defined as a percent ratio of the left-hand-side variable with respect to the right-hand-side variable of (9).

The HSB, better defined as an abrupt change in mean and variance occurring at some point in time of the series under scrutiny, may lead to erroneous statistical interpretations and inferences of (8) (McKitrick, 2006; McIntyre and McKitrik, 2006, 2009; Ventosa-Santaulària, 2009). The occurrence of $f \approx 0$ conditional on the estimated slope parameter of $|\hat{B}| \approx 0$ originates from nonstationarity of the series y_τ, and the risk of HSB increases as the series $\hat{Y}_t$ is stationary and even more so if nonnormal. This can be numerically proven by Monte Carlo simulations on two different distributional and two integration specifications regarding $\hat{Y}_t$ after letting y_τ be a Random Walk (RW).

The HSB is stated as a null hypothesis that can be tested over the entire time series Y_s of (8). Needless to specify, this procedure automatically implies testing for the abovementioned hockey-stick hypothesis with real-world data as performed in Sections 3 and 4.

Table I exhibits, for $\hat{Y}_t$ given by (7), the probability of HSB in (8) with select values of f in (9) ranging from 1 per mille to 100%, namely, for $1 \geq f \geq .001$. For this purpose, 10,000 Monte Carlo draws are produced for a normal and nonnormal, both nonstationary and stationary, Data Generation Process (DGP) of $\hat{Y}_t$ with $T = 1{,}000$. Nonnormality is set by arbitrary nonzero skewness and a kurtosis of 4. Then, for each value of f and given a prior $\sigma\left(y_\tau\right) = 0.261$ as from Table II, col. (1), the HSB probabilities are estimated for normal and nonnormal DGP of $\hat{Y}_t$ with zero mean and select skewness.

From Table I, the normal DGP (col. 1) shows that, for $f \leq 5\%$ in the nonstationarity case and for $f \leq 50\%$ in the stationary case, the probability of HSB in (8) is over 90%. With nonnormal DGPs, the first cutoff rate stays constant, but the second rises to 60% and 70% as skewness is made to rise (cols. 2-3). In practice, nonnormality and stationarity of $\hat{Y}_t$ coupled with nonstationarity of y_τ produce a sizable probability of achieving HSB. This occurs even with relatively large values of f and of the slope coefficient in the parameter set $\left|\hat{B}\right|$.

A proof of these findings is obtainable from replication of the computations performed by Mann and co-authors (Mann *et al.*, 1998) by using older BEA series (HADCRUT2) available from 1856 to 2001 together with fifteen proxy variables available from 1400 to 1980. Define these as y_τ and Y_τ, respectively. These variables are respectively found to be not significantly stationary and nonstationary at the 5% level by means of both the Augmented Dickey-Fuller (ADF) test for first-difference stationarity (Said and Dickey, 1984; Elliott *et al.*, 1996) and the KPSS test for trend stationarity (Kwiatkowski *et al.*, 1992).

Moreover, the vast majority of the proxies are found to be normally distributed by means of the Jarque-Bera (JB) test (Jarque and Bera, 1987). Finally, we have $\sigma\left(y_\tau\right) = 0.180$. From direct OLS of the synchronous observables over time $\tau \in \mathrm{T}$ the mean slope value of $\hat{B}$ is found to be equal to -0.016, and the mean absolute value of the t-statistic is 2.42. From (7), we have $\sigma\left(\hat{Y}_t\right) = 0.094$ such that finally $f = 0.524$ in (9). The value of the percent factor is almost equal to the cutoff rate of 0.5 reported in Table 1, col. 1. This implies that the variables utilized by Mann and co-authors produce a probability of HSB close to 98% in (8).

3. Bayesian Calibration and Gibbs Sampling

The occurrence of abnormally low OLS estimators together with nonstationarity of the instrumental series was shown to be conducive to HSB with high probability. To avoid this pitfall, its causes must be corrected for by producing a MTS that includes mutually statistically consistent series. The present Section discusses the properties of BC and of Monte Carlo Markov Chain (MCMC) simulation for the purpose of producing correct calibration, while the following Section illustrates the steps for obtaining statistically consistent MTS.

BC is preferable to CC since the parameter set depends on the data utilized and on their distributional properties. BC can be carried out by multi-draw Gibbs sampling if the data supplied are not sufficiently informative and the underlying model is nonlinear.

The departure point of this procedure requires an initialization parameter like the OLS-estimated set $\hat{B}$ obtained from a "training period", a flat prior or simply a scalar or vector of zeros. Normal distribution of the parameter set and Inverse-Gamma distribution of the variance are assumed, while conditional posteriors of both are made to depend on the data and on each other's priors. This avoids the intricacies of joint posterior distribution typical of Monte Carlo integration and allows simulating the conditional posteriors to produce *J* random draws of the parameter set and construct the MCMC (Koop and Korobilis, 2009).

The process is replicated at each draw to finally obtain the averaged-out desired parameter set denoted as $\overline{B}_t$. Alternatively, a TVP sequence may be obtained by means of the procedure described in Appendix B. Suffice here to mention that this procedure, due to its characteristic of sequential forward and backward estimation in a SS context, is defined as a two-way optimal KF specifically designed for Bayesian estimation (Carter and Kohn, 1994; Koop and Korobilis, 2009).

The BC method requires that the parameter MCMC meet Hadamard's three criteria of "well posedness": (i) for all admissible data, a solution exists, (ii) for all admissible data, the solution is unique, (iii) the time path of the solution is stationary, where "solution" refers in the present context to the calibration process and its underlying parameters, and "stationarity" implies no overtime explosive behavior of both parameters and their variances.

This occurrence requires meeting specific targets, as shown in Appendix B, in order to minimize the curse of uncertainty that plagues many climate reconstructions (*e.g.* Mann *et al.*, 1999). As with CC, the BC fitted time series is $\hat{Y}_t = Y_t\overline{B}$ and $\hat{Y}_t = Y_t\overline{B}_T$, respectively for the TFP and for the TVP case. In the present context, to save space and gain in efficiency, only the latter case is applied.

Finally, the BC method compares with some calibration procedures that utilize Bayesian estimation for spatiotemporal climate reconstructions (Tingley, 2010a, 2010b). In spite of data constraints given by the absence of large space and time coordinates that stem from gridded information about instrumental and target datasets, the BC is the optimal calibrating procedure because it fully exploits the advantages of Gibbs sampling and of the two-way TVP KF advanced in the Introduction and formalized in Appendix B.

It is quite obvious that additional information to the available datasets in terms of time and field records, model assumptions and priors are likely to convey richer results, as maintained by some authors (*e.g.* Tingley and Li, 2012). However, the BC method may still be defined as a direct competitor of Tingley's own Bayesian data reconstruction procedure, which is loosely defined by the author as being capable of providing results that admittedly are in some sense equivalent to the Kalman smoother.

Finally, a valuable and hotly-debated piece of paleodata reconstructions by McShane and Wyner (2011) falls short of the informational set provided by the conditional probability distribution of observables and states that is contained in the BC. By consequence, their inconclusive results regarding the hockey-stick hypothesis, obtained by an unspecified brand of CC, are still fraught with uncertainties (Rougier, 2011; Tingley 2011).

4. Data transformations in Bayesian Calibration

The MTS produced is shaped as shown in (8). However, different from CC, data transformations in a TVP setting are required to avoid the occurrence of HSB that leads to statistical inconsistencies and to the curse of "spurious regression" (Granger and Newbold, 1974).These data transformations are implemented through several steps:

1) Test for normality the series commencing in the year 1850 in (2) and (6) by means of the Jarque-Bera, the Shapiro-Wilk and/or of the D'Agostino tests (D'Agostino *et al.*, 1990; Royston, 1995). For each time series, if nonnormality cannot be rejected, proceed to ordinary normalization to ensure outlier-free estimation. The transformed and homogenized series, to avoid notational clutter, are still defined as Y_τ and y_τ, respectively. They exhibit values within the range $[0,1]$ and retain the stationarity or nonstationarity properties of the original series;

2) If necessary, proceed to first-order differencing of the above series, now denoted as ΔY_τ and Δy_τ, respectively, to perform nonspurious and robust OLS estimation, and to subsequently ensure during Gibbs sampling that the parameter MCMC meet Hadamard's three criteria;

3) Regress ΔY_τ over Δy_τ to obtain the direct OLS estimator $\hat{B}$ of (4) or alternatively regress Δy_τ over ΔY_τ to obtain the indirect OLS estimator, a practice not free from inference problems (Tellinghuisen, 2000);

4) Find by optimal Gibbs sampling the MCMC sequence of the parameter set $\hat{B}_t$ in order to produce the fitted series $\Delta\hat{Y}_t = \Delta Y_t \overline{B}_t$ prior to the year 1850.

5) Standardize both series $\Delta\hat{Y}_t$ and Δy_τ. For each time series, proceed to ordinary centering and scaling or else to centering by the median and scaling by the standard deviation obtained by the inter quartile 25%-
75% range. Both transformed series, denoted as $\Delta\hat{Y}_t^*$ and Δy_τ^*, are $NID(0,1)$ and are expected to produce the series

$$\Delta\hat{Y}_s = \left\{\Delta\hat{Y}_t^*,\ \Delta y_\tau^*\right\}' \qquad (10)$$

with no breaks in the immediate neighbourhood of time *T*, *i.e.*at the junction of their observed values. Tests for the equality of means and variances of both series are available for this occurrence. The first test may be conducted by standard paired t-tests or by the Mann-Whitney U-test for medians (Hollander and Wolfe, 1999). The second test employs textbook one-way Analysis of Variance, the Kruskal-Wallis test (Kruskal and Wallis, 1992), or the nonparametric Ansari-Bradley's test of dispersion (Ansari and Bradley, 1960), designed to incorporate outliers that may cause a nonrejection significance level beneath 90-95%.

6) If necessary, integrate overtime (10) and finally form the MTS of (8) (Banerjee *et al.*, 2005). For a smoother output, if desired, apply to (8) the Savitzky-Golay's filter (Savitzky and Golay, 1964; Orphanides, 2010) and attach to the output series the appropriate confidence intervals (*e.g.*, Loehle, 2007; Loehle and McCulloch, 2008).

7) Test the obtained MTS for structural change at an unknown date by means of the Kim-Perron procedure (Perron and Zhu, 2005; Kim and Perron, 2009). This method is based on the minimization, over all possible break dates, of the squared residual sum of a dynamic regression of the endogenous variable in (8). Thereafter, rank the highest peaks of each series, separated by a

minimum distance of 150 years. This technique is mandatory in the stochastic environment generated by Gibbs sampling to define the spread of the highest temperature dates in long-term series.

8) Repeat the steps 4–6 for *J* times in order to ensure asymptotic validity of the results as shown in Appendix B.

It should be made clear that step 1 is essential to the procedure as it enables the researcher to deal with series measured in different units, thereby producing a single homogenous MTS. Therefore, paleodata series (Appendix D) not expressed in temperature terms, like some of BÜNTGEN, PEDERSON, ME_STAHLE, GAGEN and STAMBAUGH, as well as reconstructed past temperatures (*e.g.* series 4 in G_CUBED and GUIOT) are normalized and then standardized to the BEA temperature data and thus yield Maximum-Likelihood estimators in OLS regressions and enable unfettered Gibbs sampling.

In addition, the structural change test for a break in the MTS at an unknown date that features in step 7 entails by construction the detection of the date corresponding to the minimum squared residual sum. No inferences about nonlinearity properties of the series are involved, such as fractional integration, asymmetric cyclicality and mean reversion (Fan and Yao, 2003). By consequence, in the present context, the shape of the series is irrelevant to break date detection, which merely stands as a piece of evidence of a regime change in the variability of the residuals.

In order to better describe the workings of the above procedure, four renowned time series of the G_CUBED dataset (see Appendix D) were selected. These are two series by Mann and collaborators (Mann *et al.*, 1999, 2009), the Crowley series (Crowley, 2000), and the series by D'Arrigo and collaborators (D'Arrigo *et al.*, 2006). The timelines of the original series are then visually exhibited in Fig. 1 together with their calibrated MTS counterparts. For improved comparison purposes, the coupled series are appropriately normalized and rescaled as in step 1, and also smoothed (Savitzky and Golay, 1964; Orphanides, 2010).The four original transformed series undisputedly manifest HSB by exhibiting peak dates in the late 90's, while their corresponding calibrated MTS do not exhibit HSB, and their peak dates are all distributed within the MWP. Precisely, they correspond to the years 968, 1249, 1087, and 895.

5. The Bea, the CCP datasets and the MTS results

We retain the notation of Sect. II and in particular that of (1)–(6). Hence, let y_τ, for $\tau \in [\tau_1, \mathrm{T}]$ be the BEA series where $\tau_1 = 1850$, $\mathrm{T}=2010$, and let also N = 258 be the total number of CCP series pertaining to the 19 datasets shown in Appendix D in sequential order. Each of the CCP series is characterized by a different length, that is, by different beginning and ending dates. Each of these overlaps the timeline of the BEA series and is necessary for performing OLS estimation of the parameter set $\hat{B}$. More precisely, after defining each CCP series as $Y_{s,i}$, for $s \in [t_i, \tau_i]$, $\forall i \in N$ where t_i, τ_i respectively are the *i*.th series-specific commencing and ending dates, the length of each series is $(T_i + \mathrm{T}_i)$ which is the sum of the number of observation prior to and after the year 1850. The overlapping length with the BEA series is thus $\tau \in [\tau_1, \mathrm{T}_i]$, the time stretch suitable for performing OLS estimation.

For illustrative purposes, Table II provides some descriptive statistics regarding the BEA series for the following geographical areas: Global (GL), Northern Hemisphere (NH), and the Southern Hemisphere (SH). Basic descriptive statistics and the p-values of the ADF and KPSS tests for stationarity with optimal lag selection based on the Bayesian Information Criterion (BIC) and of the Jarque-Bera statistics for normality are reported. The volatilities of the three series are almost all identical, and they share also nonrejection of the unit-root null hypothesis, in particular the KPSS test, and also a significant rejection level of normality. Also the dates of maximum achieved temperatures are almost identical, while their corresponding minima are barely similar probably because of some recording heterogeneity.

As far as the 258 CCP series are concerned, space limitations prevent us from fully reporting the same statistics as those of Table II, and even more so to produce graphs of their performances apart from those exhibited in Fig. 1. Due to space limitations, only averaged aggregate results for each dataset are exhibited in Table III.

Not unexpectedly, the series are characterized by great diversity even when pertaining to the same dataset. Out of the total number of series, the percent of stationarities computed by the BIC-corrected ADF test at a 5% level amounts to 32% (col. 4) and the percent of normalities computed by the JB test at the same level amounts to 20% (col. 5). The KPSS test results are not reported here, but they point to an even lower percent stationarities indicating a large presence of significantly trended series.

Hence, from the test results at the given level, there is no significant evidence that the majority of the CCP series are stationary and normal. However, these features do not impact the results exhibited in Tables IV and V once the methodology expounded in Sect. IV is applied with due diligence to construct the MTS.

In addition, Table III reports the average correlation coefficient of the CCP series with BEA after the year 1850 (col. 6), namely, the statistical relationship between the variables y_τ and Y_τ utilized for the OLS estimation of the parameter set $\hat{B}$. Only four series datasets, ENSO_LI, FS_LINDHOLM, ME_STAHLE and STAMBAUGH, and some series contained in LJUNGQVIST, exhibit insignificant correlation coefficients by common standards. On the other hand, some other series contained in LJUNGQVIST exhibit absolute correlation coefficients that exceed 0.70, while for the first three series of the G_CUBED dataset that were discussed in Sect.3 the correlation coefficient largely exceeds 0.80.

The reader is however warned of the relative unimportance of low correlation in a context of BC contrary to the workings of CC. In fact the OLS-estimated parameter set $\hat{B}$ is merely a prior that is utilized for initialization of the MCMC estimation process in optimal Gibbs sampling, as shown in Sect. III. Further on, the series are individually calibrated according to the criteria expounded in Sections II and III. Their most relevant results, expressed as dataset averages including variance tests, parameter values and both break and peak dates are exhibited in Tables IV and V.

In Table IV, the average results of TVP parameter estimation are produced for each dataset. The first column shows the percent value of factor f in (9). Values of the factor close to zero (*e.g.*, ENSO_YAN and TROUET) are a potential source of HSB. However, the score sums of the three equal-variance tests considered in step 5 of Sect. IV point irrefutably to nonrejection of the implied null hypothesis (col. 2). Thereby, the MTS produced according to the construct indicated in the steps of Sect. IV is mostly unlikely to produce HSB for any of the 258 CCP series, as also shown on a smaller scale in Fig. 1.

Table IV also reports the average values of the TVP parameter set (B.11) and the corresponding t-test computed statistics. The parameter set includes a slope and a constant-term coefficient (cols. 3 and 5). The majority of the slope coefficients are significantly different from zero as shown by the absolute values of their statistics (col. 4). More specifically, only the two series of STAMBAUGH, three series of the GUIOT and one series each of the BÜNTGEN and CHRISTIANSEN datasets fall beneath the customary two-tailed 5% critical value. The average t-test statistics of the constant term, instead, fare sizably worse as seven of them fall beneath the two-tailed 10% critical value (col. 6). However, the second occurrence is irrelevant for the purpose of detecting HSB in the calibrated MTS, and may thus be comfortably disregarded. GUIOT, LJUNGQVIST and a few more, meets this requirement.

Table V reports some useful results of the temperature single break dates and the ranked peak dates of the calibrated MTS. The dataset median break dates, where negatives correspond to years B.C. (col. 1), are obtained via the Kim-Perron procedure while the peak dates are obtained via the last step of the procedure shown in Sect. III and are ranked from the first to the second highest (cols. 3-6). None of the break dates is located in the 20th century or after, while the majority are concentrated, as from the data reported (col. 2), around the year 1076 A.D. ± 424 which stands well far away from the RWP. In addition, Table V reports the average ratio between the first and the second peak dates (col. 7). For the entire MTS dataset, this ratio amounts to 1.91. The ratio of the first to the third peak date value is roughly three, large enough for requiring no further consideration.

In sum, by taking into account the vast differences existing between and often within the datasets, and also the intrinsic stochastics of the peak-finding methodology, the first two ranked peaks along with their standard deviations are necessary and sufficient to produce reasonable inferences. Their mean dates respectively are the years 970 A.D. ± 407 and the years 854 A.D. ± 398, highly insufficient to enter even perchance the 20th century. Because of the large geographical scope covered by the 19 datasets analyzed, we may conclude that the MWP was a global phenomenon. Even within each dataset, the first two peak dates exceeding the year 1900 (col. 8), regarded as critical by the supporters of the anthropogenic climate change hypothesis, are very few or null in most cases. In fact, just a limited bunch of these, exactly 16, located in the datasets CHRISTIANSEN,

6. Conclusions

Bayesian Calibration applied to the reported 258 CCP series, and based on the HADCRUT4 BEA instrumental temperature records, has produced interesting results from both the theoretical and applied viewpoints. After discarding the relevance of Classical Calibration because highly likely to produce HSB, BC is shown to be highly flexible and reliable especially because of its treatment of conditional posteriors that facilitates the production of a TVP set in a long-term time series context. Its implementation however requires data transformations necessary to produce consistency between the calibrated and the instrumental series.

Mean- and especially variance-equality tests are recommended in order to correctly proceed to the Gibbs sampling routine. From the applied viewpoint, the values and the significance of the TVP estimation are satisfactory in most cases, while temperature breaks and peaks suggest widespread rejection of the hockey-stick hypothesis. In fact, single break points in no case detect structural change at or around RWP dates, while less than 10% of the highest peak dates of the CCP series enter the 20th century. Rather, temperature breaks and peaks are centered within the Middle Ages so that, given the large geographical scope covered by the available data, we may conclude that the MWP was a global phenomenon significantly warmer than the RWP, as demonstrated also by the large amount of referenced authors.

Appendix A. Classical Calibration and Hockey-Stick Behavior

In its simplest form, backward CC reconstruction for producing (8) involves two simple steps:

1) direct OLS estimation of the equation

$$y_\tau = Y_\tau B + e_\tau \qquad \text{(A.1)}$$

linking the synchronous observables y_τ (the BEA temperature readings) and the regressor Y_τ where $e_\tau \sim IID\left(0, \sigma_e\right)$, $\infty >> \sigma_e > 0$, and for consistency of the estimator $\mathrm{E}\left(Y_\tau ' e_\tau\right) = 0$. Obviously, spurious correlation and biased parameter estimation should be accounted for by stationarizing the available time series if necessary (Granger and Newbold, 1974);

2) fitting the unobservable $\hat{Y}_t$ of (3) from the parameter set $\hat{B}$ and the observable Y_t of (7) such that

$$\hat{Y}_t = Y_t \hat{B} \qquad \text{(A.2)}$$

where $\hat{Y}_t$ is the reconstructed or virtual series and

$$\hat{B} = \left(Y_\tau ' Y_\tau\right)^{-1} Y_\tau ' y_\tau \qquad \text{(A.3)}$$

is the calibration estimator of (4).

Combining in sequence the reconstructed and the actual temperature series produces the dataset of contiguous series shown in (8), here replicated

$$\hat{Y}_s = \left\{\hat{Y}_t,\ y_\tau\right\}' \qquad \text{(A.4)}$$

such that the HSB hypothesis may be tested for a stochastic change in the trend of the series $\hat{Y}_s$ at an unknown date $T^* \approx T$ possibly in the close neighborhood of the ending date of $\hat{Y}_t$ and the beginning date of y_τ.

A further test of the HSB hypothesis may be obtained by comparing the variances of the two contiguous time series in (A.4) as evidenced in step 5, Sect. IV. Substitution of eq (A.3) into (A.2) produces

$$\hat{Y}_t = Y_t \left(Y_\tau ' Y_\tau\right)^{-1} Y_\tau ' y_\tau \qquad \text{(A.5)}$$

which is the product of three variable components, for convenience rewritten in stacked form

$$\hat{Y}_t = \prod_{i=1}^{3} x_i \qquad \text{(A.6)}$$

where $x_1 = Y_t$, $x_2 = \left(Y_\tau ' Y_\tau\right)^{-1} Y_\tau$, $x_3 = y_\tau$, such that

$$\hat{Y}_t = Y_t \hat{B} \qquad \text{(A.2)}$$

where $\hat{Y}_t$ is the reconstructed or virtual series and

$$\hat{B} = \left(Y_\tau ' Y_\tau\right)^{-1} Y_\tau ' y_\tau \qquad \text{(A.3)}$$

is the calibration estimator of (4).

Combining in sequence the reconstructed and the actual temperature series produces the dataset of contiguous series shown in (8), here replicated

$$\hat{Y}_s = \left\{\hat{Y}_t,\ y_\tau\right\}' \qquad \text{(A.4)}$$

such that the HSB hypothesis may be tested for a stochastic change in the trend of the series $\hat{Y}_s$ at an unknown date $T^* \approx T$ possibly in the close neighborhood of the ending date of $\hat{Y}_t$ and the beginning date of y_τ.

A further test of the HSB hypothesis may be obtained by comparing the variances of the two contiguous time series in (A.4) as evidenced in step 5, Sect. IV. Substitution of eq (A.3) into (A.2) produces

$$\hat{Y}_t = Y_t \left(Y_\tau ' Y_\tau\right)^{-1} Y_\tau ' y_\tau \qquad \text{(A.5)}$$

which is the product of three variable components, for convenience rewritten in stacked form

$$\hat{Y}_t = \prod_{i=1}^{3} x_i \qquad \text{(A.6)}$$

where $x_1 = Y_t,\ x_2 = \left(Y_\tau ' Y_\tau\right)^{-1} Y_\tau,\ x_3 = y_\tau$, such that

$$Var\left(\hat{Y}_t\right) = Cov\left(x_i^2\right) + \prod_{i=1}^{3}\sum_{i=1}^{3} \bar{x}_i^2 \cdot Var\left(x_i\right) - \left(Cov\left(x_i\right) + \prod_{i=1}^{3} \bar{x}_i\right)^2 \qquad \text{(A.7)}$$

where $\bar{x}_i$ is the mean of x_i (Goodman, 1960, 1962). It is of particular interest computing from (A.7) the ratio that describes the impact of the variance of the last component of $\hat{Y}_t$ over its variance

$$\frac{\partial Var\left(\hat{Y}_t\right)}{\partial Var\left(x_3\right)} = \prod_{i=1}^{2}\sum_{i=1}^{2} \bar{x}_i^2 \cdot Var\left(x_i\right) \qquad \text{(A.8)}$$

which leads to

$$f = \frac{\partial \sigma\left(\hat{Y}_t\right)}{\partial \sigma\left(y_\tau\right)} \qquad \text{(A.9)}$$

where $\sigma(.)$ expresses the standard deviation of the bracketed variable. From (A.9) we can write

$$\sigma\left(\hat{Y}_t\right) = f \cdot \sigma\left(y_\tau\right) \qquad \text{(A.10)}$$

which is (8) of Sect. II here replicated, wherein the factor f is such that, by the given construct $f \propto \left|\hat{B}\right|$. In fact, $\infty > f > 0$ and obviously $\mathrm{E}(f) = 1$ if all the component series are random White Gaussian Noise (WGN). However, if the variable y_τ is not WGN, and specifically is Random Walk (RW) or RW with drift (RWD), then

$$\underset{\mathrm{T}^i \to \infty}{Lim}(f) = 0,\ i = 1, 2 \text{ if } y_\tau = \mathrm{RW}, \mathrm{RWD} \qquad \text{(A.11)}$$

which means that $f = \mathrm{O}_p\left(T^{-i}\right)$ *i.e.*, it converges to zero with order in probability of $1/\mathrm{T}$ and $1/\mathrm{T}^2$ if the variable y_τ is RW or RWD, respectively. The outcome derives, for $\hat{Y}_t$ a WGN variable where $\sum_{t=1}^{T} \hat{Y}_t^2 = \mathrm{O}_p(T)$, from $\sum_{\tau=1}^{\mathrm{T}} y_\tau^2 = \mathrm{O}_p\left(\mathrm{T}^2\right)$ if y_τ is RW, and from $\sum_{\tau=1}^{\mathrm{T}} y_\tau^2 = \mathrm{O}_p\left(\mathrm{T}^3\right)$ if y_τ is RWD (Hamilton, 1994, Ch. 17).

For $\hat{Y}_t$ to be WGN we require from (A.2) that either Y_t and Y_τ be WGN and/or $|B| \approx 0$. If Y_t and Y_τ are not WGN, then $\underset{\hat{B} \to 0}{Lim}\ \sigma\left(\hat{Y}_t\right) = 0$ which implies that from (A.3) the following holds for $\delta > 0$

$$\underset{\left|\hat{B}\right| \to 0}{Lim}\ \Pr\left[\left(\sigma\left(y_\tau\right) - \sigma\left(\hat{Y}_t\right)\right) \geq \delta\right] = 1 \qquad \text{(A.12)}$$

a corollary of which is the HSB in (A.4), namely

$$\underset{\left|\hat{B}\right| \to 0}{Lim}\ \Pr\left[\left(\min\left(y_\tau\right) - \max\left(\hat{Y}_t\right) \geq \delta\right)\right] = 1 \qquad \text{(A.13)}$$

which means that on the limit, for the necessary condition $\left|\hat{B}\right| \to 0$, the probability that the minimum observation of y_τ is at least equal to the maximum achieved observation of $\hat{Y}_t$ approaches unity.

Appendix B. Gibbs sampling for State-Space models

The Carter and Kohn procedure (Carter and Kohn, 1994) is a TVP Gibbs sampler where the coefficients are modeled as state variables following a RW and the observables are linearly tied to them. The procedure yields MCMC sampling and is designed for Bayesian estimation of select coefficient vectors or scalars. It is indeed an optimal Gibbs sampler because it exploits all the properties of the two-way KF. In fact it is implemented over the select *J* number of draws through four sequential steps:

1) setting up the underlying SS model and its prior parameters;

2) forward filtering through the sample period $1,\ldots,T$ and recovery of the estimated parameters;

3) appropriate probability conditioning and parameter sampling;

4) optimal adaptive backward smoothing of the estimated parameters by means of the Rauch-Tung-Striebel algorithm, a standard toolkit in the two-way KF procedure (Koop and Korobilis, 2009).

Let $n \geq 2$ and $p \geq 1$, where *n* is defined in (1) and *p* is the number of elements in the measurable dataset such that $Y_t : (T \times p)$. The first step of Gibbs sampling departs from the following SS model originally developed by Kalman (1960)

$$B_t = B_{t-1} + v_t \tag{B.1}$$

$$Y_t = B_t C_t^{'} + \mu_t \tag{B.2}$$

where the first and second equation respectively represent the dynamics of the state and of the measurement variables. In particular, $\forall t \in T$ we have:

$$\begin{pmatrix} v_t \\ \mu_t \end{pmatrix} \sim NID\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Q_t & 0 \\ 0 & R_t \end{pmatrix} \right) \tag{B.3}$$

and $B_t : (n \times p)$, $Q_t : (n \times n)$, Y_t, $e_t : (1 \times p)$, $C_t : (n \times p)$, and $R_t : (p \times p)$. In a univariate context, $p = 1$, such that some parameters are vectors or scalars. The matrix Q_t is diagonal such that the *n* estimated elements of B_t are mutually orthogonal.

The nonlinear SS model so conceived accommodates many kinds of estimation models of Y_t, including Vector Auto-Regressions (Hamilton, 1994, Ch. 11) where $p > 1$, and requires prior initialization of the B_t, Q_t and R_t parameters if under a regime of flat priors. If priors are informative, this kind of initialization is replaced by utilizing the OLS results over a training period of preselect length. In both cases, the pre-sample parameters are denoted respectively as B_0, Q_0 and R_0.

Define $P_t : (n \times n)$ as the Riccati matrix expressing the state covariance matrix of B_t while Q_t and R_t from (B.3) are defined as the covariance matrix of the state and of the measurement error, respectively. Finally, C_t is a Kronecker Product (KP) of the measurement variable for each *t* observation and preselected lags $k \geq 1$

$$C_t = I_p \otimes \left(Y_{t-k}, I_p \right) \tag{B.4}$$

where I_p is the identity matrix of size *p*. Inclusion of the latter into the bracketed expression of (B.4) adds constant terms to the estimation process. As an example to better describe the construction of C_t, by assuming $k = 1$ for simplicity, as $t = 1$ the bracketed expression includes the first row of observations of Y_t. Likewise as $t = 2$, the second row of Y_t is included in the KP equation and so on until *T* is reached.

The Kalman gain is a $p \times p$ matrix defined as

$$H_t = \left(C_t P_t C_t^{'} + R_t\right)^{-1} \qquad \text{(B.5)}$$

and the dynamic Riccati matrix is

$$P_t = P_{t-1} + Q_t - P_t^{'} C_t^{'} H_t C_t P_t \qquad \text{(B.6)}$$

where the sequence of the vectorized diagonal elements of Q_t, is sampled once from a multivariate normal distribution because $Q_t \sim NID\left(0,1\right)$. In addition, the dynamic evolution of the state variable is expressed as

$$B_t = B_{t-1} + \left(P_t C_t H_t \mu_t^{'}\right)^{'} \qquad \text{(B.7)}$$

and finally the measurement errors and their covariance matrix are

$$\mu_t = Y_t - B_t C_t^{'},\ \mathrm{R}_t = \mu_t^{'} \mu_t \qquad \text{(B.8)}$$

The second step of Gibbs sampling is accomplished by estimating in sequence (B.3)–(B.8) after providing the necessary initializations. Estimation is performed via forward KF and during the process (B.5) to (B.7) are stored. The third step consists of retrieving the time series of the estimated parameters of length $T^* = T-1$ and to produce by single sample the following conditional posteriors:

$$\begin{aligned} \hat{P}_t &= P_t \,\middle|\, B_{t-1}, \\ \hat{B}_t &= B_t \,\middle|\, \hat{P}_t, \\ \hat{v}_t &= \hat{B}_t - B_{t-1}, \\ \hat{Q}_t &= Q_t \,\middle|\, \hat{v}_t. \end{aligned} \qquad \text{(B.9)}$$

In accordance with Gibbs sampling, the conditional posterior distributions are

$$\begin{aligned} \hat{P}_t &\sim IW\left(Var\left(\hat{B}_t\right),\ T\right), \\ B_t &\sim \left(\hat{B},\ \hat{P}_t\right), \\ \hat{v}_t &\sim NID\left(0,\ \hat{Q}_t\right), \\ \hat{Q}_t &\sim IW\left(Var\left(\hat{v}_t\right),\ T\right) \end{aligned} \qquad \text{(B.10)}$$

where *IW* refers to the Inverse Wishart distribution, and

$$Var\left(\hat{B}_t\right) = \hat{B}_t^{'} \hat{B}_t \qquad \text{(B.11)}$$

which, together with $\hat{P}_t$ and $\hat{Q}_t$ is of size $\left(n \times n\right)$. The state and the observable covariances are instead Inverse-Gamma distributed.

The fourth and final step of the TVP Gibbs sampling procedure is represented by adaptive backward KF estimation of the parameters of (B.9). In practice, the process entails running in reverse time, from $T^* = T-1$, down to the first observation, the following two equations

$$\begin{aligned} \hat{P}_{T^*-t+1} &= P_{T^*-t+1} \,\middle|\, B_{T^*-t}, \\ \hat{B}_{T^*-t+1} &= B_{T^*-t+1} \,\middle|\, \hat{P}_{T^*-t+1} \end{aligned} \qquad \text{(B.12)}$$

which are appropriately modified versions of (B.6)–(B.7), respectively. The end result of interest is $\hat{B}_T : \left(T^* \times np\right)$, where the parameter vector $\hat{B}_t$ of (B.9) is sampled once under the assumption posited in (B.10). Hadamard's three criteria for "well posedness" are then tested at this stage by tracking the evolution of the norms of the error and covariance parameters of (B.9). These should converge toward zero or at least manifest no nonstationarity within their own timespan.

Once a stationary solution is found, the above process is replicated *J* times in Gibbs sampling after averaging out the parameter set of (B.12), denoted as B^*. This in turn at each *j*th draw is conditioned on its own Riccati covariance matrix similarly denoted as P^* such that, eventually, we have the required Gibbs-sampled MCMC parameter series denoted as $B_J^*:(J,n)=B_j^*|P_j^*$. Averaging out this new outcome over all draws eventually produces the desired BC-estimated time series which is $\hat{Y}_t = Y_t\bar{B}$, where the estimator

$$\bar{B} = J^{-1}\sum_{j=1}^{J} B_j^* \qquad \text{(B.13)}$$

is of size $(n \times p)$ and $Y_t,\ \hat{Y}_t:(T \times n)$. Asymptotically, for $\tilde{B}$ the true population parameter vector, the following applies

$$\underset{J\to\infty}{Lim}\ \Pr\left(\bar{B}=\tilde{B}\right)=1 \qquad \text{(B.14)}$$

as expected from optimal Gibbs sampling.

Appendix C. Table of acronyms used in the text

ADF: Augmented Dickey-Fuller test

BC: Bayesian Calibration

BEA: Best Estimated Anomaly

BIC: Bayesian Information Criterion

CC: Classical Calibration

CCP: Climate Change Proxy

DGP: Data Generation Process

GL: Global

HSB: Hockey-Stick Behavior

IPCC: Intergovernmental Panel on Climate Change

JB: Jarque-Bera test

KF: Kalman Filter

KP: Kronecker Product

KPSS: Kwiatkowski-Phillips-Schmidt-Shin test

MCMC: Monte Carlo Markov Chain

MTS: Millennial-scale Time Series

MWP: Medieval Warming Period

NH: Northern Hemisphere

OLS: Ordinary Least Squares

RW: Random Walk

RWD: Random Walk with Drift

RWP: Recent Warming Period

SH: Southern Hemisphere

SS: State-Space

TFP: Time Fixed Parameter

TVP: Time Variable Parameter

WGN: White Gaussian Noise

Appendix D. Table of acronyms and sources of the 19 datasets

1. BÜNTGEN, Büntgen U. *et al.*, 2011: Central Europe 2,500 year tree ring summer climate reconstructions, includes two series: Reconstructed April-May-June precipitation (398 B.C.-2008 A.D.), and Reconstructed June-July-August temperature anomaly (499 B.C.-2003 A.D.); Büntgen U. *et al.*, 2010, Reconstructed precipitation (996–2005 A.D.); Büntgen U. *et al.*, 2006, Reconstructed temperature (755–2004 A.D.).

2. CHRISTIANSEN, Christiansen B. and Ljungqvist F.C., 2011: Northern Hemisphere extratropical 1,000 year temperature reconstruction (1000–2000 A.D.).

3. COOK, Cook E.R. *et al.*, 2000: Tasmania temperature reconstruction (1600 B.C.-1991 A.D.).

4. CUVEN, Cuven S. *et al.*, 2011: East Lake, Melville Island, Canada 4,200 year varve thickness data (2192 B.C.-2005 A.D.).

5. ENSO_LI, Li J. *et al.*, 2011: 1,100 Year El Niño/Southern Oscillation (ENSO) index reconstruction (900–2002 A.D.).

6. ENSO_YAN, Yan H. *et al.*, 2011: 2,000 year precipitation-based Southern Oscillation index reconstruction (50–1955 A.D.).

7. FS_GAGEN, Gagen M. *et al.*, 2007: Fennoscandia 1,100 year summer (July-August) sunshine reconstruction (886–2001 A.D.).

8. FS_LINDHOLM, Lindholm M.R. *et al.*, 2010: Fennoscandia 1,250 year height increment summer temperature reconstruction (745–2007 A.D.).

9. GUIOT, Guiot J. *et al.*, 2010: Latitude-longitude point gridded panel (1408 x 125) (600–2007 A.D.) made up of different proxies (tree-ring width, historical and ice-core data).

10. LJUNGQVIST, Ljungqvist F.C., 2009: Proxy series and assembled from different published sources and with at least centennial sample resolution covering the last two millennia, most of which commencing the year 0 A.D. and quite a few ending in the late 90s.

11. LOSO, Loso M.G., 2008: Iceberg Lake, Alaska varve thickness data and temperature reconstruction (442–1998 A.D.).

12. ME_STAHLE, Stahle, D.W. *et al.*, 2011: Mesoamerican 1,238 year June PDSI reconstructions.

13. NEUKOM, Neukom *et al.*, 2010: Southern South America multiproxy 1,100 year temperature reconstructions (900–1995 A.D.).

14. PEDERSON, Pederson *et al.*, 2011. North American Cordillera (Northern Rockies and Greater Yellowstone Region) 1,600 year snowpack reconstruction (1252–2007 A.D.), North American Cordillera (Upper Colorado, and Southern Cordillera basins) reconstructed HUC6 watershed April 1 SWE records (369–2005 A.D.).

15. SINHA, Sinha A. *et al.*, 2011: Central and Northeast India 1,000 year stalagmite oxygen isotope data, composite $\delta 180$ (625–2007 A.D.).

16. STAMBAUGH, Stambaugh, M.C. *et al.*, 2011: Central United States 1,000 year summer PHDI reconstructions (992–2004 A.D.).

17. TROUET, Trouet V. *et al.*, 2009: Multi-decadal winter North Atlantic Oscillation reconstruction (1049–1995 A.D.).

18. WILSON, Wilson R. *et al.*, 2006: Gulf of Alaska 1,300 year temperature reconstruction (724–1999 A.D.).

19. G_CUBED, Wahl E.R. *et al.,* 2010: NOAA 92 PCN temperature reconstructions, amongst which ten selected 1K+ uninterrupted series including: (1) Hantemirov and Shiyatov, Yamal peninsula multimillennial summer temperature reconstruction, 2002, (2066 B.C.-1996 A.D.); (2) Salzer and Kipfmueller, 2005, Southern Colorado Plateau temperature and precipitation reconstructions (251 B.C.-1996 A.D.); Tand *et al.*, 2003, Shihua Cave, Beijing stalagmite temperature reconstruction (665 B.C.-1985 A.D.); (4) Mann *et al.*, 2008, 2,000 year hemispheric and global surface temperature reconstructions, global: land and ocean: error-in-variables method (500–2006 A.D.); (5) Mann *et al.*, 1999, Northern Hemisphere temperatures during the past millennium (1000–1980 A.D.); (6) Crowley, 2000, Northern Hemisphere temperature reconstruction (1000–1993 A.D.); (7) Esper *et al.*,

2002, Northern Hemisphere extratropical temperature reconstruction (831–1992 A.D.); (8) D'Arrigo *et al.*, 2006, Northern Hemisphere tree-ring-based temperature reconstruction: regional curve standardization (713–1995 A.D.); (9) Moberg *et al.*, 2005, 2,000-year Northern Hemisphere temperature reconstruction (1–1979 A.D.); (10) Moore *et al.*, 2001, Baffin Island 1,250 year summer temperature reconstruction (752–992 A.D.).

Acknowledgments

The author wishes to thank R. Fiorito, N.-A. Mörner, and three anonymous referees for their very helpful comments. Any errors remain solely within his personal responsibility.

References

Abdussamatov, H.: About the long-term coordinated variations of the activity, radius, total irradiance of the sun and the Earth's climate, In Proceedings of the International Astronomical Union, 223, 541-542, 2004.

Alanko-Huotari, K., Usoskin, I.G., Mursula, K., and Kovaltsov, G.A.: Global heliospheric parameters and cosmic ray modulation: an empirical relation for the last decades, Solar Phys., 238, 391–404, 2006.

Ansari, A.R. and Bradley, R.A.: Rank-sum tests for dispersion, Ann. Math. Stat., 31, 1174–1189, 1960.

Appenzeller, C., Stocker, T.F., and Anktin, M.: North Atlantic oscillation dynamics recorded in Greenland ice cores, Science, 282, 446–449, 1998.

Baliunas, S., Soon, W.: Lessons and limits of climate history: was the 20th. century climate unusual? The George C. Marshall Institute: Washington, D.C., 1-25, 2003.

Banerjee, A., Marcellino, M., and Masten, I.: Leading indicators for Euro-area inflation and GDP growth, Oxf. Bull. Econ. Stat., 67, 785–813, 2005.

Bard, E. and Frank, M.: Climate change and solar variability: What's new under the sun?, Earth Planet. Sci. Lett., 248, 1–14, 2006.

Bradley, R.S.: Paleoclimatology: Reconstructing climates of the Quaternary, Harcourt Academic Press: San Diego, CA, pp. 613, 1999.

Bradley, R.S., Briffa, K.R., Crowley, T.J., Hughes, M.K., Jones, P.D., and Mann, M.E.: Scope of medieval warming, Science, 292, 2011–2012, 2001.

Broecker, W.S.: Was the Medieval Warm Period global?, Science, 291, 1497–1499, 2001.

Büntgen, U., Tegel, W., Nicolussi, K., McCormick, M., Frank, D., Trouet, V., Kaplan, J.O., Herzig, F., Heussner, K.U., and Wanner, H.: 2500 Years of European climate variability and human susceptibility, Science, 331, 578–583, 2011.

Büntgen, U., Trouet, V., Frank, D.C., Leuschner, H.H., Friederichs, D., Luterbacher, J., and Esper, J.: Tree-ring indicators of German summer drought over the last millennium, Quat. Sci. Rev., 29, 1005–1016, 2010.

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

Bürger, G.: Comment on the spatial extent of 20th-century warmth in the context of the past 1200 years, Science, 316/5833, 1844, 2007.

Carter, C.K. and Cohn, P.: On Gibbs sampling for state space models, Biometrika, 81, 541–553, 1994.

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

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

Christiansen, B. and Ljungqvist, F.C.: Reconstruction of the extratropical NH mean temperature over the last millennium with a method that preserves low-frequency variability, J. Clim., 24, 6013–6034, 2011.

Cook, E.R., Palmer, J.G., and D'Arrigo R.D.: Evidence for a 'Medieval Warm Period' in a 1,100 year tree-ring reconstruction of past austral summer temperatures in New Zealand, Geophys. Res. Lett., 29, 12.1-12.4, 2002.

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

Cook, E.R., Buckley, B.M., and D'Arrigo, R.D.: Inter-decadal climate variability in the Southern Hemisphere: Evidence from Tasmanian tree rings over the past three millennia, In Climate variations and forcing mechanisms of the last 2,000 years. NATO ASI Series, I, Jones P.D., Bradley R.S., Jouzel J., Eds., Springer-Verlag: Berlin, Germany, 41, 141–160, 1996.

Cooley, T.F.: Calibrated models, Oxf. Rev. Econ. Policy, 13, 55–69, 1997.

Crowley, T.J. and North, G.R.: Paleoclimatology, Oxford University Press: New York, N.Y., pp. 339, 1991.

Crowley, T.J.: Causes of climate change over the past 1000 years, Science, 289, 270–277, 2000.

Crowley, T.J. and Lowery, T.: How warm was the medieval warm period?, Ambio, 29, 51–54, 2000.

Cuven, S., Francus, P., and Lamoureux, S.: Mid to Late Holocene hydroclimatic and geochemical records from the varved sediments of East Lake, Cape Bounty, Canadian High Arctic, Quat. Sci. Rev., 30, 2651–2665, 2011.

D'Agostino, R.B., Belanger, A., and D'Agostino, R.B. Jr.: A suggestion for using powerful and informative tests of normality, Am. Stat., 44, 316–321, 1990.

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

Elliott, G., Rothenberg, T.J., and Stock, J.H.: Efficient tests for an autoregressive unit root, Econometrica, 64, 813–836, 1996.

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

Esper, J. and Frank D.: The IPCC on a heterogeneous Medieval Warm Period, Clim. Change, 94, 267–273, 2009.

Evans, M.N., Kaplan, A., and Cane, M.A.: Pacific sea surface temperature field reconstruction from coral $\delta 180$ data using reduced space objective analysis, Paleoceanography, 17, doi: 10.1029/ 2000PA000590, 2002.

Fan, J. and Yao, Q.: Nonlinear time series: Nonparametric and parametric methods, Springer-Verlag, N.Y., pp. vii+551, 2003.

Folland, C.K., Karl, T.R., Christy, J.R., Clarke, R.A., Gruza, G.V., Jouze, J., Mann, M.E., Oerlemans, J., Salinger, M.J., and Wang S.W.: Observed climate variability and change, In Climate change 2001: The

scientific basis, Houghton J.T., Griggs D.J., Nogner M., van der Linden P.J., Dai X., Maskell K., and Johnson C.A. Eds., Cambridge University Press: Cambridge, UK, pp. 99–181, 2001.

Fouka, P., Fröhlich, C., Spruit, H., and Wigley, T.M.: Variations in solar luminosity and their effect on the Earth's climate, Nature, 443, 161–166, 2006.

Fritts, H.C., Blasing, T.J., Hayden, B.P., and Kutzbach, J.E.: Multivariate techniques for specifying tree-growth and climate relationships and for reconstructing anomalies in paleoclimate, J. Appl. Meteorol. 10, 845–864, 1971.

Fritts, H.C.: Reconstructing large-scale climatic patterns from tree-ring data, The University of Arizona Press: Tucson, AZ., pp. 286, 1991.

Gagen, M., McCarroll, D., Loader, N.J., Robertson, I., Jalkanen, R., and Anchukaitis, K.J.: Exorcising the ‘segment length curse’: Summer temperature reconstruction since A.D. 1640 using non-detrended stable carbon isotope ratios from pine trees in northern Finland, Holocene, 17, 435–446, 2007.

Goodman, L.: On the exact variance of products, J. Am. Stat. Assoc., 55, 708–713, 1960.

Goodman, L.: The variance of the product of K random variables, J. Am. Stat. Assoc., 57, 54–60, 1962.

Graham, N.E., Ammann C.M., Fleitmann D., Cobb K.M., and Luterbacher J.: Support for global climate reorganization during the “Medieval Climate Anomaly”, Clim. Dyn., doi:10.1007/s00382-010-0914-z, 2010.

Granger, C.W.J. and Newbold, P.: Spurious regression in econometrics, J. Econom., 2, 111-120, 1974.

Gray, W.M., Sheaffer, J.D., and Landsea, C.W.: Climate trends associated with multidecadal variability of Atlantic hurricane activity, In Hurricanes: Climate and Socioeconomic Impacts, Diaz H.F.and Pulwarty R.S., Eds., Springer-Verlag: New York, NY, USA, pp. 15–53, 1997.

Guiot, J., Corona, C., and ESCARSEL members: Growing season temperatures in Europe and climate forcings over the past 1400 years, PLoS ONE, 5, e9972, 2010.

Hamilton, J.D.: Time Series Analysis, Princeton University Press, Princeton, NJ, USA, pp. 799, 1994.

Hantemirov, R.M. and Shiyatov, S.G.: A continuous multimillennial ring-width chronology in Yamal, Northwestern Siberia, Holocene, 12, 717–726, 2002.

Hendy, E.J., Gagan, M.K., Alibert, C.A., McCulloch, M.T., Lough, J.M., and Isdale, P.J.: Abrupt decrease in Tropical pacific sea surface salinity at end of Little Ice Age, Science, 295, 1511–1154, 2002.

Hollander, M. and Wolfe, D.A.: Nonparametric statistical methods, John Wiley & Sons, Inc.: Hoboken, N.J., pp. 816, 1999.

IPCC, 2001: Third Assessment Report, Available at: http://www.grida.no/publications/other/ipcc_tar/.

IPCC, 2007, Fourth Assessment Report, Available at: http://www.ipcc.ch/publications_and_data//ar4/syr/en/contents.html.

Jarque, C.M. and Bera, A.K.: A test for normality of observations and regression residuals, Int. Stat. Rev., 55, 163–172, 1987.

Jones, P.D., Osborn, J., and Briffa, K.R.: The evolution of climate over the last millennium, Science, 292, 662–667, 2001.

Juckes, M.N., Allen, M.R., Briffa, K.R., Esper, J., Hegerl, G.C., Moberg, A., Osborn, T.J., and Weber, S.L.: Millennial temperature reconstruction intercomparison and evaluation, Clim. Past, 3, 591–609, 2007.

Kalman, R.E.: A new approach to linear filtering and prediction problems, Trans. ASME, 82, 35–45, 1960.

Kennedy, M. and O’Hagan, A.: Bayesian calibration of computer models (with discussion), J. R. Stat. Soc. Ser. B, 63, 425–464, 2001.

Kim, D. and Perron, P.: Unit root tests allowing for a break in the trend function at an unknown time under both the null and alternative hypotheses, J. Econom.., 148, 1–13, 2009.

Koop, G. and Korobilis, D.: Manual to accompany Matlab package for Bayesian VAR models, Mimeo, University of Strathclyde, Glasgow, UK, 1-27, 2009.

Kruskal, W.H. and Wallis, W.A.: Use of ranks in one-criterion variance analysis, J. Am. Stat. Assoc., 47, 583–621.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., and Shin, Y.: Testing the null hypothesis of stationarity against the alternative of a unit root, J. Econom., 54, 159–178, 1992.

Li, J., Xie, S.P., Cook, E.R., Huang, G., D'Arrigo, R., Liu, F., Ma, J., and Zheng, X.-T.: Interdecadal modulation of El Niño amplitude during the past millennium, Nat. Clim. Change, 1, 114–118, 2011.

Lindholm, M., Jalkanen, R., Salminen, H., Aalto, T., and Ogurtsov, M.: The height-increment record of summer temperature extended over the last millennium in Fennoscandia, Holocene, doi: 10.1177/0959683610378875, 2010.

Ljungqvist, F.C.: Temperature proxy records covering the last two millennia: a tabular and visual overview, Geogr. Ann., 91, 11–29, 2009.

Loehle, C.: A 2000-year global temperature reconstruction based on non-tree ring proxies, Energy Environ., 18, 1049–1058, 2007.

Loehle, C., McCulloch, J.H.: Correction to: A 2000-year global temperature reconstruction based on non-tree ring proxies, Energy Environ., 19, 93–100, 2008.

Loso, M.G.: Summer temperatures during the Medieval Warm Period and Little Ice Age inferred from varved proglacial lake sediments in Southern Alaska, J. Paleolimnol., 41, 117–128, 2008.

Mann, M.E., Bradley, R.S., and Hughes, M.K.: Global-scale temperature patterns and climate forcing over the past six centuries, Nature, 392, 779–787, 1998.

Mann, M.E., Bradley, R.S., and Hughes, M.K.: Northern Hemisphere temperature during the past millennium: inferences, uncertainties and limitations, Geophys. Res. Lett., 26, 759–762, 1999.

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

Mann, M.E., Zhang, Z., Rutherford, S., Bradley, R.S., Hughes, M.K., Shindell, D., Ammann, C., Faluvegi, G., and Ni, F.: Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly, Science, 326, 1256–1260, 2009.

McIntyre, S. and McKitrick, R.: The M&M critique of the MBH98 Northern Hemisphere climate index: update and implications, Energy Environ., 16, 69–100, 2005.

McIntyre, S. and McKitrick, R.: Proxy inconsistency and other problems in millennial paleoclimate reconstructions, Lett. Proc. Natl. Acad. Sci., 106, E10, 2009.

McKitrick, R.: The Mann *et al.* Northern Hemisphere "Hockey- Stick" climate index: A Tale of due diligence, In Shattered consensus: the true state of global warming, Michaels P., Ed., Rowman and Littlefield: Washington D.C., pp. 20–48, 2006.

McShane, B.B. and Wyner, A.J.: A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable?, The Annals of Appl. Stat., 5, 5-44, 2011.

Moberg, A., Sonechkin, D.M., Holmgren, K., Datsenko, N.M., and Karlén, W.: Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data, Nature, 433, 613–617, 2005.

Montford, A.W.: The hockey-stick illusion, climategate and the corruption of science, Stacey International, London, UK, pp. 482, 2010.

Moore, J.J., Hughen, K.A., Miller, G.H., and Overpeck, J.T.: Little Ice Age recorded in summer temperatures from varved sediments of Donard Lake, Baffin Island, Canada, J. Paleolimnol., 25, 503–517, 2001.

Morice, C.P., Kennedy, J.J., Rayner, N.A., and Jones, P.D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, J. Geophys. Res., doi: 10.1029/2011JD017187, 2012.

Neukom, R., Luterbacher, J., Villalba, R., Küttel, M., Frank, D., Jones, P.D., Grosjean, M., Wanner, H., Aravena, J.-C., Black, D.E., Christie, D.A, D'Arrigo, R., Lara, A., Morales, M., Soliz-Gamboa, C., Srur, A., Urrrutia, R., and von Gunten, L.: Multiproxy summer and winter surface air temperature field reconstructions for Southern South America covering the past centuries, Clim. Dyn., doi:10.1007/s00382-010-0793-3, 2010.

Noriega, A.E. and Ventosa-Santaulària, D.: Spurious regression and trending variables, Oxf. Bull. Econ. Stat., 69, 439–444, 2007.

O'Keefe, W. and Kueter, J.: Climate models: A primer, The George C. Marshall Institute, Washington D.C., 1-22, 2004.

Orfanidis, S.J.: Introduction to signal processing, Sophocles J. Orfanidis, Rutgers University, pp. xiii+795, 2010.

Pederson, G.T., Gray, S.T., Woodhouse, C.A., Betancourt, J.L., Fagre, D.B., Littell, J.S., Watson, E., Luckman, B.H., and Graumlich, L.J.: The unusual nature of recent snowpack declines in the North American Cordillera, Science, 333, 332–335, 2011.

Perron, P. and Zhu, X.: Structural breaks with deterministic and stochastic trends, J. Econom.129, 65–119, 2005.

Rougier, J.: Discussion of "A statistical analysis of multiple temperature proxies: are reconstructions of surface temperatures over the last 1000 years reliable?" by Mcshane and Wyner, The Annals of Appl. Stat., 5, 1-3, 2011.

Royston, P.: Algorithm AS R94, Appl. Stat., 44, 547-550, 1995.

Said, E. and Dickey, D.A.: Testing for unit roots in autoregressive moving average models of unknown order, Biometrika, 71, 599–607, 1984.

Salzer, M.W. and Kipfmueller, K.F.: Reconstructed temperature and precipitation on a millennial timescale from tree-rings in the Southern Colorado Plateau, U.S.A., Clim. Change, 70, 465-487, 2005.

Sansó, B. and Forest, C.: Statistical calibration of climate system properties, J. R. Stat. Soc., 58, 485–503, 2009.

Savitzky, A. and Golay, M.J.E., Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem., 36, 1627–1639, 1964.

Scafetta, N.: Discussion on common errors in analyzing sea level accelerations, solar trends and global warming, Pattern Recogn. Phys.1, 37–57, 2013.

Shaviv, N.: On climate response to changes in the cosmic ray flux and radiative budget, J. Geophys. Res., 110, 1–15, 2005.

Shindell, D.T., Schmidt, G.A., Mann, E.M., and Faluvegi, G.: Dynamic winter climate response to large tropical volcanic eruptions since 1600, J. Geophys. Res., 109, 1-12, 2004.

Sinha, A., Berkelhammer, M., Stott, L., Mudelsee, M., Cheng, H., and Biswas, J.: The leading mode of Indian summer monsoon precipitation variability during the last millennium, Geophys. Res. Lett., 38, doi: 10.1029/2011GL047713, 2011.

Smerdon, J., Kaplan, A., and Chang, D.: On the origin of the standardization sensitivity in RegEM climate field reconstructions, J. Clim., 21, 6710–6723, 2008.

Stahle, D.W., Diaz, J.V., Burnette, D.J., Paredes, J.C., Heim, R.R. Jr., Fye, F.K., Acuna Soto, R., Therrell, M.D., Cleaveland, M.K., and Stahl, D.K.: Major Mesoamerican droughts of the past millennium, Geophys. Res. Lett., 38, doi: 10.1029/2010GL046472, 2011.

Stambaugh, M.C., Guyette, R.P., McMurry, E.R., Cook, E.R., Meko, D.M., and Lupo, A.R.: Drought duration and frequency in the U.S. Corn Belt during the last millennium (A.D. 992–2004), Agric. For. Meteorol., 151, 154–162, 2011.

Stock, J.H. and Watson, M.W.: New indexes of coincident and leading economic indicators, In NBER Macroeconomics Annual, Blanchard O.J. and Fischer S. Eds., NBER, Cambridge, MA, 351–409, 1989.

Svensmark, H. and Frijs-Christensen, E.: Reply to Lockwood and Fröhlich. The persistent role of the sun in climate forcing, Danish National Space Center Scientific Report, 3, 1-6, 2007.

Tand, M., Liu, T.S., Hou, J., Qin, X., Zhang, H., and Li, T.: Cyclic rapid warming on centennial-scale revealed by a 2650-year stalagmite record of warm season temperature, Geophys. Res. Lett., 30, doi: 10.1029/2003GL017352, 2003.

Tellinghuisen, J.: Inverse vs. classical calibration for small data sets, Fresenius J. Anal. Chem., 368, 585–588, 2000.

Tingley, M.P.: BARCAST and the Kalman smoother, Manuscript, 2010a.

Tingley, M.P. and Huybers, P.: A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: Development and applications to paleoclimate reconstruction problems, J. Clim., 23, 23, 2759–2781, 2010b.

Tingley, M.P.: Spurious predictions with random time series: The LASSO in the context of paleoclimatic reconstructions. A Discussion of “A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable?, The Annals of Appl. Stat., 5, 83-87, 2011.

Tingley, M.P. and Li, B.: Comments on “Reconstructing the NH mean temperature: Can underestimation of trends and variability be avoided?”, J. Clim. 2012, 25, 3441–3446, 2012.

Trouet, V., Esper, J., Graham, N.E., Baker, A., Scourse, J.D., and Frank, D.C.: Persistent positive North Atlantic oscillation mode dominated the medieval climate anomaly, Science, 324, 78-80, 2009.

Usoskin, I.G., Mursula, K., Solanki, S.K., Schüssler, M., and Alanko-Huotari, K.: Reconstruction of solar activity for the last millennium using ^{10}Be data, Astron. Astrophys.,413, 745–751, 2004a.

Usoskin, I.G., Mursula, K., Solanki, S.K., Schüssler, M., and Alanko-Huotari, K.: Millennium-scale sunspot number reconstruction: Evidence for an unusually active sun since the 1940s, Phys. Rev. Lett., 91, 211101:1–211101:4, 2004b.

Usoskin, I.G., Solanki, S.K., and Korte, M.: Solar activity reconstructed over the last 7000 years: The influence of geomagnetic field changes, Geophys. Res. Lett., 33, L08103:1–L08103:4, 2006.

Ventosa-Santaulària, D.: Spurious regression, J. Probab. Stat., 2009, 1–27, 2009.

Von Storch, H., Zorita, E., Jones, J.M., Dimitriev, Y., González-Rouco, F., and Tett, S.F.B.: Reconstructing past climate from noisy data, Science, 306, doi:10.1126/science.1096109, 2004.

Wahl, E.R., Anderson, D.M., Bauer, B.A., Buckner, R., Gille, E.P., Gross, W.S., Hartman, M., and Shah, A.: An archive of high-resolution temperature reconstructions over the past 2+ millennia, Geochem. Geophys. Geosyst., 11, doi:10.1029/2009GC002817, 2010.

Wilson, R., Wiles, G., D’Arrigo, R., and Zweck, C.: Cycles and shifts: 1,300 years of multi-decadal temperature variability in the Gulf of Alaska, Clim. Dyn., 28, 425–440, 2006.

Yan, H., Sun, L., Wang, Y., Huang, W., Qiu, S., and Yang, C.: A record of the Southern Oscillation Index for the past

2,000 years from precipitation proxies, Nat. Geosci., 4, 611–614, 2013.

Zunli, L., Rickaby, R.E.M., Kennedy, H., Kennedy, P., Pancost, R.D., Shaw, S., Lennie, A., Wellner, J., and Anderson, J.B.: An Ikaite record of late Holocene climate at the Antarctic Peninsula, Earth Planet. Sci. Lett., 325-326, 108–115, 2012.

TABLE I. Probabilities of "hockey-stick" behaviour (HSB) for select factor f and data-generating processes of stationary (STA) and nonstationary (NON) series $\hat{Y}_t$ together with their computed mean standard deviations (SD)

Normal DGP, (1)				Nonnormal DGP, skewness = 0.95, (2)				Nonnormal DGP, skewness = 1.95, (3)			
Prob. HSB of NON $\hat{Y}_t$	Prob. HSB of STA $\hat{Y}_t$	Mean SD of NON $\hat{Y}_t$	Mean SD of STA $\hat{Y}_t$	Prob. HSB of NON $\hat{Y}_t$	Prob. HSB of STA $\hat{Y}_t$	Mean SD of NON $\hat{Y}_t$	Mean SD of STA $\hat{Y}_t$	Prob. HSB of NON $\hat{Y}_t$	Prob. HSB of STA $\hat{Y}_t$	Mean SD of NON $\hat{Y}_t$	Mean SD of STA $\hat{Y}_t$
1.000	1.000	0.003	0.000	1.000	1.000	0.003	0.000	1.000	1.000	0.003	0.000
0.999	1.000	0.074	0.007	1.000	1.000	0.080	0.007	0.999	1.000	0.074	0.007
0.928	1.000	0.144	0.013	0.900	1.000	0.160	0.013	0.932	1.000	0.144	0.013
0.832	1.000	0.224	0.020	0.780	1.000	0.234	0.020	0.810	1.000	0.225	0.020
0.719	1.000	0.296	0.026	0.670	1.000	0.318	0.026	0.724	1.000	0.295	0.026
0.458	1.000	0.878	0.078	0.440	1.000	0.952	0.078	0.471	1.000	0.879	0.078
0.396	0.984	1.470	0.130	0.400	1.000	1.534	0.130	0.404	1.000	1.480	0.130
0.348	0.339	2.036	0.182	0.310	0.190	2.272	0.183	0.369	1.000	2.033	0.183
0.374	0.183	2.661	0.235	0.330	0.160	2.814	0.234	0.344	0.159	2.660	0.234
0.351	0.193	2.893	0.261	0.360	0.110	3.271	0.262	0.327	0.135	2.981	0.261

Table II. Basic descriptive statistics of the BEA series, p-values of the ADF and KPSS tests for stationarity and of the Jarque-Bera test statistic for normality, and critical dates

Statistics	Hadcrut4_GL	Hadcrut4_NH	Hadcrut4_SH
Mean	−0.125	−0.081	−0.168
Variance	0.068	0.084	0.061
Volatility	2.082	3.570	1.465
ADF test p-value	0.432	0.515	0.242
KPSS test p-value	0.011	0.010	0.010
Max. temperature date	2005	2005	1998
Min. temperature date	1911	1862	1911

Table III. Summary information and aggregate test statistic results of the CCP datasets

Order and Series Name	Number of series included (1)	Maximum length (2)	Minimum length (3)	Percent nonstationarities (4)	Percent nonnormalities (5)	Average absolute correlation with BEA beyond 1850 (6)
1. BÜNTGEN	4	2503	1010	50	25	0.214
2. CHRISTIANSEN	19	1001	971	0	32	0.294
3. COOK	2	3592	3592	100	0	0.327
4. CUVEN	2	4198	4198	0	0	0.141
5. ENSO_LI	1	1103	1103	0	0	0.091
6. ENSO_YAN	1	1906	1906	0	0	0.269
7. FS_GAGEN	3	1116	1116	100	0	0.345
8. FS_LINDHOLM	2	1263	1263	50	0	0.035
9. GUIOT	125	1408	1408	0	34	0.346
10. LJUNGQVIST	66	2001	1738	62	6	0.374
11. LOSO	1	1557	1557	0	0	0.408
12. ME_STAHLE	5	1238	1238	0	0	0.038
13. NEUKOM	3	1096	1096	0	33	0.528
14. PEDERSON	9	1637	1637	0	33	0.146
15. SINHA	1	1383	1383	100	100	0.401
16. STAMBAUGH	2	1013	1013	0	0	0.052
17. TROUET	1	959	959	0	0	0.244
18. WILSON	1	1276	1276	100	100	0.340
19. G_CUBED	10	4062	981	40	0	0.551

Table IV. Averaged results of TVP estimation: factors, scores of variance tests, coefficients and t-statistics

Order and Series Name	Percent value of factor *f* (1)	Score of the three equal-variance tests (2)	Slope coefficient (3)	*t*-stat of slope coefficient (4)	Constant coefficient (5)	*t*-stat of constant coefficient (6)
1. BÜNTGEN	1.218	3.00	0.005	7.968	0.022	2.668
2. CHRISTIANSEN	1.254	3.00	−0.008	21.804	0.018	1.827
3. COOK	0.734	3.00	0.083	16.943	0.020	3.856
4. CUVEN	0.513	3.00	−0.105	48.686	0.020	3.590
5. ENSO_LI	2.197	3.00	0.191	24.049	0.014	1.479
6. ENSO_YAN	0.059	3.00	−1.098	155.127	0.016	1.108
7. FS_GAGEN	0.443	3.00	−0.044	4.750	0.010	1.224
8. FS_LINDHOLM	1.724	3.00	0.124	15.410	0.008	0.908
9. GUIOT	1.712	3.00	0.046	20.889	0.016	2.146
10. LJUNGQVIST	0.312	3.00	0.618	158.173	0.058	4.411
11. LOSO	1.278	3.00	0.261	34.426	0.011	1.213
12. MESO_STAHLE	1.054	3.00	−0.187	21.555	0.017	1.694
13. NEUKOM	1.137	3.00	0.278	29.755	0.018	1.542
14. PEDERSON	0.708	3.00	−0.141	30.614	0.013	1.906
15. SINHA	0.543	3.00	−0.241	28.046	0.016	1.540
16. STAMBAUGH	1.626	3.00	0.011	1.269	0.040	4.828
17. TROUET	0.138	3.00	0.272	24.916	0.028	2.185
18. WILSON	1.309	3.00	0.319	36.658	0.027	2.461
19. G_CUBED	0.813	3.00	0.449	66.184	0.010	1.907

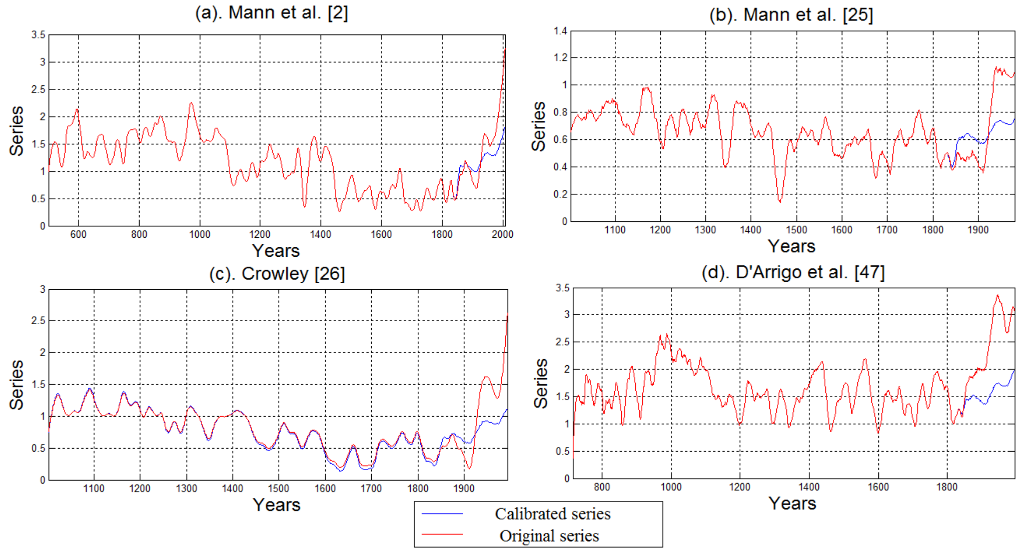

Fig.1. Comparative plots of select normalized and smoothed original series from G_CUBED and corresponding calibrated series