INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1825

RMetS
Royal Meteorological Society

Short Communication

How do we tell which estimates of past climate change are correct?

Steven C. Sherwood,[a]* Holly A. Titchner,[b] Peter W. Thorne[b] and Mark P. McCarthy[b]
[a] *Department of Geology and Geophysics, Yale University, New Haven, CT, USA*
[b] *Met Office Hadley Centre, Exeter, UK*

ABSTRACT: Estimates of past climate change often involve teasing small signals from imperfect instrumental or proxy records. Success is often evaluated on the basis of the spatial or temporal consistency of the resulting reconstruction, or on the apparent prediction error on small space and time scales. However, inherent methodological trade-offs illustrated here can cause climate signal accuracy to be unrelated, or even inversely related, to such performance measures. This is a form of the classic conflict in statistics between minimum variance and unbiased estimators. Comprehensive statistical simulations based on climate model output are probably the best way to reliably assess whether methods of reconstructing climate from sparse records, such as radiosondes or paleoclimate proxies, actually work on longer time scales. Copyright © 2008 Royal Meteorological Society

KEY WORDS climate change; statistics; climate reconstruction

Received 10 March 2008; Revised 28 October 2008; Accepted 31 October 2008

1. Introduction

During the last decade, significant attention has focused on quantitative reconstruction of past climate change. This includes efforts to remove artefacts (specifically, bias changes) from ground-, balloon- and satellite-based instrumental records (CCSP, 2006) as well as efforts to estimate pre-instrumental climate changes using proxy records (NRC, 2006). Invariably this involves fitting the available data to some kind of statistical model. Recent experience tells us that in such cases the uncertainty in the resulting climate change is dominated by structural uncertainty (that is associated with analysis assumptions) rather than parametric uncertainty (that is due to limitations in the quantity or quality of data, given correct assumptions) (e.g. Thorne *et al*., 2005a). Unfortunately there is no obvious way of confirming the veracity of a reconstruction, except perhaps by consensus among a number of truly independent efforts. And even this may not be reliable since similar mistakes could be made by all groups.

A common way of evaluating individual reconstruction efforts, and the assumptions behind them, is by looking at their impact on apparent inconsistencies in the raw record at short time and/or space scales. This has been especially prevalent in evaluating the success of 'homogenization' efforts designed to remove artefacts in instrumental records, but may still be tempting for other problems with similar characteristics such as proxy reconstruction. The purpose of this note is to explain why this is often a poor strategy if characterizing the longer-term changes is the main aim, and to support an alternative.

2. Prototypical examples

We identify two types of underlying problem one typically faces: inhomogeneous data and imprecise data. Inhomogeneous data possess a relationship with the desired climate measure that may be precise but is not stationary over time: for example a satellite orbit might drift, a replacement sensor may have a different calibration from the original, or a proxy measure of temperature may be affected by some other time-varying influence such as changes in Earth's orbit. Imprecise data are imperfectly correlated with the desired climate measure. For example, a proxy temperature record with random fluctuations because of local influences, or indirect satellite estimates of rainfall, may have a steady but uncertain relationship to the desired observable. In this case the statistics of the underlying data may be stationary, but hard to estimate from the available data.

In general, one can face both of these issues to some degree. To illustrate them, we consider two examples each isolating one of the two issues. As is typical, there will be key parameters that one cannot quantify from first principles and must estimate empirically.

* Correspondence to: Steven C. Sherwood, Department of Geology and Geophysics, Yale University, New Haven, CT, USA.
E-mail: Steven.Sherwood@yale.edu

Copyright © 2008 Royal Meteorological Society

2.1. An inhomogeneous-data prototype: bias change

Suppose one has a data series with an underlying climate signal plus stochastic variability. To keep our example as simple as possible we suppose a linear trend and white noise, respectively. Midway through the record (at time t_0), the observing bias changed by an amount h. The existence and time of this is known, but not the amount. We consider three possible approaches to calculate the underlying trend:

1. Perform linear trend analysis, ignoring the bias change.
2. Estimate h by subtracting the mean of all the data before t_0 from all the data after t_0, then adjust the data by adding h to all data before t_0. Perform linear trend analysis on the revised data.
3. Perform bivariate linear regression of the data onto two functions, a linear function of time and a step function with change at t_0, retaining the first regression coefficient as the trend and the second as h.

The probability distribution of retrieved trend, for random realizations of the natural variability and for a negative value of h against a positive total trend of similar magnitude, is shown in Figure 1. Only method (3) returns an unbiased trend estimate; under the circumstances (e.g. with no additional information on the natural variability), this method is in fact the optimal unbiased method – it is not possible to do any better without accepting bias. Method (1) underestimates the trend because of the unrecognized artefact, while (2) underestimates it because the trend itself is taken to be partly artificial. Nonetheless, (2) shows so much less scatter than the others that its root-mean-square (RMS) departure from the truth over all realizations is the best of the three. By both measures, (2) is preferable to (1), but whether (2) or (3) is better depends on whether one is more worried about bias or random error (The bias and RMS error levels obviously depend on the parameters (trend, h and noise), here chosen arbitrarily for illustration. Method (3) is never biased, while the others are always biased for non-zero trend and h, and (3) will always have the greatest variance. While the RMS error of (3) could be smaller than that of the others if the noise were small enough compared with h or the trend, this would be unlikely in practice since it would imply either intolerable instrumental problems or a trivial trend estimation situation.). If this process was to be repeated at many observing sites, and the trends averaged, then bias would be more important.

2.2. An imprecise-data prototype: gain estimation

Again suppose that one seeks a linear climate signal amid white noise. This time, one has two climate records of significantly different length and quality. For simplicity we suppose the short one a perfect record of the climate $T(t)$, while the longer one $X(t)$ is related to T by $X = gT + e$ where e is random noise and g an unknown gain. Recovery of good climate signals on long time scales requires correct estimation of g. Here we consider two possible approaches:

1. (Forward regression) regress T onto X during the period of overlap, setting g to the regression coefficient;
2. (Reverse regression) regress X onto T, setting g to the reciprocal of the regression coefficient.

Performing a similar analysis as in the first example, one obtains a similar result (Figure 2): approach (1) yields the smaller RMS difference (0.71 vs 0.99) between gX and T during the reference period, but yields a long-term trend that is too small by a factor of two (because

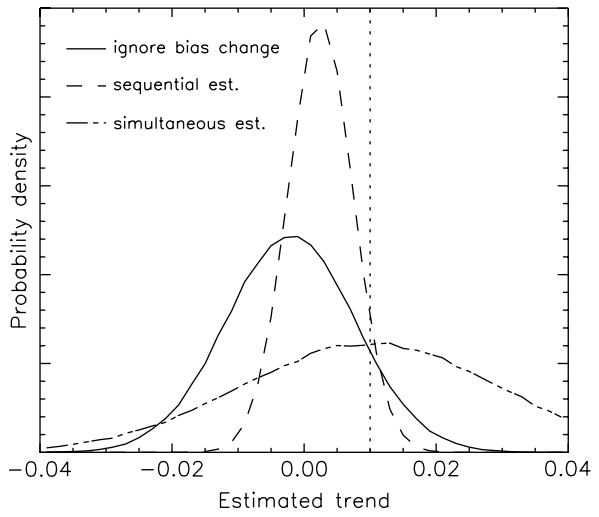

Figure 1. Probability distribution of the trend estimated by three procedures described in the text, with correct trend indicated by vertical dotted line. Results shown are for a time series of 50 data points, with an underlying true total trend of +0.5 units, a downward bias change of 0.4 units midway through the record and white noise of unit variance.

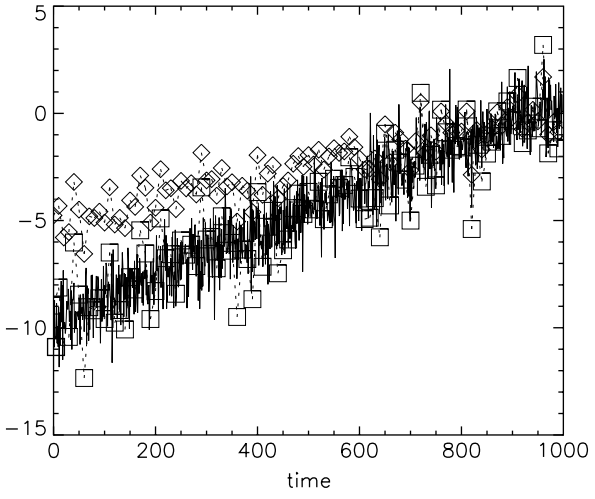

Figure 2. Correct climate time series (solid line), series reconstruction by forward regression (diamonds) and reconstruction by reverse regression (boxes). Forward regression has the best RMS agreement with the actual climate during the reference period (last 10% of the time interval), but yields a long-term trend roughly half as large as the true one.

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/joc

regression is biased when the explanatory variable is noisy). Despite having larger RMS error, however, (2) yields the correct trend (since regression works fine when the response variable is noisy).

2.3. A more practical example

To show that the above characteristics do not necessarily depend on naive or simplistic approaches, we present one final example. Recently, to test the robustness of the procedure used previously by Thorne *et al.* (2005b) to produce the HadAT (Met Office Hadley Centre Atmospheric Temperature) homogenized radiosonde dataset, McCarthy *et al.* (2008) created an ensemble of 100 different versions of an automated variant of that procedure. The basic methodology rested on a strategy of iteratively adjusting data by comparing station series with neighbour-based composites, thereby identifying and adjusting breaks. It therefore tends to minimize the RMS differences between station series that are relatively close together. In each version, the automated variant was altered by significantly changing one or more of its key parameters.

Each version was run on each of four different simulated, inhomogeneous datasets, making a total of 400 trials. The four datasets were created by sampling the output of a run of the HadAM3 climate model with prescribed SSTs and anthropogenic and natural forcings (Pope *et al.*, 2000) in the spatio-temporal data-availability pattern of the actual radiosonde database, adding white noise to approximate sampling effects, then imposing several thousand bias-change artefacts whose character varied substantially among the four datasets (Titchner *et al.*, In Press). Thus, the four simulations were all based on the same simulated 'truth' but had different underlying assumptions as to the type of artefacts occurring, with some much more pessimistic (thus difficult to homogenize) than others.

Since in these test cases the truth is known *a priori*, performance can be assessed unambiguously as with our more idealized examples above. The results of each homogenization trial were evaluated according to three criteria: the accuracy of the trend in the adjusted data, the RMS difference between adjusted and known actual temperatures and the internal consistency of the adjusted data (RMS difference between temperatures and composites of nearby neighbours). The RMS difference between adjusted and actual temperatures and internal consistency measures are very highly correlated in all cases (not shown). However, in ensemble results for all four simulations, despite the wide range of recovered trends (Titchner *et al.*, In Press), minimization of the RMS error of individual station records (whether measured against the truth or against neighbours) was weakly or unrelated to successful recovery of large-scale trends (see one example in Figure 3). On the basis of correlations between neighbour-based RMS rankings and trend skill recovery rankings, in no case did the RMS rankings explain more than one-third of the variance in the trend

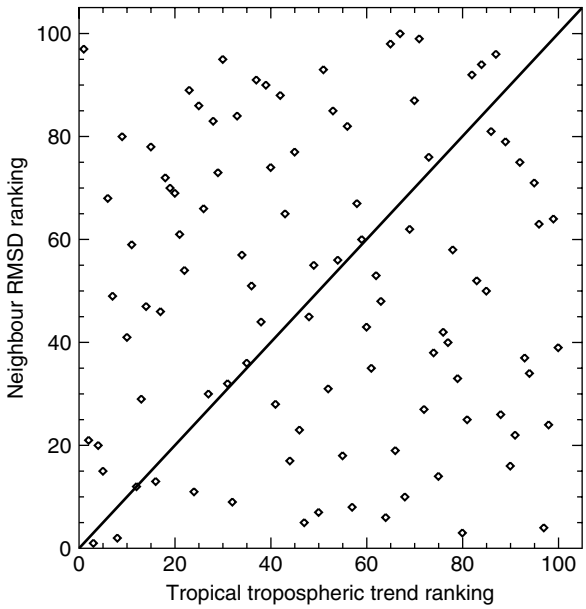

Figure 3. Relationship between the success ranks according to (x-axis) tropical tropospheric (MSU 2LT) trend error and (y-axis) RMS deviation between estimated temperatures and those of neighbours, of 100 different implementations of the automated radiosonde homogenization procedure described by McCarthy *et al.* (2008) applied to a set of simulated data. Correlation coefficient is 0.01; those for three other datasets ranged from 0.15 to 0.58.

ranking skill for a 100 member ensemble. The correlation between rankings was better for those simulated datasets which *a priori* should be easiest to adjust (few, large breaks), and reduced to zero for the more pessimistic case shown in Figure 3 (many, small breaks).

3. Discussion

Our intuition tells us that a better method for retrieving past climate should improve all measures of success. There is clearly some truth in this, as a genuinely poor procedure will indeed produce bad results across the board (for example, method (1) in our first example). However, once one has rejected such clearly inferior approaches, one begins to encounter a tradeoff in which improving one measure of success potentially sacrifices another.

This situation may be familiar to students of statistics in the form of a conflict between unbiased and minimum-variance estimators. The most familiar example is probably the sample variance estimator

$$\mathrm{var}(X) = \frac{\sum (X - \langle X \rangle)^2}{n - a} \qquad (1)$$

which is unbiased for $a = 1$ but has minimum RMS error when $a = 0$. Since, in practice, one usually has $n \gg 1$, the choice of estimator hardly matters in this particular case. This may leave the impression that the distinction is unimportant in general. Our examples above clearly demonstrate, however, that this is not the

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/joc

case even for very simple time-series analysis problems. In records with broad-spectrum variability, successful retrieval of change on long time scales (unbiased estimation) is not necessarily well predicted by performance on shorter ones (minimum variance estimation).

4. Conclusion

While the setup of our two prototypes suggests applications, respectively, to instrumental data homogenization or paleoclimate reconstruction by proxy data, the results illustrate a general principle that may apply to any dataset whose relationship to the underlying target quantity is uncertain. For example, global rainfall estimation relies on satellite proxy information to extrapolate horizontally from ground observations in a manner analogous to the temporal extrapolation required for paleoclimate reconstruction. Intercomparisons of satellite rainfall products have sometimes found that those performing best on short time and space scales are inferior to others on longer scales (Ebert *et al*., 1996), and useful estimation of changes in global-mean rainfall remains a challenge. Similarly, weather forecasts with lower mean-squared error often have larger biases and lead to larger systematic errors in decision-making (Ehrendorfer and Murphy, 1988).

Recent efforts to better estimate past climate change have sometimes been judged on the basis of variance minimization metrics such as RMS error or internal consistency checks. Efforts to quantify atmospheric warming from inhomogeneous radiosonde data, in particular, have nearly all done this (Lanzante *et al*., 2003; Thorne *et al*., 2005b; Haimberger, 2007) and at least one recent comparison of two efforts to homogenize satellite temperatures has picked a winner, in part, on the basis of such a metric (Christy and Norris, 2006). It is our hope that the present discussion will bring about a reassessment of such assumptions of transferability between different measures of success.

In its place, we propose that methods for reconstructing climate should be evaluated by testing them on simulated datasets constructed as realistically as possible, including suspected proxy/instrumental artefacts or biases and sampling patterns. Those attempting to reconstruct the climate of the past millennium from proxies now recognize this (at least from the standpoint of sampling) and some have begun to test their methodologies in this way (Mann and Rutherford, 2002; von Storch *et al*., 2004; Mann *et al*., 2005). These efforts reveal that success can depend on the details of either the actual climate change (von Storch *et al*., 2004) and/or the location of artificial instrument problems (Titchner *et al*., In Press). They have nonetheless been informative as to the role of sampling and analysis limitations on results. Examination of the more difficult cases can lead to methodological improvements and quantitative reassessment of the findings when methods are applied to real-world data, where we are not afforded such a luxury of knowing the answer *a priori*.

Acknowledgements

This study was performed while the first author was visiting and partially supported by the Hadley Centre. Met Office Hadley Centre authors were supported by the Joint Defra and MoD Programme, (Defra) GA01101 (MoD) CBC/2B/0417_Annex C5.

References

CCSP. 2006. *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences, A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research, National Oceanic and Atmospheric Administration*, Karl TR, Hassol S, Miller C, Murray W. (eds). National Climatic Data Center: Asheville; 164.

Christy JR, Norris WB. 2006. Satellite and VIZ-radiosonde intercomparisons for diagnosis of nonclimatic influences. *Journal of Atmospheric and Oceanic Technology* **23**: 1181–1194.

Ebert EE, Manton MJ, Arkin PA, Allam RJ, Holpin GE, Gruber A. 1996. Results from the GPCP algorithm intercomparison programme. *Bulletin of the American Meteorological Society* **77**: 2875–2887.

Ehrendorfer M, Murphy AH. 1988. Comparative-evaluation of weather forecasting systems–sufficiency, quality, and accuracy. *Monthly Weather Review* **116**: 1757–1770.

Haimberger L. 2007. Homogenization of radiosonde temperature time series using innovation statistics. *Journal of Climate* **20**: 1377–1403.

Lanzante JR, Klein SA, Seidel DJ. 2003. Temporal homogenization of monthly radiosonde temperature data. Part II: Trends, sensitivities, and MSU comparison. *Journal of Climate* **16**: 241–262.

Mann ME, Rutherford S. 2002. Climate reconstruction using 'pseudoproxies'. *Geophysical Research Letters* **29**: Art. No. 1501. DOI: 10.1029/2001GL014554.

Mann ME, Rutherford S, Wahl E, Ammann C. 2005. Testing the fidelity of methods used in proxy-based reconstructions of past climate. *Journal of Climate* **18**: 4097–4107.

McCarthy MP, Titchner HA, Thorne PW, Tett SF, Haimberger L, Parker DE. 2008. Assessing bias and uncertainty in the HadAT adjusted radiosonde climate record. *Journal of Climate* **21**: 817–832.

NRC. 2006. *Surface Temperature Reconstructions for the Last 2,000 Years*, Technical report. National Academies: Washington, DC.

Pope VD, Gallani ML, Rowntree PR, Stratton RA. 2000. The impact of new physical parameterizations in the Hadley Centre climate model: HadAM3. *Climate Dynamics* **16**: 123–146.

Thorne PW, Parker DE, Christy JR, Mears CA. 2005a. Uncertainties in climate trends–Lessons from upper-air temperature records. *Bulletin of the American Meteorological Society* **86**: 1437.

Thorne PW, Parker DE, Tett SFB, Jones PD, McCarthy M, Coleman H, Brohan P. 2005b. Revisiting radiosonde upper-air temperatures from 1958–2002. *Journal of Geophysical Research* **110**(D18): 105.

Titchner HA, Thorne PW, McCarthy MP, Tett SFB, Haimberger L, Parker DE. Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments. *Journal of Climate* In Press.

von Storch H, Zorita E, Jones JM, Dimitriev Y, Gonzalez-Rouco F, Tett SFB. 2004. Reconstructing past climate from noisy data. *Science* **306**: 679–682.

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/joc