Reassessing changes in diurnal temperature range: A new data set and characterization of data biases

Reassessing changes in Diurnal Temperature Range: A new dataset and characterization of data biases.

P. W. Thorne[1,2], M. J. Menne[3], C. N. Williams[3], J. J. Rennie[4], J. H. Lawrimore[3], R. S. Vose[3], T. C. Peterson[5], I. Durre[3], R. Davy[2], I. Ezau[2], A. M. G. Klein-Tank[6], A. Merlone[7]

[1] National University of Ireland Maynooth, Maynooth, Ireland
[2] Nansen Environmental and Remote Sensing Center, Bergen, Norway
[3] NOAA's National Center for Environmental Information - Asheville, Asheville, NC, USA
[4] Cooperative Institute for Climate and Satellites, Asheville, NC
[5] Asheville, NC
[6] KNMI, De Bilt, Netherlands
[7] Istituto Nazionale di Ricerca Metrologica (INRiM), Torino, Italy

Corresponding author:
Peter Thorne
National University of Ireland Maynooth
Maynooth
Ireland
peter@peter-thorne.net

Key points

- Breakpoints are found to be more prevalent in DTR than other elements
- DTR has decreased since the early 20th Century but decrease is not linear
- Effects of homogenization change many details of global and regional DTR

Abstract

It is almost a decade since changes in Diurnal Temperature Range (DTR) globally have been explicitly assessed in a stand-alone data analysis. The present study takes advantage of substantively improved basic data holdings arising from the International Surface Temperature Initiative's databank effort and applies the National Climatic Data Center's automated pairwise homogeneity assessment algorithm to reassess global and regional DTR records. It is found that breakpoints are more prevalent in DTR series than other temperature elements and that the resulting adjustments have a broader distribution. This strongly implies that there is an over-arching tendency across the global meteorological networks for non-climatic artifacts to impart either random or anti-correlated rather than correlated biases in maximum and minimum temperature series. Future homogenization efforts would likely benefit from a consideration of DTR, maximum and minimum, in addition to average temperatures. Estimates of change in DTR are relatively insensitive to whether adjustments are calculated directly or inferred from adjustments returned for the maximum and minimum temperature series. The homogenized series exhibit a reduction in DTR since the early 20th Century globally. Adjustments serve to roughly half the magnitude of the long-term global reduction in DTR in the basic 'raw' data. Most of the estimated reduction in globally-averaged DTR occurred over 1960-1980. In several regions DTR has apparently increased since the 1990s, whilst globally it has exhibited very little change. Estimated changes in DTR are an order of magnitude smaller than in maximum and minimum temperatures, which have both been increasing rapidly on multi-decadal timescales.

1. Introduction

Diurnal Temperature Range (DTR) is defined as the daily maximum (*T*x) minus the daily minimum (*T*n) temperature. Herein consideration of DTR is restricted to land regions where DTR is far more dynamic than over the oceans. Over land areas DTR varies enormously both seasonally and geographically [Wang and Dillon, 2014]. The nature of DTR variability is important from a theoretical perspective for myriad reasons including in understanding microclimate impacts and the nature of changes within the deeper boundary layer [e.g. Christy et al., 2009, Pielke and Matsui, 2005, Zhou and Ren, 2011, Parker, 2006, Steenveld et al., 2011, McNider et al., 2012], and potentially as a determinant between forcings that have different Short Wave and Long Wave radiative fingerprints but may otherwise be similar [e.g. Jackson and Forster, 2013; Wang and Dickinson, 2013]. Trends and variability in DTR also have important practical implications for human health [Paaijmans et al., 2010], ecology [Peng et al., 2013, Vasseur et al., 2014], and agriculture [Battisti et al., 2009] amongst others.

Meteorlogical records have been undertaken at observing stations that extend back to the late 18th Century regionally and to the late 19th Century quasi-globally [Rennie et al., 2014]. Efforts have been made for at least three quarters of a Century [Callendar, 1938, Hawkins and Jones, 2013] to collate these data, apply homogeneity assessments and ascertain the nature of changes in Land Surface Air Temperatures (LSAT) over the globe. Today, there exist several such datasets globally [Lawrimore et al., 2011 (see also Williams et al., 2012a,b, 2013), Jones et al., 2012, Rohde et al., 2013] and regionally [e.g. Bohm et al., 2010; Tietavainen et al., 2010, Li et al., 2010, Jain and Kumar, 2012, Trewin, 2012, Vincent et al., 2012, Falvey and Garreaud, 2009, Christy et al., 2009; Van der Schrier et al., 2013]. Many of these analyses have been limited to a consideration of changes in average temperatures (*T*m), in part because records for average temperatures are more complete (Figure 1). Proportionately the effect becomes substantial prior to about 1950 and critical prior to 1895 (Figure 1 lower panel). Most US series post-1895 have been digitized to include *T*x and *T*n elements as part of the Climate Database Modernization Program. Elsewhere the situation is substantially more mixed and depends upon the data source.

Although DTR has been discussed as part of more general analyses globally [Rohde et al., 2012, Donat et al., 2013] and regionally [e.g. Makowski et al., 2008, Sen Roy and Balling, 2005, Christy et al., 2009, Zhou and Ren, 2011], it is almost a decade since the last stand-alone comprehensive analysis of global DTR data and its homogeneity was produced [Vose et al., 2005] and over twenty years since the first such assessment [Karl et al., 1993]. The IPCC in the most recent working group 1 assessment [Hartmann et al., 2013] noted that there was only '*medium confidence*' (see Mastandrea et al. [2010] for an interpretation of the specific meaning of this term in an IPCC context) in available records of observed changes in DTR due to the presence of a number of unresolved issues raised in the literature [Fall et al., 2011, Williams et al., 2012c, Christy et al., 2009] and the lack of recent studies and analyses.

In the last decade substantial progress has been made in:

- Creating better more complete records of daily data holdings of *T*x and *T*n with better provenance and quality control [Menne et al., 2012];
- In combining disparate global holdings of monthly records with the improved daily holdings to provide a more robust data basis from which to undertake analyses of long-term LSAT changes [Rennie et al., 2014]; and
- The creation of automated monthly climatic timeseries homogeneity assessment methods and their performance benchmarking and assessment [Venema et al., 2012, Williams et al., 2012c, Menne and Williams, 2009].

This paper aims to take advantage of these methodological and data innovations to create a new estimate of long-term changes in DTR globally and regionally. A subsequent companion paper compares these results to a broad range of other observationally based estimates [Thorne et al. submitted]. These subsequent analyses permit an assessment of sensitivity to both structural and parametric uncertainties [Thorne et al., 2005] in DTR estimation. A holistic assessment of DTR and its changes is stayed to the companion piece. This paper focuses instead upon the effects of the Pairwise Homogenization Algorithm (PHA) technique upon the data and a characterization of the resulting series and a consideration of implications for trends in *T*x and *T*n.

The remainder of the paper is structured as follows. In section 2 the data and homogenization methods employed in this study are briefly introduced. Section 3 summarizes the impacts of running the PHA algorithm on the data and discusses potential implications for the nature of non-climatic artifacts in the record. Section 4 describes the spatial and temporal evolution of the homogenized series for the spatially incomplete global mean and a subset of regions for which data are complete enough to analyze back to 1901 (Europe, N. America and Australia). Section 5 provides a brief discussion. Section 6 contains details on the dataset availability and Section 7 concludes.

2. Data and homogeneity assessment method

2.1 Source data

The present analysis is exclusively based upon the version 1 'recommended merge' release of the Global Land Surface Databank [Rennie et al., 2014] at monthly data resolution. This databank release is a result of efforts by many international collaborators under the auspices of the International Surface Temperature Initiative [Thorne et al., 2011]. It has combined holdings from over 50 constituent sources ranging from single stations to holdings of many thousand stations. These sources have been merged hierarchically with merge decisions based upon both metadata and data similarity metrics. Sources with *T*x and *T*n and better provenance and believed to be closer to the original recorded 'raw' basic data have been prioritized. The merge creates a single unique version per station that is as long as possible while minimizing potential discontinuities through false imputation of short period data. In total this version consists of just over 32,000 stations, most of which have *T*x and *T*n series for at least part of their records and many of which extend over at least 100 years (although not necessarily continuously).

The processing of the databank series merged *T*x and *T*n series stations first and only then went back to look for record segments for which solely Tm records exist. Despite this deliberate effort to maximize the amount of *T*x and *T*n data pull-through, availability for these elements is always lower than for *T*m (Figure 1). It is all but certain that *T*x and *T*n data, or at least observations at intervals over the day, were associated with the original records for which in the digital archives now only *T*m data exist in most cases. These data have either been lost or more likely were never digitized. This attests to the real importance of data rescue efforts, even for those stations which nominally already have records but for which the records are incomplete in important aspects such as availability of daily summaries which serve to inhibit understanding [e.g. Allan et al., 2011].

To facilitate the analysis herein a fourth field – *T*dtr – the difference between *T*x and *T*n has also been calculated and analyzed. In addition, for those analyzes of homogenization performance (Section 3) which include recourse to results for *T*m these consider solely *T*m values derived directly from the *T*x and *T*n elements as their average. This avoids conflation of data completeness and data characteristics in the analysis, which would otherwise ensue from use of the more temporally complete merged *T*m series (Figure 1). In many cases for remaining *T*m reports the archived *T*m may not result simply from averaging *T*x and *T*n. For example in at least Australia (and perhaps many other regions) in recent years the monthly average reported in CLIMAT messages is the average of hourly reports. Regardless, given that PHA is a neighbor-based procedure it is important to have the same networks for each element to perform a fair comparison and evaluation.

Both the DTR and the *T*m fields result from direct calculation from the monthly mean *T*x and *T*n series. So, the basic data used herein are internally consistent in that in the data presented to the homogenization algorithm DTR will always be the difference between *T*x and *T*n, *T*m will always be their average, and these elements are only ever calculated when both *T*x and *T*n are present. However, for months where either *T*x and / or *T*n have missing daily values this is not going to be equivalent to the average of the calculable daily DTRs (or *T*m's) within the month. While a more restrictive criteria of calculation of these values from the dailies could be applied to the subset of the databank arising from daily sources [Rennie et al., 2014] it would result in considerably fewer candidate station records, particularly prior to the 1950s. This comes at a potential cost regarding the monthly statistical mean and / or standard deviation characteristics for those stations where data is patchy on an intra-month basis due either to frequent missing days or frequent quality control flagging on the daily reports.

2.2 Pairwise Homogeneity Assessment

The data are presented to the exact same processing suite as those for Global Historical Climatology Network Monthly (GHCN, currently GHCN-Mv3.2.0) [Lawrimore et al., 2011, Williams et al. 2012a,b, 2013]. This consists of a set of quality control checks followed by application of a Pairwise Homogeneity Assessment (PHA) breakpoint identification and adjustment procedure [Menne and Williams, 2009]. The interested reader is directed to these papers for a fuller exposition of the methodology than is possible here if technical details are required.

The data are submitted separately for each of the four data streams considered (*T*x, *T*n, *T*m and DTR). No attempt is made herein to consider these data jointly to ensure consistency in returned adjustments across the elements when assessing homogeneity of the series, although such an approach is being actively developed for new versions of GHCNM. This is likely to yield inconsistencies at the station level between elements herein, which may occasionally be substantial (Section 3). The PHA algorithm analyzes timeseries of pairwise differences between nearby stations. It uses a Standardized Normal Homogeneity Test (SNHT) test statistic [Alexandersson, 1986] which is a t-test class of test, to identify potential discontinuities in each station pair. After doing so for all identified neighbor combinations the very large matrix of potential breakpoints is decomposed such that breakpoints are assigned iteratively to those stations in which they arise concurrently across multiple inter-comparisons with the resulting counts reduced accordingly until no further plausible breakpoint candidates exist. Then adjustments are inferred for the resulting population of identified candidate real breakpoints through comparisons to apparently homogeneous neighbor segments and applied if the distribution of returned adjustment estimates is substantively non-zero. The process is run solely once and the resulting set of applied and rejected adjustments are returned. The stations have been adjusted based upon the adjustment estimates and quality control decisions returned by the PHA in its operational version settings. The ensemble analysis of Williams et al. [2012c] highlights potential impacts from giving different, plausible, parameter settings to a number of the uncertain parameters within the PHA algorithm. For the present analysis consideration of such ensembles is deemed beyond scope.

2.3 Station gridding

For subsequent analysis only stations and months with sufficient data to create a 1971-2000 climatology under a Climate Anomaly Method have been retained in the gridded fields. As is discussed in the accompanying paper [Thorne et al., submitted] this is one of several possible approaches to gridding. For each station and calendar month, the minimum data requirement for calculating a climatology is 2/3 of data in the 30 year period taken as a whole and at least 1/2 in each decade (1971-1980, 1981-1990 and 1991-2000). This implies that a climatology may have been computed for some, but not all, calendar months at a particular station. For example, if the station's operator always took a vacation in July, then an insufficient amount of data may have been available for July while data for the other months of the year were sufficiently complete. In practice stations tend to be either substantively complete over the climatology period or have a marked data paucity that precludes their inclusion, meaning this affect is relatively minor in the retained station set. Stations for which a climatology can be calculated for any month tend to have climatologies for all twelve calendar months.

The climatology value has only been calculated with a trimmed mean based upon solely months within 3 standard deviations (σ) of the climatology period data mean for the given calendar month. An additional simple 5σ anomaly QC check has then been applied to the resulting anomaly series on a calendar month basis to remove gross outliers. Data between 3 and 5 standard deviations are retained but do not inform the climatological estimate. In stations with a strong secular trend this quality control step may remove real points far away in time from the climatology period. A high critical threshold of 5σ was chosen to mitigate this risk while still ensuring grossly questionable data did not get gridded. The check removes solely a handful of grossly questionable data points.

Resulting anomalies have simply been gridded, without any further weighting, into bins of 5 degrees latitude by 5 degrees longitude. Data have been gridded for all *T*x, *T*n and DTR for both the raw and adjusted series. Gridded *T*m series are not considered herein but will be documented in forthcoming GHCNM analyses instead.

For DTR it is possible to estimate the adjustments and resulting gridded series both directly from applying PHA to the timeseries and indirectly, through applying the net effect of the returned adjustments to *T*x and *T*n. The latter approach will yield a set of physically consistent estimates by construction, but at a potential cost if it misses breaks more amenable to identification and / or adjustment in DTR. Regardless, differences arising between 'directly adjusted' and 'indirectly adjusted' series provide some indication of likely uncertainties / sensitivities of the resulting analyses using the PHA method. However, these are very much an incomplete indicator of the likely true uncertainties. Comparisons to other estimates, constructed using distinct methods for all processing choices including quality control, adjustment, climatology calculation and gridding, will likely give a more realistic assessment of the true magnitude of the uncertainties in DTR estimates and are discussed further in the accompanying paper [Thorne et al., submitted].

3. Analysis of homogeneity adjustments

3.1 A consideration of the potential structure and magnitude of breakpoints

The four sets of series submitted to PHA consist of the two primary elements (*T*x and *T*n), their average (*T*m), and their difference (DTR). To ascertain the possible effects of the different data artifact characteristics on breakpoint magnitudes and distributions all possible combinations of *T*x and *T*n breakpoints between -5 and 5 K have been considered in Figure 2. By construction breakpoints in *T*m are always smaller than the break in either *T*x or *T*n except in the special case where the breaks in both elements are identical in sign and magnitude (perfectly correlated). Because DTR is the difference between the two elements there is no such cancellation in breakpoints of DTR and absolute breakpoint magnitudes reach 10K at [-5K, 5K] and [5K, -5K]. Hence DTR has twice as large a potential breakpoint magnitude for combinations explored as any of the other elements. By construction breakpoint magnitudes in DTR and *T*m are orthogonal. In the limit of perfectly correlated breakpoints in *T*x and *T*n (*T*x break = *T*n break) there will be no breakpoints in DTR. Similarly for perfectly anti-correlated breakpoints (*T*x break = -*T*n break) there will be no breakpoints in *T*m.

In cases where the breakpoints in *T*x and *T*n are correlated (both of the same sign) one or other of the breakpoints in *T*x and *T*n will always be the largest breakpoint. Where the breakpoints in *T*x and *T*n are anti-correlated (one positive, one negative) the largest breakpoint will always be in DTR. Restricting to a consideration of solely *T*m and DTR, the breakpoint in DTR will be largest both when the breakpoints in *T*x and *T*n are anti-correlated, and when they are only weakly correlated (same sign but substantially distinct magnitude whereby the difference is greater than their mean).

Assuming that the inter-station noise arising from random effects and real physical effects is similar across the elements such that Signal-to-Noise Ratios (SNRs) are similar in all resulting pairwise comparisons for breakpoint detection (Section 2.2) there is therefore a set of *a priori* expectations that can be made:

1. If the breakpoints in *T*x and *T*n are entirely randomly distributed and not conditionally dependent such that the break in *T*x has no *a priori* distributional basis given a break in *T*n, then it would be expected that there would be more and larger breaks in DTR than in *T*x or *T*n and fewest in *T*m.
2. If the breaks in *T*x and *T*n are conditionally dependent such that if the break in *T*n is positive it is more likely that *T*x is also positive and vice-versa then most and larger breakpoints would be expected to be found in *T*x and *T*n with fewest in DTR or *T*m (depending upon whether the conditioning was weak (*T*m) or strong (DTR))
3. If the breaks in *T*x and *T*n are conditionally independent such that a negative break in *T*n has a tendency to lead to a positive break in *T*x and vice-versa then it would be expected that most breaks would be found in DTR and they would be substantially larger than in *T*x and *T*n with fewer, much smaller breaks in *T*m

3.2. Analysis of returned breakpoint adjustments from the PHA algorithm

The PHA algorithm (Section 2.2) was run on the subset of stations which had sufficiently long records and for which sufficient neighbor estimates existed. The data masks are exactly equivalent for *T*m and DTR as they require *T*x and *T*n to both be available (Section 2.1). For *T*x and *T*n some additional data exists for some stations. However, to a first approximation the number of stations and record length are equivalent for all four elements presented to PHA. Despite this similarity in input data availability there exist marked differences in the estimated frequency, magnitude and distribution of adjustments returned across the 4 elements (Figure 3). There are more adjustments returned for DTR (66,572) than for *T*n (62,013), for which there are more again than for both *T*x (51,777) and *T*m (50,378). The standard deviation of the returned adjustment estimates is largest for DTR (1.24K), roughly equivalent for *T*x (0.98K) and *T*n (1.00K), and smallest for *T*m (0.75K). There is no obvious substantial departure for any element from Gaussian distributional assumptions. In all cases there is a 'missing middle' of undetectable / unadjustable real-world breakpoints that must in reality exist.

Following from Section 3.1 if there is no difference in effective power of PHA to detect and adjust for breaks between elements then the implication is that the breakpoints in *T*x and *T*n are either entirely random or conditionally independent. However, there are also reasons why DTR may be expected to exhibit lower noise as it is the difference between two variables, *T*x and *T*n, which tend to co-vary on monthly timescales. If the noise in the pairwise station comparators, which form the basis for the breakpoint statistical assessment, was lower then it may simply be that PHA can more efficiently detect smaller breakpoints from the 'missing middle' clearly evident in all panels of Figure 3. It is obvious given the broader distribution of DTR adjustments from Figure 3 that the increased number of breakpoints found and adjusted in DTR results from larger discontinuities rather than any difference in efficacy of breakpoint identification.

The breakpoint behavior can be further investigated by consideration of directly inferred and indirectly inferred adjustment estimates for DTR and *T*m (Figures 4 and 5). Breaks in the derived variables would be expected to be coincident in timing and resulting magnitude with those estimated from the *T*x and *T*n analyses. Comparing direct and indirect adjustment estimates therefore provides a check on internal consistency of results. The direct and indirect adjustment estimates should be correlated and show no overall offset from one another. Scatter would be expected to arise due to variations in breakpoint date assignments and neighbor segments used to adjust. The degree of scatter provides some indication of the probable uncertainty in the resulting station series estimates.

For DTR these comparisons exhibit substantial scatter, even when a collocation error of 12 months in the breakpoint locations found is allowed for (Figure 4 left hand panel). There are many cases where either a DTR adjustment is made without a corresponding adjustment to either *T*x or *T*n and vice-versa (points along either $y=0$ $x\neq0$ or $x=0$ $y\neq0$ respectively). In numerous cases the adjustments differ in sign (top left and lower right quadrants). Overall, however, there is a tendency to broadly agree with the cloud of points scattered around the 1:1 line rather than entirely randomly. The histogram of adjustment comparators (Figure 4 right hand panel) is zero mean and broadly Gaussian, albeit with a large sigma such that almost

23% of differences exceed 1K in magnitude. A similar analysis of *T*m (Figure 5) exhibits far less scatter between directly and indirectly inferred adjustments (left hand panel, points lie much closer to the 1:1 line) with only just under 5% of differences exceeding 1K in magnitude (right hand panel).

Both direct and indirect adjustments to DTR act to reduce the apparent spread in individual station linear trend fit estimates over 1901-2012 and 1951-2012 (Figure 6). This is consistent with what would be expected if reasonable adjustments were being applied to data containing inhomogeneities. Individual station series in the basic data contain systematic data errors. Such systematic effects are equivalent to adding units of red noise to the time-series, causing artificial dispersion in the distribution of long-term station series behavior. Figure 6 suggests that many such systematic biases are being effectively removed in a reasonable manner by the PHA algorithm.

3.3 Synthesis of adjustments analysis

Breakpoints are more easily discoverable using PHA in DTR than they are in *T*x or *T*n which in turn are somewhat more discoverable than in *T*m. Earlier analyses over the European domain [Wijngaard et al., 2003] and globally using HadISD [Dunn et al., 2014] found similarly that breakpoints in DTR were somewhat more amenable to detection. Not only were more breakpoints found in DTR but they were on average larger and had a broader standard deviation than other elements. When calculated directly from DTR or indirectly from *T*x and *T*n adjustments, individual adjustment estimates show similar behavior but with substantial dispersion. Therefore care should be taken in interpretation of individual adjusted station DTR series. However, the overall distribution of station trend estimates is less dispersive following application of adjustments with many obviously questionably large station trends removed. Taken as a whole this analysis provides confidence in the efficacy of PHA when applied to DTR series at least at regional or global scales.

Overall, results from PHA strongly imply that breakpoints in *T*x and *T*n are either randomly distributed or conditionally independent. Strong conditional dependence whereby *T*x and *T*n breakpoints are almost always of the same sign and similar magnitude can be ruled out by the present analysis. Reasons and implications are returned to in the discussion (Section 5).

4. Analysis of gridded fields and regional averages

4.1 Data completeness

As with most preceding analyses of DTR [e.g. Vose et al., 2005] data is globally incomplete and the data density in those areas sampled varies over at least two orders of magnitude. Figure 7 shows gridbox DTR station data counts for the month when data density is globally maximal (October 1987). Sampling is dense over much of Australia, China and Japan, Europe and in particular North America.

Sampling is particularly poor (or even non-existent) over much of Africa, SE Asia, the Arabian Peninsula, the Amazon basin and the ice sheets of Antarctica and Greenland. Sampling varies substantively through time both globally and regionally in those regions with records that extend back to the early 20th Century (Figure 8). Outside North America there exists a step-change in availability in 1960 with far fewer stations prior to this. As a result trends and variability in DTR for analyses across 1960 may be an artifact of coverage changes rather than true changes. As discussed further in Section 2.1 there likely exist records which if rescued digitized and shared could mitigate this issue.

4.2 Diurnal Temperature Range

Herein analysis is made of changes in DTR from the original 'raw' data records and following adjustments calculated directly and indirectly from applying the adjustments returned to *T*x and *T*n and then calculating DTR from these series as outlined in Section 2.3. The analysis starts with spatial patterns of trends over increasingly shorter periods to present. Recourse is then made to regionally averaged timeseries behavior and linear trend estimates.

4.2.1 Spatial trends

Trends calculated since the beginning of the 20th Century greatly reduce coverage if a data completeness mask is applied to ensure early and late period data availability in addition to total timeseries completeness (Figure 9 c.f. Figure 7). Data remain only for N. America, Europe, parts of Australia, E. China and Japan and a handful of dispersed additional locations. The spatial domains sampled in Figure 9 govern the designation of sub-domains considered in subsequent regional analyses and denoted henceforth by geographic shorthand as: N. America (45W-135W, 25-60N); Europe (10W-60E, 25-60N); and Australia (110E-155E, 10S-45S). The cluster over Japan and E. China is deemed too small to calculate a reasonable regional average.

Century timescale trends in DTR (Figure 9) are of the order 0.1K/decade at most across the sampled gridboxes in the raw data and in the two adjusted products. Trends are significant at the gridbox level in many of the gridboxes sampled in the input data, but this decreases substantially following application of adjustments either using the direct or the indirect approach. In the input data most gridboxes exhibit a reduction in DTR over time. Although a majority of gridboxes still indicate a reduction in DTR following the application of adjustments, the magnitude of the DTR reduction is far less significant. Adjustments change the sign of the DTR trends in much of the South Western / Western United States from negative to positive and reduce the negative trends elsewhere in N. America. This change is more marked when adjustments are calculated indirectly than when they are calculated directly. There are less spatially consistent changes in remaining regions with many gridboxes experiencing large changes including changing the sign of the DTR trend.

Starting in 1951 as expected from Figure 8, spatial sampling is much more complete although Africa, the Indian sub-continent and S. America remain substantively incompletely sampled in addition to Greenland and Antarctica (Figure 10). Over this 62 year period in the input data records the vast majority of gridboxes exhibit substantial reductions in DTR that are particularly marked over much of Asia and N. America. Application of adjustments substantively changes the trend behavior over N. America where trends are reduced with a sign change in many gridboxes west of the Rockies to an increasing DTR and very few gridbox series remain significant. In Southern Europe adjustments indicate small increases in DTR. Overall, adjusted series are visually somewhat more spatially homogeneous than the input data trends lending some support to the findings detailed in Section 3 regarding the efficacy of the PHA when applied either directly or indirectly to DTR records.

The last period for which geographical trends are considered is from 1979, a start date typically used in climate studies because it is the advent of regular polar-orbiter satellite measurements. Although the current analysis is in-situ only it is still potentially informative to other studies to document changes over this period (Figure 11). Over this period sampling is more complete again, particularly so over South America although large areas remain data void. Since 1979 trends are substantively larger in magnitude and of more mixed sign. That trends over shorter periods are larger, more spatially heterogeneous, and of mixed sign is to be expected as shorter periods increasingly reflect decadal-scale regional variability [Santer et al., 2011]. Over this shorter period, the application of adjustments leads to large changes in apparent sign and magnitude of DTR trends in many regions. This is particularly marked in the United States, in parts of Europe and over much of China and SE Asia.

Over the United States the adjustments in the post-1979 era lead to a change from a slight reduction in DTR to a larger increase in many gridboxes. The adjusted DTR increases are significant in several gridboxes in the South Western states. This adjustment is consistent with understanding of the transition from Cotton Region Shelters (CRS, termed Stevenson Screens elsewhere) to electronic Maximum Minimum Temperature Sensor (MMTS) starting in the 1980s and substantively completed by the late 1990s. In this change both the instrument and its shielding were changed substantively, often associated with a change in measurement location. This change affected roughly 70% of the COOP network, which is the backbone of the US records. Field based studies and statistical analyses have variously concluded that the CRS to MMTS transition led to a positive bias in Tn and a negative bias in Tx artificially reducing DTR in the raw data [Fall et al., 2011, Williams et al., 2012c and references therein]. Assuming that the PHA algorithm is adequate the effect of this change is larger than the underlying real-world DTR signal over much of the United States. The size of the effect found and adjusted for here is consistent in magnitude with understanding from various side-by-side comparisons under the assumption that c.70% of the network experienced the change.

In Europe adjustments lend support to the propensity for increased DTR in recent years [Vautard et al., 2009]. In China and SE Asia, although gridbox trends remain significant the reductions in DTR are generally less following adjustment than is implied by the raw data.

4.2.2 Regional and global timeseries and trends

As is visually obvious from Figures 9-11 linear trend estimates do not describe all facets of the timeseries behavior globally or regionally. Timeseries for global (Figure 12) and regional (Figure 13) DTR averages serve to highlight the presence of substantial interannual to multi-decadal variability in DTR even globally. In all cases these timeseries have been derived from averaging all available gridded data at each timestep using cos(lat) area weighting. As noted earlier, given the varying station count and gridbox availability care should be taken in interpretation in particular of pre-1960 data. The effects of different completeness inclusion criteria for this step are further discussed and analyzed in the accompanying paper [Thorne et al., submitted].

Following adjustments it is estimated that globally averaged DTR was elevated relative to present day until the late 1950s, declined by of the order 0.2C by the early 1980s and has then been relatively steady since according to both adjusted series considered. There are substantial differences between directly and indirectly adjusted series estimates prior to around 1950. Overall the adjusted series are more similar to each other than they are to the input data both in terms of the long-term trend and also decadal timescale variability. Globally adjustments have a substantial impact in the most recent period since 2000 when (semi-)automation has been prevalent across the global network as a whole (although some regions experienced this change 10-20 years earlier), and prior to the 1970s.

Global and regional average trends are substantively impacted by the PHA homogenization procedures. Adjusting either directly or indirectly the net effect is to reduce the magnitude of the apparent long-term trends in global DTR (Table 1). Nonetheless, trends towards globally reduced DTR are statistically significant over the period 1901 to 2012 and the shorter sub-period 1951 to 2012 for the 'raw' series and remain so for the adjusted series. Over the period 1979 to 2012 the global mean trend reverses from a significant reduction in the 'raw' data, to a slight increase in both of the adjusted series neither of which are statistically significant (c.f. Figure 11 and associated discussion).

In North America the adjustments reduce DTR prior to 1950 and increase DTR since the 1980s yielding a large reduction in the apparent narrowing of DTR implied by the basic 'raw' data (Figure 13, top panel). As discussed previously post-1980 changes are consistent with understanding of the effects of transition from CRS to MMTS across roughly 70% of the US observing network. Earlier period adjustments may relate either to the effects of changes in time of observation [Karl et al., 1986] or a propensity to relocate from city to airport locations. Trends over 1901-2012 are significantly negative in the basic 'raw' data and both adjusted series, but are halved in magnitude following adjustments. Over the two shorter periods considered neither adjusted series exhibits significant trend behavior. Estimates are slightly negative over 1951-2012 and slightly positive over 1979-2012 (Table 1). The two adjusted series are very similar to each other and very distinct from the basic 'raw' data behavior.

Over the European domain adjustments act to increase DTR both since the 1980s and prior to the 1950s (Figure 13, middle panel). This yields a marked change in multi-decadal variability in this region removing an apparent trend of increasing DTR in the first half of the twentieth Century in the basic 'raw' data. On the longest timescales this leads to an increased negative trend in DTR following adjustments, which is significant in both adjusted estimates but not the basic data (Table 1). Over 1951-2012 again all estimates are significantly negative. Since 1979 both adjusted series imply positive trends in DTR over the European domain taken as a whole but these are not statistically significant. As is the case globally and over N. America the adjusted series are much more similar to each other than they are to the basic 'raw' data.

Australian DTR series exhibit far greater variability than those over Europe and America (Figure 13, lower panel). Variability appears to be highly correlated with continental scale aridity / rainfall (and by extension ENSO). For example the very wet year of 2010/11 is associated with a marked negative DTR anomaly, consistent with basic theoretical understanding of partitioning of fluxes [Peterson et al., 2011]. The effect of the adjustments is more muted for this region with slight increases in DTR in the mid-20th Century and reductions in the early 20th Century. Trends are generally not significant in the adjusted series with the exception of indirectly adjusted series for 1901-2012 (Table 1) and confidence intervals are larger than for other regions considered reflecting the much greater year to year variability in the series. Over this region there is less obvious concordance between the adjusted series.

4.3 Maximum and minimum temperatures

For *T*x and *T*n only direct adjustments exist so analysis is limited to the raw and directly adjusted series. Trends over 1951-2012 for *T*x (Figure 14) and *T*n (Figure 15) both exhibit strong warming in the vast majority of the gridboxes that are sampled. Adjustments remove an apparent cooling in *T*x in the eastern United States consistent with the United States Historical Climatology Network (USHCN) [Menne et al., 2010] and our understanding of US biases arising from the CRS to MMTS transition. Cooling in *T*x in Southern China is also reduced and several obviously erroneous gridbiox series look more similar to surrounding series after homogenization. Adjustments to *T*n adjust several obviously erroneous gridbox trends and increase slightly the apparent warming in eastern North America but otherwise have little obvious effect at the gridbox scale.

Global average timeseries of *T*x and *T*n are strongly positive (Figure 16), particularly since the early 1970s. Adjustments serve to narrow the difference in trends (which is consistent with a reduction in the estimated rate of decrease in DTR in the preceding subsection). The overall effect of PHA adjustments is to increase the long-term trend in both *T*x and *T*n with the effect being larger for *T*x (although the *T*x trend is still smaller than that for *T*n, Table 2). Trends in *T*x and *T*n are highly significant over all three periods considered in the present analysis and, in the adjusted series, roughly an order of magnitude larger than DTR trends. Trends in *T*x and *T*n are consistent with GHCNv3.2.0 trends for Tm even though the station basis set differs substantially.

5. Discussion

The adjustments returned by the PHA algorithm strongly imply that breakpoints in *T*x and *T*n are either random or conditionally independent. Random breaks would mean that the break size and magnitude in *T*n on average had no influence upon the resulting break size and magnitude in *T*x. Conditionally independent would imply an overall tendency for *T*x and *T*n breakpoints to be of opposite sign such that they partially or completely cancel in the mean. This raises two interesting questions: first whether there are more optimal approaches to homogenization than analyzing *T*m as is commonly the case for global centennial timescale LSAT reconstructions to date; and second why, metrologically, the over-arching tendency may be so.

5.1 Future homogenization efforts considerations

Homogenization of surface meteorological station records is inherently a signal-to-noise problem. Small, relative to meteorological and climatological variability, breakpoints arising for myriad reasons must be found and then accurately quantified. Therefore it is important to search in an optimal direction. State of the art algorithms like PHA perform pairwise comparisons that act to remove common real-world variations between candidate nearby stations and leave a difference series that in the absence of any biases in the two comparators should behave as iid white noise arising from random measurement errors and real inter-site variability. The white noise places a hard lower limit on signal detectability. No break will be discoverable that is of comparable magnitude to the standard deviation of the series. Yet, small breaks arguably matter substantively because they are systematic effects that do not cancel, so methods should try to optimize breakpoint detectability and adjustments whilst simultaneously minimizing false alarm rates. All breakpoint algorithms return bivariate distributions (cf. Figure 3) that in reality are the two wings of the true Gaussian distribution of real-world breaks with breaks around zero not being found and / or adjusted for.

If the breakpoints in *T*x and *T*n were strongly conditionally dependent (similar sign and magnitude) then searching for breakpoints in *T*m would be quasi-optimal. The further towards conditional independence of *T*x and *T*n breakpoints the less optimal use of *T*m series to locate and adjust for breakpoints will become as the dominant direction of breaks becomes increasingly orthogonal to *T*m (Figure 3). Section 3 strongly implies breakpoints are at best random, if not conditionally independent. If the breakpoints are random then a search should be made in all four elements. If the breakpoints are mainly conditionally independent then consideration could be limited to DTR, *T*x and *T*n. Thus in future, homogenization procedures that search for breakpoints in *T*m, *T*x, *T*n and DTR simultaneously will very likely yield a more accurate and optimal set of breakpoint locations.

Finding the breakpoints is just the first part of the problem. The resulting adjustment estimates then need to be reconciled. Here, no such effort has been made and instead the difference between direct DTR and indirect DTR adjustments has been used to illustrate potential sensitivities. In future, efforts could be made given a set of 4 adjustment estimates (or better still conditional density functions of the adjustments) and a closure condition that the adjustments to *T*x and *T*n must average to the adjustment of *T*m and difference to the adjustment to DTR to form a combined set of adjustments. Such an approach is being pursued to develop future versions of GHCNM.

All of the above considerations are moot if the station series are only available as *T*m, as is the case for many of the stations in the current databank (Figure 1, lower panel). Therefore to optimize future analyses of surface temperature changes over land efforts should be made to recover *T*x and *T*n records for stations and periods of record for which currently only *T*m records exist in addition to rescuing that data for new stations to improve both coverage and station periods of record [Allan et al., 2011].

5.2 Why metrologically may breakpoints in *T*x and *T*n be random or conditionally independent?

All meteorological temperature measurements are undertaken by a proxy that is correlated with the target measurand be that the expansion of liquid, electrical resistance or some other means. Ideally, the calibration processes for thermometers would be defined by robust and well documented procedures, under highly controlled conditions, leading to a full evaluation and definition of calibration uncertainty components budgets and total values, according to the kind of sensors used and environments experienced.

Far from being in thermal adiabatic condition, a thermometer used to measure air temperature actually measures the mix of convective, radiative and contact heat transfers. All of these thermodynamic effects are difficult to be corrected with an uncertainty on the correction. Some devices permitting evaluation of the influence of such parameters on the sensors under calibration are being developed, but are still under experimental prototype status [Lopardo et al 2014, Merlone et al. 2014, Musacchio et al. 2014]. Moreover, since the calibration is performed in stable temperature conditions, while the measurement of daily air temperature fluctuations is anything but stable, sensor dynamics can introduce deviations due to the response inertia and delay, not evaluated during calibration. For example, the behavior of two different thermometers calibrated both in a climatic chamber and in a liquid bath, was compared to their performance in a Stevenson Screen (CRS) (Grykalowska, 2014). While both the controlled calibration methods resulted in consistency within uncertainty, when placed in the Stevenson Screen, the readings of the two thermometers differed by substantially more than the sum of their calibration uncertainties, demonstrating that hitherto unaccounted for sensor dynamics effects remained.

In the atmosphere there are two critical aspects: the response to heat transfer effects; and dynamic behavior in capturing temperature fluctuations. Having long established and recognized the difficulties in estimating the errors induced by these quantities of influence on the sensors there have been the attempts to reduce the effects through e.g. screens protecting from direct radiation on the sensing element, reduced contact surface with the supporting structure, models to minimize the convective effects, and ventilation to reduce extra heating due to stagnant air. The range of measurement, shielding and mounting techniques likely yields differing error characteristics across the meteorological networks, which further are likely to be climatically dependent.

In principle, three physical co-variates shall influence the temperature measurements: radiation, wind speed and humidity. In days with wind blowing and limited sun radiation these effects are expected to be of low amplitude regardless of instrument configuration whereas in days with sun, absence of wind and larger night-day temperature fluctuations the effects would be maximal. Such conditions amplify the possible differences in DTR recording arising from changes in instrumentation and practices through time.

There are two broad classes of instrumentation: artificially aspirated and non-aspirated. Artificially aspirated measurements exhibit substantially lower sensitivity to prevailing meteorological conditions so long as adequately screened from direct and indirect radiative effects. They may tend to read slightly high during daytime due to imperfect shielding from radiation or thermal contact and slightly low during nighttime due to cooling effects from condensation of the drawn air. Non-aspirated measures will exhibit substantially greater sensitivity to prevailing meteorological conditions. On average the measures may be warm biased for both *T*x and *T*n due to a mix of radiative and ventilation effects. The biases will be highly dependent upon configuration and site micro-environment. The change from CRS to MMTS (both non-aspirated but very distinct) had differential effects on *T*x and *T*n with *T*x decreasing and *T*n increasing. Changing from non-aspirated to aspirated measurements will tend to yield an apparent and spurious increase in DTR that is larger than any concurrent change in *T*m.

5.3 Caveats pertaining to use of current data products

For analyses of DTR using the dataset constructed herein, the effects of the changing station availability through time are potentially an insidious effect. The primary effects are two-fold. Firstly changing the neighbor constraint substantively through time will affect the efficacy of any homogenization algorithm and PHA is not immune to this. Secondly, the changing data mask may confound a clean interpretation of global and regional trends even if the data were perfect (which they are not). Care should be taken in interpreting pre-1960 records when the station mix changes substantively both globally and regionally.

6. Dataset availability

The dataset is made available through [website to be appended here once decided, can we host through NUIM?]. The following series shall be made available:

- Adjusted station series as CF-compliant netcdf files (one per station) containing several timeseries fields.
- Gridded raw and adjusted series for Tx, Tn and DTR (including indirectly adjusted) as CF-compliant netcdf files (a total of 7 files)

At this time there are no plans to update the series beyond 2012. Dataset users should cite this paper.

7. Conclusions

The present analysis has re-examined changes in DTR globally and regionally using improved holdings and NCDC's PHA algorithm. Adjustments to the basic 'raw' data have a non-negligible impact upon the resulting series behavior on multi-decadal timescales and are comparable in magnitude to the apparent trend in the basic 'raw' data globally and regionally. DTR is estimated to have decreased globally since the mid-twentieth Century but the adjustments reduce by half the trend compared to that in the basic 'raw' data. Both maximum and minimum temperatures have increased rapidly and changes in these elements are an order of magnitude greater than in DTR globally. Adjustments are more prevalent in DTR than in *T*x or *T*n, which in turn are more common than in *T*m. This implies that overall the biases in *T*x and *T*n are either random or conditionally independent and has potentially important implications for future homogenization strategies. It implies that searching for and adjusting breaks in average temperatures is likely to be sub-optimal as the signal to noise ratio will tend to be a minimum in average temperatures. Instead efforts that search in addition for breakpoints in DTR, *T*x ,and *T*n would likely be more efficient at finding and adjusting for non-climatic artifacts in the records.

Acknowledgements

We thank 2 NOAA NCEI internal reviewers for their insights. Fabio Bertiglia, provided useful input to Section 5.2. Details as to how to ascertain data and materials are given in Section 6 and can also be attained from the lead author.

References

Alexandersson, H., 1986: A homogeneity test applied to precipitation data. Journal of Climatology, 6, 661-675

Allan, R. J., et al., 2011: The International Atmospheric Circulation Reconstructions over Earth (ACRE) initiative. *Bull. Amer. Meteor. Soc.,* **92,** 1421–1425.

Battisti, D. S., and R. L. Naylor, 2009, Historical warnings of future food security with unprecedented seasonal heat. Science, 323, 240-244

Bohm, R., P. D. Jones, J. Hiebl, D. Frank, M. Brunetti, and M. Maugeri, 2010: The early instrumental warm-bias: A solution for long central European temperature series 1760–2007. *Clim. Change*, **101**, 41–67.

Callendar, G. S. (1938), The artificial production of carbon dioxide and its influence on temperature. Q.J.R. Meteorol. Soc., 64: 223–240. doi: 10.1002/qj.49706427503

Christy, J. R., W. B. Norris, and R. T. McNider, 2009: Surface temperature variations in East Africa and possible causes. *J. Clim.*, **22**, 3342–3356.

Donat, M. G., L. V. Alexander, H. Yang, I. Durre, R. Vose, and J. Caesar, 2013a: Global land-based datasets for monitoring climatic extremes. *Bull. Am. Meteor. Soc.*, **94**, 997-1006.

Falvey, M., and R. D. Garreaud, 2009: Regional cooling in a warming world: Recent temperature trends in the southeast Pacific and along the west coast of subtropical South America (1979–2006). *J. Geophys. Res. Atmos.*, **114,** D04102.

Grykalowska A. et al., The basics for definition of calibration procedure of temperature sensors for weather station Submitted to Meteorological Application as MMC2014 Proceedings

Hawkins, E., and P. D. Jones, 2013: On increasing global temperatures: 75 years after Callendar *Quarterly Journal of the Royal Meteorological Society*

Jackson, L. S. and P. M. Forster, 2013: Modeled rapid adjustments in diurnal temperature range response to CO_2 and solar forcings. *J. Geophys. Res.*, **118**, 2229-2240, doi: 10.1002/jgrd.50243

Jain, S. K., and V. Kumar, 2012: Trend analysis of rainfall and temperature data for India. *Curr. Sci.*, **102**, 37–49.

Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice, 2012: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *J. Geophys. Res. Atmos.*, **117,** D05127.

Karl, Thomas R., Philip D. Jones, Richard W. Knight, George Kukla, Neil Plummer, Vyacheslav Razuvayev, Kevin P. Gallo, Janette Lindseay, and Thomas C. Peterson, 1993: A new perspective on recent global warming: Asymmetric trends of daily maximum and minimum temperatures. *Bulletin of the American Meteorological Society,* **14**, 1007-1023.

Karl, T.R., C.N. Williams, Jr., P.J. Young, and W.M. Wendland, 1986: A model to estimate the time of observation bias associated with monthly mean maximum, minimum, and mean temperature for the United States, *J. Climate Appl. Meteor.*, **25**, 145-160.

Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie, 2011: An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *J. Geophys. Res. Atmos.*, **116**, D19121.

Li, Q., W. Dong, W. Li, X. Gao, P. Jones, J. Kennedy, and D. Parker, 2010: Assessment of the uncertainties in temperature change in China during the last century. *Chin. Sci. Bull.*, **55**, 1974–1982.

Lopardo G. et al. 2014. Traceability of Ground-Based Air-Temperature Measurements: A Case Study on the Meteorological Observatory of Moncalieri (Italy); International Journal of Thermophysics; available as 'Online First' on SpringerLink: http://link.springer.com/article/10.1007/s10765-014-1806-y

Makowski, K., M. Wild, and A. Ohmura, 2008: Diurnal temperature range over Europe between 1950 and 2005. *Atmos. Chem. Phys.*, **8**, 6483–6498.

Mastrandrea, M.D., C.B. Field, T.F. Stocker, O. Edenhofer, K.L. Ebi, D.J. Frame, H. Held, E. Kriegler, K.J. Mach, P.R. Matschoss, G.-K. Plattner, G.W. Yohe, and F.W. Zwiers, 2010: Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. Intergovernmental Panel on Climate Change (IPCC). Available at <http://www.ipcc.ch>

McNider, R. T., et al., 2012: Response and sensitivity of the nocturnal boundary layer over land to added longwave radiative forcing. *J. Geophys. Res.*, **117**, D14106.

Menne, M. J., and C. N. Williams, 2009: Homogenization of temperature series via pairwise comparisons. *J. Clim.*, **22**, 1700–1717.

Menne, M.J., C.N. Willaims, Jr., and M.A. Palecki, 2010: On the reliability of the U.S. surface temperature record. *J. Geophys. Res.* , doi:10.1029/2009JD013094

Menne, M.J., I. Durre, R.S. Vose, B.E. Gleason, and T.G. Houston, 2012: An overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, 29, 897-910, doi:10.1175/JTECH-D-11-00103.1.

Merlone A. et al. 2014, In situ calibration of meteorological sensor in Himalayan high mountain environment. Submitted to Meteorological Application as MMC2014 Prioceedings

Musacchio C. et al. METROLOGY ACTIVITIES IN NY-ÅLESUND (SVALBARD), Submitted to Meteorological Application as MMC2014 Prioceedings

Parker, D. E., 2006: A demonstration that large-scale warming is not urban. *J. Clim.*, **19**, 2882–2895.

Paaijmans, K. P., et al., 2009: Influence of climate on malaria transmission depends on daily temperature variation. PNAS, 107, 15135-15139

Peng, S. et al., 2013: Asymmetric effects of daytime and night-time warming on Northern Hemisphere Vegetation. *Nature*, **501**, 88-94, doi:10.1038/nature12434

Peterson, T. C., K. M. Willett, and P. W. Thorne, 2011: Observed changes in surface atmospheric energy over land. *Geophys. Res. Lett.*, **38,** L16707.

Pielke, R. A., and T. Matsui, 2005: Should light wind and windy nights have the same temperature trends at individual levels even if the boundary layer averaged heat content change is the same? *Geophys. Res. Lett.*, **32,** L21813.

Rennie, J. J. et al., 2014: The International Surface Temperature Initiative global land surface databank: monthly temperature data release description and methods. *Geoscience Data Journal*, doi: 10.1002/gdj3.8

Rohde, R., et al., 2012: A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinfor. Geostat.: An Overview*, **1**.

Rohde, R., et al., 2013: Berkeley Earth temperature averaging process. *Geoinfor Geostat: An Overview*, **1**.

Santer, B. D., et al., 2011: Separating signal and noise in atmospheric temperature changes: The importance of timescale. *J. Geophys. Res. Atmos.*, **116**, D22105.

Sen Roy, S., and R. C. Balling, 2005: Analysis of trends in maximum and minimum temperature, diurnal temperature range, and cloud cover over India. *Geophys. Res. Lett.*, **32,** L12702.

Steeneveld, G. J., A. A. M. Holtslag, R. T. McNider, and R. A. Pielke, 2011: Screen level temperature increase due to higher atmospheric carbon dioxide in calm and windy nights revisited. *J. Geophys. Res. Atmos.*, **116**.

Thorne, P. W., et al 2005: Uncertainties in climate trends - Lessons from upper-air temperature records. BAMS **86**(10): 1437+

Thorne, P. W. et al., 2011: "Guiding the Creation of a Comprehensive Surface Temperature Resource for 21st Century Climate Science.", Bulletin of the American Meteorological Society, doi: 10.1175/2011BAMS3124.1

Thorne et al. submitted

Tietavainen, H., H. Tuomenvirta, and A. Venalainen, 2010: Annual and seasonal mean temperatures in Finland during the last 160 years based on gridded temperature data. *Int. J. Climatol.*, **30**, 2247–2256.

Trewin, B., 2012: A daily homogenized temperature data set for Australia. *Int. J. Climatol.*, 33, 1510-1529.

van der Schrier, G., E. J. M. van den Besselaar, A. M. G. Klein Tank, and G. Verver (2013), Monitoring European average temperature based on the E-OBS gridded data set, J. Geophys. Res. Atmos., 118, 5120–5135, doi:10.1002/jgrd.50444.

Vasseur, D. A., et al ., 2014, Increased temperature variation poses a greater risk to species than climate warming. *Proc. Roy. Soc. B.*, 281, 20132612

Vautard, R., P. Yiou, and G.J. Van Oldenborgh, 2009: Decline of fog, mist and haze in Europe over the past 30 years. *Nature Geoscience*, 2, 115-119, doi:10.1038/ngeo414

Venema, V. K. C., et al., 2012: Benchmarking homogenization algorithms for monthly data. *Clim. Past*, **8**, 89–115.

Vincent, L. A., X. L. L. Wang, E. J. Milewska, H. Wan, F. Yang, and V. Swail, 2012: A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis. *J. Geophys. Res. Atmos.*, **117,** D18110.

Vose, R. S., D. R. Easterling, and B. Gleason, 2005: Maximum and minimum temperature trends for the globe: An update through 2004. *Geophys. Res. Lett.*, **32,** L23822.

Wang, K. and R. E. Dickinson, 2013: Contribution of solar radiation to decadal temperature variability over land. *PNAS*, doi:10.1073/pnas.1311433110

Wang, G. and M. E. Dillon, 2014, Recent Geographic convergence in diurnal and annual temperature cycling flattens global thermal profiles. Nature Climate Change, doi: 10.1038/NCLIMATE2378

Wijngaard, J. B, A. M. G. Klein-Tank and G. P. Konnen, 2003: Homogeneity of 20th Century European daily temperature and precipitation series. Int. J. Clim, 23: 679-692 doi:10.1002/joc.906

Williams, C. N., M. J. Menne, J. H. Lawrimore, 2012a: Modifications to Pairwise Homogeneity Adjustment software to improve run-time efficiency. NCDC Technical Report. NCDC No. GHCNM-12-01R [http://www1.ncdc.noaa.gov/pub/data/ghcn/v3/techreports/Technical%20Report%20NCDC%20No12-01R-27Jul12.pdf]. Accessed 11/13/14

Williams, C. N., M. J. Menne, and J. H. Lawrimore, 2012b: Modifications to Pairwise Homogeneity Adjustment software to address coding errors and improve run-time efficiency. NCDC Technical Report. NCDC No. GHCNM-12-02 [http://www1.ncdc.noaa.gov/pub/data/ghcn/v3/techreports/Technical%20Report%20NCDC%20No12-02-3.2.0-29Aug12.pdf]. Accessed 11/13/14

Williams, C. N., M. J. Menne, and P. W. Thorne, 2012c: Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *J. Geophys. Res. Atmos.*, **117**.

Zhou, Y. Q., and G. Y. Ren, 2011: Change in extreme temperature event frequency over mainland China, 1961–2008. *Clim. Res.*, **50**, 125–139.