Is temperature or the temperature record rising?

David R.B. Stockwell*

September 26, 2012

Abstract

In this paper, we prove a logical circularity undermines the validity of a commonly used method of homogenizing surface temperature networks. High rates of type I error due to circularity may explain the exaggeration of surface warming found in official temperature networks.

1 Introduction

Homogenization consists of adjusting the baseline of sections of a temperature or rainfall series up or down in an effort to mitigate the effects of changes in location or instrumentation. Recent audits of surface temperature networks have found that official, homogenized networks show more warming than the raw temperature data: in Australia +0.9C vs +0.7C per century [1], in New Zealand +0.9C vs +0.3C per century [2], and globally +0.7C vs +0.4C [3] respectively. A recent study by the Australian Bureau of Meteorology (BoM) also reported a similar variation of +1.09C vs +0.69C between the homogenized ACORN and the non-homogenized WNAWAP networks respectively [4, 5].

These differences between the trends of homogenized and unadjusted data are quite large. Homogenization seems to be favored in official meteorological networks, ostensively to repair micro-site shifts and temporal inhomogeneities due to changes in observing practices, instrumentation, or reporting. The BoM acknowledges, however, "clear evidence in favour of this hypothesis is yet to be obtained" [5].

*Central Queensland University, Australia. d.stockwell@cqu.edu.au

There are potentially valid reasons to adjust raw data, particularly where artificial discontinuities and outliers can be proven and quantified. There are also potentially valid reasons for a non-climatic warm bias in early records, justifying adjustments to maintain consistency. Demonstrable improvements in the quality of shelter, instrumentation and software [6], or a consistent pattern of micro-site shifts to cooler, higher elevation sites away from creeping urbanization are two such non-climatic factors.

There are also compelling reasons not to alter data. Adjustments destroy the ascendency of record high and low temperatures. Adjustments may add more errors than they remove. Adjustments that amplify warming are clearly not acceptable if their only justification is a departure from the overall trend.

Here, we show that homogenization by comparison of a target site with a reference climatology can introduce a bias into a temperature network. Such biases contaminate other alarming global warming studies [7, 8, 9, 10].

2 Analysis

In the Standard Normal Homogenization Test (SNHT), individual target series are examined for jumps indicative of station moves or other problems. The SNHT often optimizes information metrics such as the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) in conjunction with break-tests based on F statistics using a Chow and supF test [11]. These tests determine if, where and by how much the target series should be adjusted.

Another approach is to compare the target series S with a regional climatology Rp, using either a weighted average of neighbors [12] or an exhaustive pairwise comparison [13]. This amounts to applying SNHT to the difference series $D = S - Rp$. The breaks identified in the difference data determine the adjustments to S.

While there is a known tendency for aliasing whereby the target series becomes indistinguishable from the reference [13], the circular reasoning is not widely appreciated.

The aliasing of the reference homogenization method is evident from the mathematical expression for the homogenized target series $H(S)$. The reference series replaces the target series.

$H(S) = S - D = S - (S - Rp) = Rp$

Spurious breaks in D erroneously coerce the trend of S towards the trend of the regional climatology Rp. Even if the breaks are real, aliasing biases estimate of the magnitude of the jump towards the reference trend.

To demonstrate aliasing on historical data, we selected a surface station whose trend deviates from the continental mean temperature trend, but does not show any obvious inhomogeneity. Deniliquin was not chosen to be representative. Many stations will display the average trend. Deniliquin was chosen to show the coercion of the trend of *any* station that deviates from the average global warming trend into the warming trend, irrespective of its trend or quality.

Figure 1 illustrates the steps in the homogenization process on the raw minimum temperature for Deniliquin Post Office site number 074128 (series offset for clarity). After subtracting the raw data ("Raw") from the Australian temperature average, an iterative Chow test finds a significant break in 1975. Adjustment to the cooling "Raw" series produces a series with a warming trend similar to the version of Deniliquin in the Bureau of Meteorology (BoM) High Quality dataset "BoM HQ".

The metadata for Deniliquin report a large adjustment to the minimum in 1971 by -0.8C in concert with a station relocation of 1km to the north west and another in 1951. However, the metadata does not indicate whether the move changed the stations average temperature. The neighboring stations of Eucha and Hay are cooling or flat (-0.14C and 0.08C per decade respectively) and do not seem to have any obvious discrepancies with Deniliquin to explain the 1971 adjustment. Moreover, there does not seem to be anything unusual about the diurnal range around 1971. The Deniliquin adjustments do not seem to be justified on any basis other than departure from the global warming trend.

Breaks can potentially be justified using a robust statistical test for structural change on autocorrelated series, like the empirical fluctuation process (EFP), in particular the recursive CUSUM tests [14]. Figure 2 shows the EFP (dashed lines) on raw and difference series for Deniliquin where crossing the red line indicates a significant change in level. The cooling raw temperatures at Deniliquin are not yet significant, but the difference series is significant.

Is the change in level at Deniliquin significant? It could be argued that comparison with a reference increases the power of the test, so that breaks

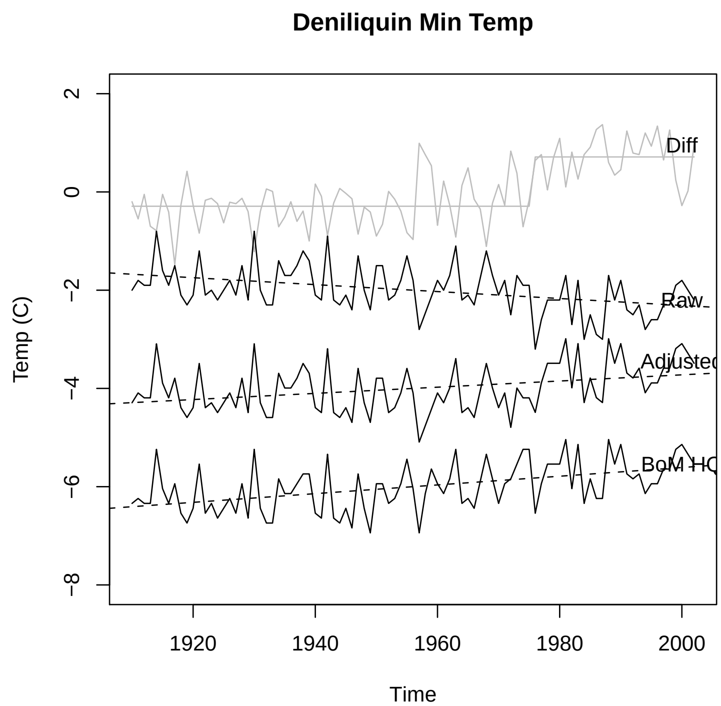

Figure 1: Homogenization of the temperature data at Deniliquin, Australia between 1910 and 2007 (”Raw”) (series offset for clarity). ”Diff” in gray, is the subtraction of ”Raw” from the reference series, the average annual Australian temperature. An iterative Chow test in the package strucchange in R [14] identified a break in the level of the difference series, shown by a segmented gray line. ”Raw” plus the segmented line equals the ”Adjusted” Deniliquin series. The trend of the adjusted series matches the trend on the official Bureau of Meteorology High Quality series (”BoM HQ”) for Deniliquin.

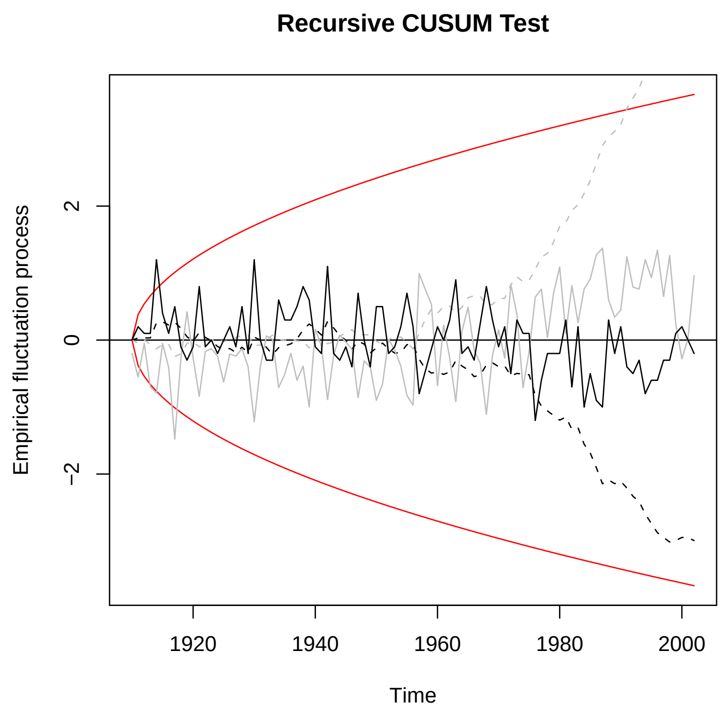

Figure 2: Robust test of a change in level of the raw temperature in Deniliquin (black) and its difference series (gray) using an empirical fluctuation process (dashed lines). The EFP of the difference series exceeds the 99% significance level (red) but the EFP of the raw data does not.

of borderline significance become significant. As we have shown above, and demonstrate in the following figures, the homogenization methodology would find breaks in any series with a trend that deviates sufficiently from the regional average. Such breaks are not true breaks; they are false positives or type I errors.

In Figure 3 the trend in the "Raw" Deniliquin series was increased by 0.5C degrees per decade. As previously, subtracting the raw data from the Australian temperature average yielded a significant break, adjusting "Raw" down to match the Australian temperature reference series "BoM HQ". Figure 4 demonstrates that segmented lines cause stronger coercion of trends. The target series matches the trend of the reference series if both breaks and trends are used.

3 Discussion

Unlike the policy arena where the type II error (false negative) is often a concern, scientific protocols demand type I error be reduced below 5% or preferably less than 1%.

The false alarm rate (or FAR) in a typical temperature network (Case 4, two random change points in all series, Table 4 in [13]) for the pairwise comparison and reference methods was 8.5% and 46.0% respectively. If a series has two breaks on average, this entails a type I error rate of 12% and 70% respectively. These error rates greatly exceed the acceptable rate in science and should severely limit the use of adjusted series in studies of observed temperature change. Methods should be rejected as 'not-fit-for-purpose' if the FAR exceeds conventionally acceptable levels of error.

High type I error is indicative of circular reasoning and 'data peeking'. Peeking at the regional standard to achieve a more powerful test may seem harmless, but is highly inappropriate. The only valid sources of information for adjusting a series are wholly independent of the regional average: the site metadata, significant breaks in the series itself (SNHT), or comparison with a small number of near neighbors not in the final network. Selection of reference stations by a high correlation (even though they are 1000's of kilometers from the target series) is also a variety of circularity [8].

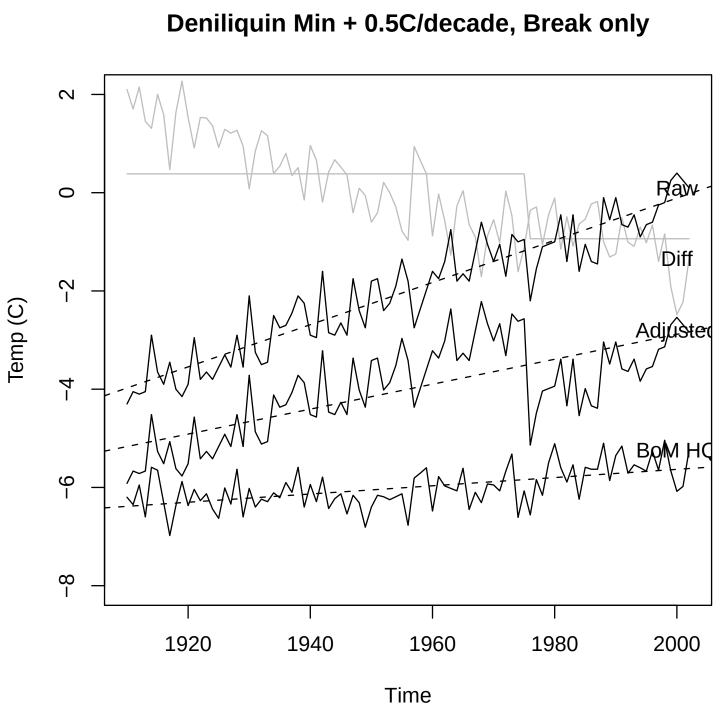

Figure 3: Homogenization of Deniliquin with an artificial warming trend of 0.5C per decade. The break down in the level of the difference series coerces the "Adjusted" Deniliquin series towards the trend of the Australian mean temperature used as the reference series ("Bom HQ").

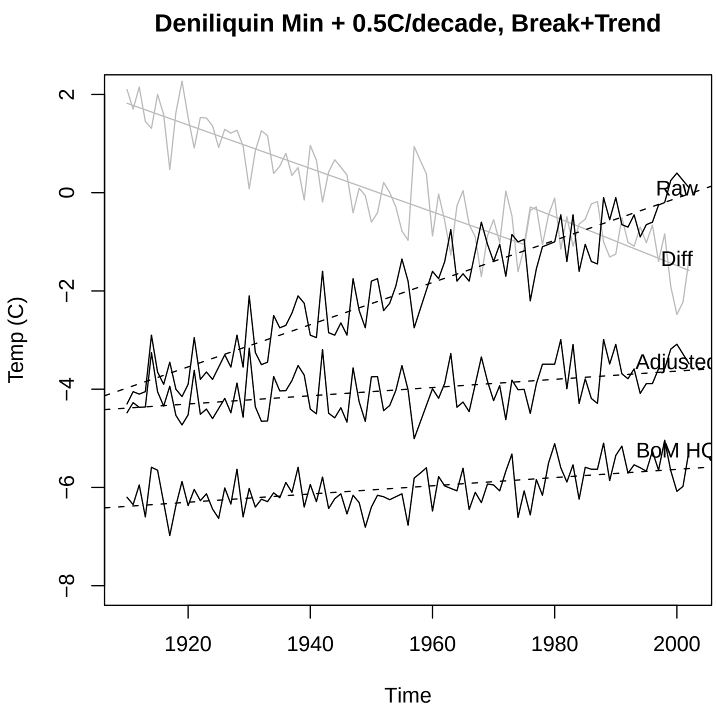

Figure 4: Homogenization of Deniliquin with an augmented warming trend of 0.5C per decade with the difference series fit to a segmented line. The adjustments coerce the Deniliquin series to the trend of the reference series ("Bom HQ") perfectly.

4 Acknowledgements

Reviews of drafts by Anthony Cox, Geoff Sherrington and Ken Stewart are gratefully acknowledged. Data and code for the study in R are available on the author's website http://landshape.org/enm.

References

[1] D.R.B. Stockwell and Kenneth Stewart. Biases in the australian high quality temperature network. *Energy & Environment*, in press, 2012.

[2] Anon. Statistical audit of the niwa 7-station review. Technical report, The New Zealand Climate Science Coalition, 2011.

[3] E. Steirou and D. Koutsoyiannis. Investigation of methods for hydroclimatic data homogenization. In A. Abbasi and N. Giesen, editors, *EGU General Assembly Conference Abstracts*, volume 14 of *EGU General Assembly Conference Abstracts*, page 956, April 2012.

[4] David A. Jones, William Wang, and Robert Fawcett. High-quality spatial climate data-sets for australia. *Australian Meteorological and Oceanographic Journal*, 58:233–248, 2009.

[5] Anon. On the sensitivity of australian temperature variability and trends to analysis methods and observation networks. Technical report, The Centre for Australian Weather and Climate Research, July 2011.

[6] Leopold Haimberger, C. Tavolato, and S. Sperka. Toward elimination of the warm bias in historic radiosonde temperature records—some new results from a comprehensive intercomparison of upper-air data. *J. Climate*, 21:4587–4606, 2008.

[7] David R.B. Stockwell. Guest Editorial: Biased Towards Extinction. *CO2 Science*, 7(19), 2004.

[8] D.R.B. Stockwell. Reconstruction of past climate using series with red noise. *Australian Institute of Geologists News*, 8:314, 2006.

[9] David R B Stockwell. Recent climate observations: disagreement with projections. *Energy & Environment*, 20(4):2, 2009.

[10] D.R.B. Stockwell. Critique of Drought Models in the Australian Drought Exceptional Circumstances Report (DECR). *Energy & Environment*, 21(5):425–436, 2010.

[11] Achim Zeileis and Christian Kleiber. Validating multiple structural change models : a case study. Report 12, Institut f{ü}r Statistik und Mathematik, 2004.

[12] D. A. Rhoades and M. J. Salinger. Adjustment of temperature and rainfall records for site changes. *International Journal of Climatology*, 13(8):899–913, December 1993.

[13] Matthew J. Menne and Claude N. Williams. Homogenization of Temperature Series via Pairwise Comparisons. *Journal of Climate*, 22(7):1700–1717, April 2009.

[14] Achim Zeileis, Friedrich Leisch, Kurt Hornik, and Christian Kleiber. strucchange: An R package for testing for structural change in linear regression models, 2001.