AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2015JD024583

This article is a companion to *Thorne et al.* [2016] doi:10.1002/2015JD024584.

Key Points:
- Breakpoints are more prevalent in DTR than other temperature elements
- DTR has decreased since the early twentieth century, but decrease is not linear
- Effects of homogenization alter many details of global and regional changes

Correspondence to:
P. W. Thorne,
peter@peter-thorne.net

Citation:
Thorne, P. W., et al. (2016), Reassessing changes in diurnal temperature range: A new data set and characterization of data biases, *J. Geophys. Res. Atmos.*, *121*, 5115–5137, doi:10.1002/2015JD024583.

Received 30 NOV 2015
Accepted 11 APR 2016
Accepted article online 22 APR 2016
Published online 17 MAY 2016

Reassessing changes in diurnal temperature range: A new data set 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. Esau[2], A. M. G. Klein-Tank[6], and A. Merlone[7]

[1]Department of Geography, National University of Ireland Maynooth, Maynooth, Ireland, [2]Nansen Environmental and Remote Sensing Center/Bjerknes Centre for Climate Change, Bergen, Norway, [3]NOAA's National Centers for Environmental Information-Asheville, Asheville, North Carolina, USA, [4]Cooperative Institute for Climate and Satellites, Asheville, North Carolina, USA, [5]Asheville, North Carolina, USA, [6]KNMI, De Bilt, Netherlands, [7]Istituto Nazionale di Ricerca Metrologica, Torino, Italy

Abstract It has been a decade since changes in diurnal temperature range (DTR) globally have been 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 Centers for Environmental Information's automated pairwise homogeneity assessment algorithm to reassess DTR records. It is found that breakpoints are more prevalent in DTR than other temperature elements and that the resulting adjustments have a broader distribution. This strongly implies that there is an overarching tendency, across the global meteorological networks, for nonclimatic artifacts to impart either random or anticorrelated rather than correlated biases in maximum and minimum temperature series. Future homogenization efforts would likely benefit from simultaneous consideration of DTR and maximum and minimum temperatures, 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 midtwentieth century globally (−0.044 K/decade). Adjustments serve to approximately halve the long-term global reduction in DTR in the basic "raw" data. Most of the estimated DTR reduction occurred over 1960–1980. In several regions DTR has apparently increased over 1979–2012, while globally it has exhibited very little change (−0.016 K/decade). Estimated changes in DTR are an order of magnitude smaller than in maximum and minimum temperatures, which have both been increasing rapidly on multidecadal timescales (0.186 K/decade and 0.236 K/decade, respectively, since the midtwentieth century).

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 for understanding microclimate impacts, changes in stratification propensity, 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; *Steeneveld 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 and Naylor*, 2009] amongst others.

The daily maximum temperature tends to occur a few hours after maximum incipient solar radiation and when, typically, the boundary layer is well mixed. Values of *T*x hence tend to be fairly representative of at least local, if not regional, thermal properties of the near-surface atmosphere and indicative of the deeper boundary layer. Conversely, the daily minimum temperature tends to occur around dawn, when surface radiative cooling is drawing to a close. Values of *T*n will occur when the atmosphere is at its most vertically stratified. Particularly when synoptically driven atmospheric mixing is weak, *T*n will therefore tend to be spatially heterogeneous. Particularly in regions of complex surface characteristics or topography *T*n measured at a point may not even be locally representative, e.g., if a site is located in a frost hollow it will read systematically colder than its surroundings. Analysis of *T*x and *T*n and, in particular, their difference, DTR, may therefore

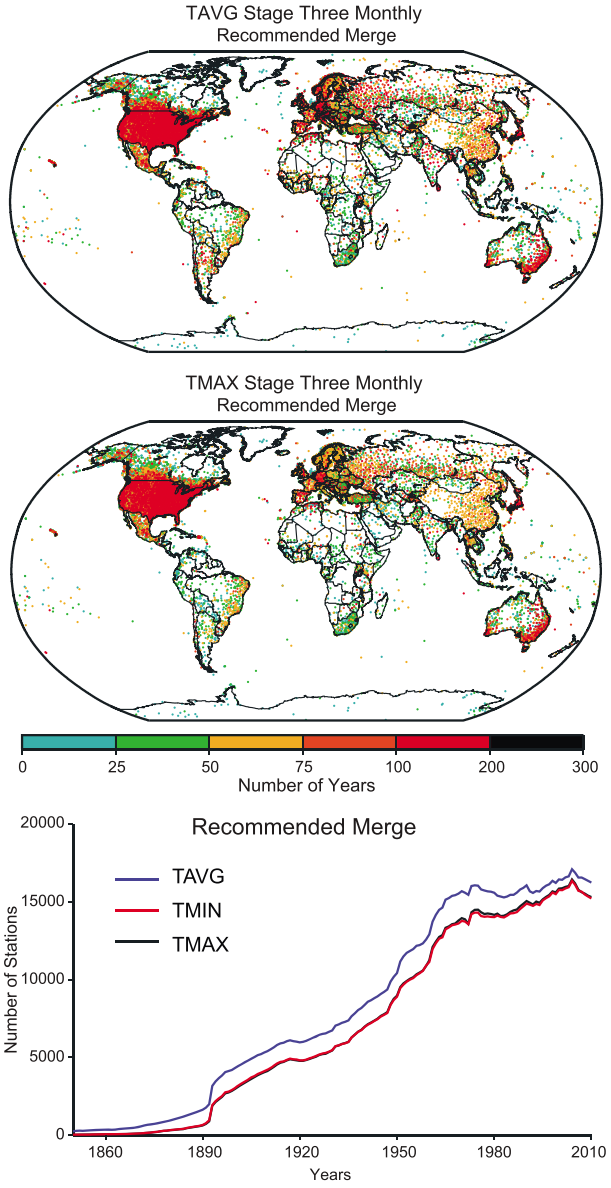

Figure 1. Summary of station holdings in the v1 global and surface air temperature databank release [*Rennie et al.*, 2014]. Station availability for (top) *T*m and (middle) *T*x. Longer series overplot shorter series. *T*n is similar to *T*x. (bottom) The station count through time by element (time axis truncated to midnineteenth century on for presentational purposes).

elucidate both on important climate processes and provide additional insight beyond more typical considerations of mean temperature, *T*m, into the nature of nonclimatic influences in the available observational records.

Meteorological records have been undertaken at observing stations that extend back to the late eighteenth century regionally and to the late nineteenth 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 data sets globally [*Lawrimore et al.*, 2011; see also *Williams et al.*, 2012a, 2012b, 2012c; *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). It is important to recognize that in many parts of the world, *T*m is derived from fixed-hour observations rather than *T*x and *T*n, with the (*T*x + *T*n)/2 method to derive *T*m historically being used mostly in the English-speaking world and former British colonies [*Trewin*, 2010]. In many places, therefore, historical observations of *T*x and *T*n (which, in the pre-automatic weather station era, required separate instruments) may not exist for at least long periods of the record when *T*x and *T*n were not required reporting variables.

The dropout of *T*x and *T*n data from the holdings considered herein becomes substantial prior to about 1950 and critical prior to 1895 (Figure 1, bottom). Most U.S. 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 [*Rennie et al.*, 2014].

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 has been a decade since the last stand-alone comprehensive analysis of global DTR data and its homogeneity was produced [*Vose et al.*, 2005] and over 20 years since the first such assessment [*Karl et al.*,

1993]. The Intergovernmental Panel on Climate Change (IPCC) in the most recent working group 1 assessment [*Hartmann et al.*, 2013] noted that there was only "medium confidence" (see *Mastrandrea 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 the following: (1) 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]; (2) 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 (3) creating automated monthly climatic time series 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.*, 2016]. 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 enhanced data holdings, a characterization of the resulting series, and a consideration of implications for trends in DTR, *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 nonclimatic artifacts in the record. Section 4 describes the spatial and temporal evolution of the homogenized series for both the spatially incomplete global mean and a subset of regions for which data are complete enough to analyze back to 1901 (Europe, North America, and Australia). Section 5 provides a brief discussion. Section 6 contains details on the data set 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 thousands of 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, better provenance, and daily data 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 several thousands of which extend over at least 100 years (although not necessarily continuously). Where daily data sources are available, monthly averages have been calculated only where sufficient days (per present-day World Meteorological Organization guidelines) were reported [*Rennie et al.*, 2014]. For sources available only as monthlies the submonthly completeness criteria are unknown (hence their lower priority in the merge).

The processing of the databank series merged the *T*x and *T*n series stations first and only then went back to look for record segments for which solely *T*m 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 (in the English-speaking world in particular where maximum and minimum thermometers have been predominant), or at least observations at regular intervals over the day, were associated with the original records for which, in most cases, only *T*m data now exist in the digital archives. These data either have 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 or subdaily observations, which serves to inhibit understanding and scientific utility of the records [e.g., *Allan et al.*, 2011].

To facilitate the analysis herein, a fourth field—*T*dtr (or DTR), the difference between *T*x and *T*n—has also been calculated and analyzed. For those analyses 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 spatiotemporally complete merged *T*m series (Figure 1). It is important to note that in many cases for these remaining *T*m reports in the databank, the archived *T*m may not result simply from averaging *T*x and *T*n. For example, at least in Australia and in several European countries, in some recent years the monthly average reported in CLIMAT messages is the average of a number of subdaily reports, which can impact *T*m series homogeneity [*Fawcett et al.*, 2012]. 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 used herein result from direct calculation from the monthly mean *T*x and *T*n series. So the basic data used 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 monthly *T*x and *T*n are present. However, for months where either *T*x and/or *T*n have missing daily values, the monthly means of *T*x and *T*n are not consistent with the average of the calculable daily DTRs (or *T*ms) within the month. While a more restrictive criteria of calculation of these *T*m and DTR values from the daily data 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 choice, which maximizes data availability, comes at a potential cost regarding the monthly statistical mean (which could be somewhat biased) and/or standard deviation (which would be somewhat inflated) characteristics, for those monthly resolution input data stations where data are patchy on an intramonth basis, due either to frequent missing days or frequent quality control flagging on the daily reports. Where days are missing, it may be that the value returned on the day of resumption was the maximum/minimum observed over the interruption in reporting, further complicating analysis. In general, however, records from sites that report on a sustained basis in the daily archives report regularly with few data interruptions across most of the globe. So, assuming that this reporting performance is representative of the reporting for those sites only available as monthly averages, the effects should be relatively minor.

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, 2012b, 2012c]. 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]. Data are input as monthly means, but returned adjustments are seasonally invariant estimates. The interested reader is directed to the method 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 future 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 time series 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 (or unconfounded), such that breakpoints are assigned iteratively to those stations in which they arise concurrently across multiple intercomparisons, 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 nonzero. 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 [*Lawrimore et al.*, 2011]. 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. Consideration of such ensembles is deemed beyond the scope of the present analysis.

2.3. Station Gridding

For the 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. This is one of several possible approaches to gridding, as is discussed in the accompanying paper [*Thorne et al.*, 2016]. For each station and calendar month, the minimum data requirement for calculating a climatology is two thirds of reported monthly mean data in the 30 year period taken as a whole and at least one half 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 effect 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 12 calendar months.

The climatology value has been calculated through a trimmed mean based solely upon months within 3 standard deviations (σ) of the climatology period data mean for the given calendar month. An additional simple 5σ anomaly quality control 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 (or a highly skewed distribution arising from real physical effects such as a high dependency on wind direction) this quality control step may remove real points, especially so far away in time from the climatology period. A high critical threshold of 5σ was chosen to mitigate this risk while still ensuring that grossly questionable data did not get gridded. The check only removes a handful of grossly questionable data points from individual series. There were no cases where an entire station was deemed necessary to be removed.

Resulting anomalies have simply been gridded, without any further weighting, into bins of 5° latitude by 5° 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 time series 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 across the three series 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. 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. Such analyses are discussed further in the accompanying paper [*Thorne et al.*, 2016].

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). We assume that a nonclimatic effect imparts a breakpoint onto all four elements, even if it is not detectable in one or more of them. The only exception would be replacement of one or other of a maximum/minimum thermometer for those sites which employ this method, in which case the break would not impact the other primary variable (*T*x in the case of a minimum thermometer replacement and vice versa) unless the new instrument changed the thermal and radiative microenvironment properties within the

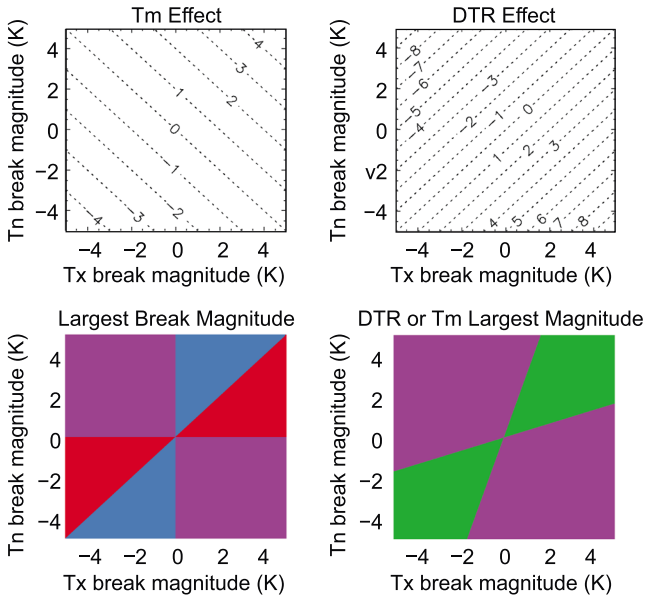

Figure 2. Summary of potential breakpoint structure and magnitude for different combinations of break magnitudes in *T*x and *T*n. (top row) The effects in (left) *T*m and (right) DTR. (bottom left) The greatest magnitude break (*T*x—red, *T*n—blue, DTR—purple; there are no cases where *T*m is uniquely the largest breakpoint). (bottom right) Same as Figure 2 (bottom left) but restricting to DTR and *T*m (DTR—purple, *T*m—green).

instrument shelter/housing. For all remaining cases the a priori expectation is that a true break affects both primary and hence both derived variables.

To ascertain from a purely theoretical standpoint 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 cancelation in breakpoints of DTR and absolute breakpoint magnitudes reach 10 K at [−5 K, 5 K] and [5 K, −5 K]. Hence, DTR has twice as large a potential breakpoint magnitude for combinations explored as any of the other elements. Importantly, 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 anticorrelated breakpoints (*T*x break = −*T*n break), there will be no breakpoints in *T*m. It further follows that if one assumes that breakpoints of *T*x and *T*n both have distributions with standard deviation σ, the expected standard deviation of breakpoints of DTR would be approximately $\sqrt{2}\sigma$, and of *T*m would be $\sigma/\sqrt{2}$ assuming an entirely random distributional relationship between the *T*x and *T*n breaks.

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 anticorrelated (one positive, one negative), the largest breakpoint will always be in DTR (Figure 2, bottom left). Restricting to a consideration of solely *T*m and DTR breakpoint behaviors, the breakpoint in DTR will be largest both when the breakpoints in *T*x and *T*n are anticorrelated, and when they are only weakly correlated (same sign but substantially distinct magnitude, whereby the difference is greater than their mean, Figure 2 bottom right).

Making a simplifying assumption that the interstation 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 based upon Figure 2 (bottom row):

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. In Figure 2 (bottom row) there is a 50% chance of DTR breaks being largest and a 25% chance each of *T*x and *T*n being largest from an entirely random draw (which is unlikely in the real world).
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)).

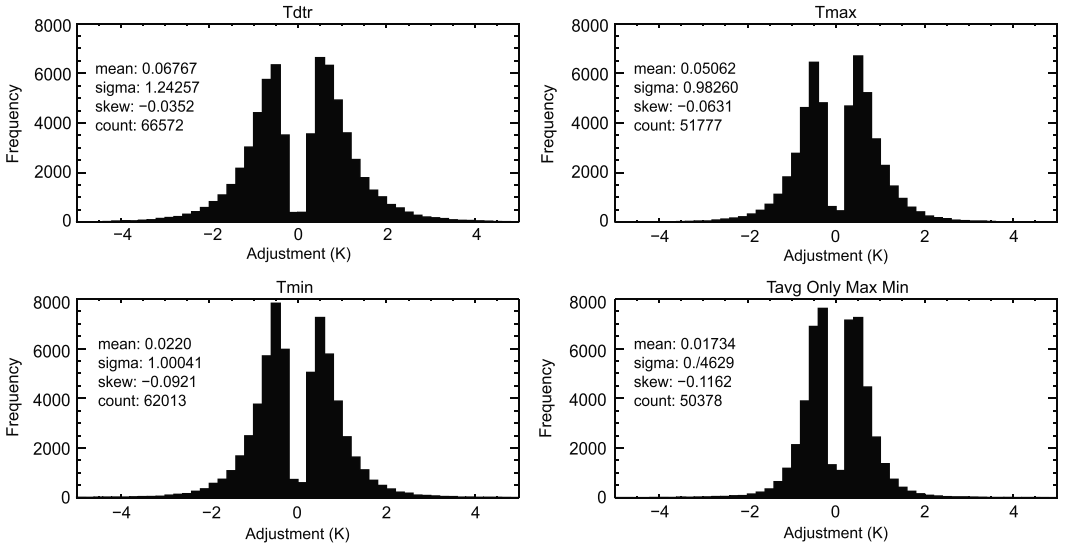

Figure 3. Distribution of applied adjustments by running the PHA on (top left) diurnal temperature range, (top right) maximum temperatures, (bottom left) minimum temperatures, and (bottom right) average temperatures derived from maximum and minimum measures only. In each panel is given the mean of the adjustments population, its standard deviation, skew, and the count of returned adjustment estimates.

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 found 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 exist 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 four 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 (1.24 K), roughly equivalent for *T*x (0.98 K) and *T*n (1.00 K), and smallest for *T*m (0.75 K). 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.

As discussed in section 4, much of the U.S. network underwent a transition from Cotton Regional Shelter measurements (Stevenson screens) to Maximum Minimum Temperature Sensors. The transition imparted a cooling artifact into *T*x and a warming artifact into *T*n (see section 4 for further discussion and references). The change occurred at roughly two thirds of U.S. stations concentrated over a period of about 3 years in the mid-1980s. Given the density of U.S. networks compared to the rest of the world, the resultant behavior in Figure 3 may be dominated by this effect. To ascertain whether this was the case, the Federal Information Processing Standard country identifier in the International Surface Temperature Initiative's databank station identifiers was used to separate the U.S. from the rest of the world stations. Statistics for each subset (identical to those in Figure 3 panels) are presented in Table 1. It is clear that there is a strong degree of congruence between U.S. stations breakpoint adjustments behavior and those for the rest of the world. The biggest difference in behavior relates to the skew of the distribution, with stations outside the U.S. exhibiting greater skew in all four elements. The transition from Cotton Region Shelters (CRS) to Maximum Minimum Temperature Sensor (MMTS) is therefore not a primary explanation for the observed global adjustments behavior in terms of either the frequency of breakpoints or the standard deviation of applied adjustments across the four elements.

Table 1. Summary of Breakpoint Adjustment Characteristics Broken Down Between the U.S. Stations and the Rest of the World Stations[a]

Temperature Element	DTR	*T*x	*T*n	*T*m
U.S. Stations				
Mean adjustment (K)	0.084	0.046	−0.040	0.003
Adjustment standard deviation	1.214	0.925	0.970	0.713
Adjustment skew	0.020	−0.062	−0.045	−0.051
Adjusted breakpoint count	42173	34771	40294	33712
Rest of the World Stations				
Mean adjustment (K)	0.040	0.060	0.012	0.047
Adjustment standard deviation	1.290	1.091	1.053	0.809
Adjustment skew	−0.108	−0.070	−0.175	−0.230
Adjusted breakpoint count	24399	17006	21719	16666

[a]Many U.S. stations changed from CRS to MMTS, and this change is known to impart a large break into DTR (sections 3 and 4). Despite this, U.S. and rest of the world adjustments exhibit similar behavior. The collated results for the entire network are given in Figure 3.

Following from section 3.1, if there is no difference between elements in effective power of PHA to detect and adjust for breaks, 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 covary 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 of the bimodal distribution of adjustments, 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 principally results from larger discontinuities rather than any substantial 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 adjusted for 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, if the techniques used were perfect. 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). 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 $y = 0, x \neq 0$ or $x = 0, y \neq 0$, 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) is zero mean and broadly Gaussian, albeit with a large sigma, such that almost 23% of applied adjustment differences exceed 1 K in magnitude. A similar analysis of *T*m (Figure 5) exhibits far less scatter between directly and indirectly inferred adjustments (Figure 5, left, points lie much closer to the 1:1 line), with $<5\%$ of differences exceeding 1 K in magnitude (Figure 5, right).

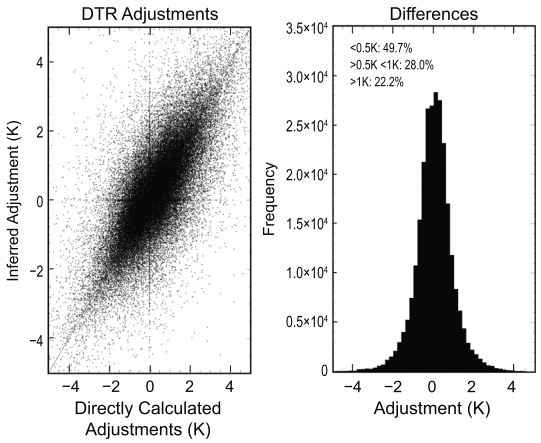

Figure 4. Analysis of applied against indirectly inferred adjustments returned by running the PHA algorithm on DTR (directly adjusted) and maximum and minimum (indirectly inferred) temperatures. Screening has been applied such as to include only those differences that have at least 12 data points, to account for uncertainty inherent in ascertaining the unique locations of breakpoints under the PHA algorithm. (left) A scatterplot of all retained pairs. (right) A histogram summary of the differences between directly applied and inferred adjustments.

Both direct and indirect adjustments to DTR act to reduce the apparent spread in individual station linear

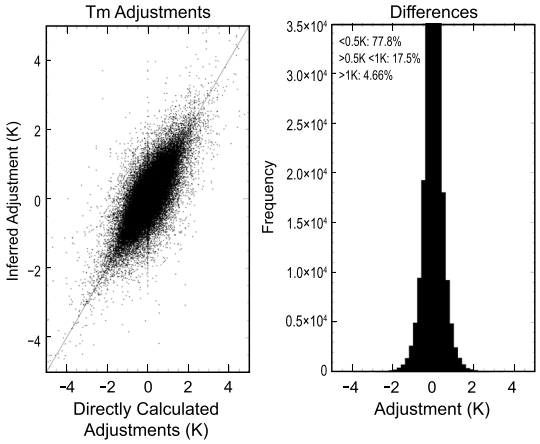

Figure 5. As in Figure 4 but for average temperature adjustments. Axes ranges in both figures are identical.

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], also found that breakpoints in DTR were somewhat more amenable to detection. Not only were more breakpoints found in the present analysis in DTR

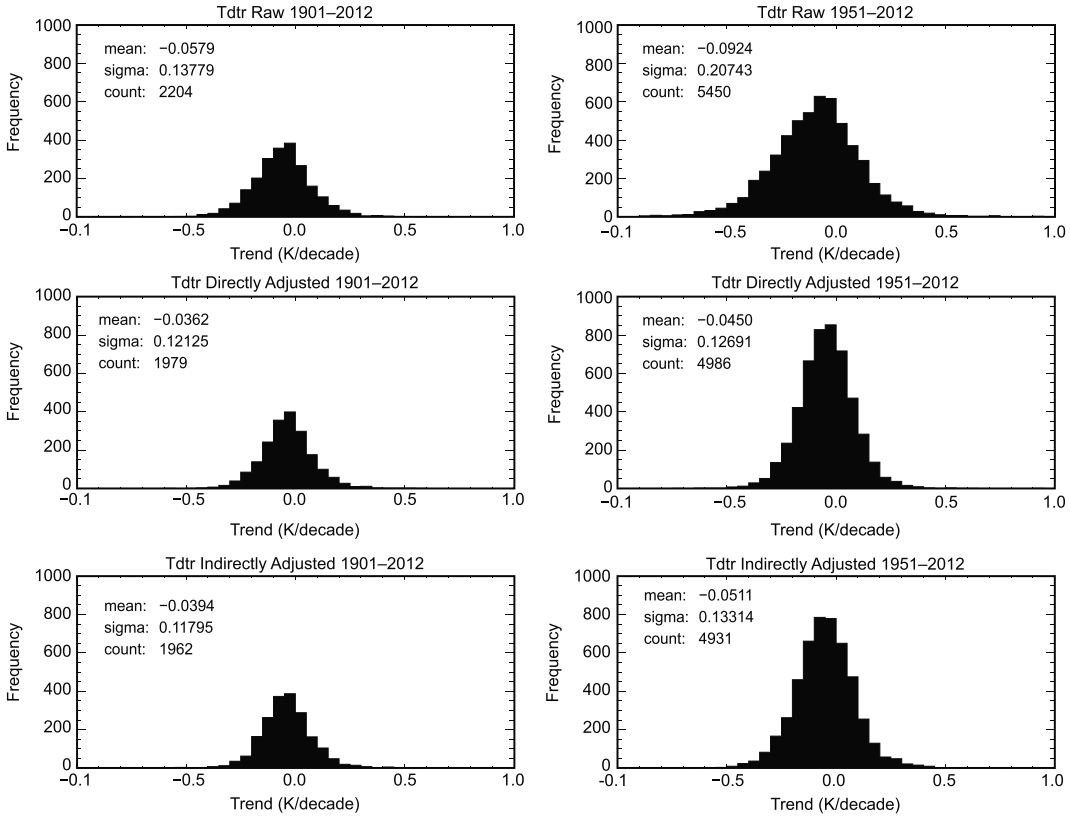

Figure 6. Histogram of station trends over the periods (left column) 1901–2012 and (right column) 1951–2012, for stations with sufficient data completeness. Trends are ordinary least squares (OLS) fits in K/decade. (top row) Raw series calculated directly from the databank holdings [*Rennie et al.*, 2014], (middle row) analysis when directly calculated adjustments are applied, and (bottom row) analysis when indirectly calculated adjustments are applied. In each panel is included the mean, sigma, and station count. Station count varies due to degree of quality control and deletion of short segments in different PHA realizations.

but they were on average larger (and hence 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 globally, 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 (although this does not rule out some breaks of this nature or their potential prevalence regionally and/or in certain periods). Split analysis between U.S. and remaining stations rules out the CRS to MMTS transition, which is known to have a strong effect on DTR, as the dominant cause of the observed behavior. Reasons and implications for this global behavior 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 are globally incomplete, and the data density in those areas sampled varies over at least 2 orders of magnitude. Figure 7 shows grid box 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 nonexistent) 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 twentieth century (Figure 8). Data dropoff outside of the climatology period to some extent is inevitable whatever climatology period is chosen as stations open, close, and get relocated/reallocated for other purposes. However, outside North America, there exists a step change in availability in 1960, with far fewer stations prior to this. As a result, outside North America trends and variability in DTR for analyses spanning across 1960 may, to some extent, 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 both directly and indirectly as outlined in section 2.3. The analysis starts with spatial patterns of trends over increasingly shorter periods to the present. Recourse is then made to regionally averaged time series behavior and linear trend estimates.

4.2.1. Spatial Trends

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

Century timescale trends in DTR (Figure 9) are at most of the order of 0.1 K/decade across the sampled grid boxes in the raw data, and in the two adjusted products. Trends are significant at the grid box level in many of the grid boxes 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 grid boxes exhibit a reduction in DTR over time. Although a majority of grid boxes still indicate a reduction in DTR following the application of adjustments, the magnitude of the DTR reduction is statistically significant across fewer grid boxes. Adjustments change the sign of the DTR trends in much of the southwestern/western United States from negative to positive and reduce the negative trends elsewhere in North America. This change is more marked when

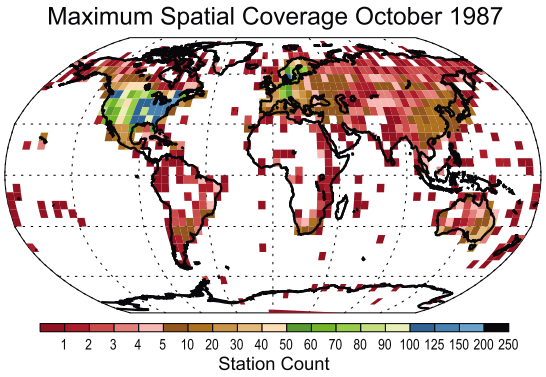

Figure 7. Maximum spatial coverage (October 1987) for the data set, after application of a 1971–2000 climatology period requirement on all station records and gridding onto a 5° by 5° grid. The majority of grid boxes contain fewer than 10 stations, but some well-sampled regions contain many more. Coverage varies substantially through time (cf. Figure 8).

adjustments are calculated indirectly than when they are calculated directly. There are less spatially consistent changes in remaining regions, with many individual grid boxes experiencing large changes including changing the sign of the apparent DTR trend.

Starting in 1951, as expected from Figure 8, spatial sampling is much more complete although Africa, the Indian subcontinent, and South 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 grid boxes exhibit substantial reductions in DTR that are particularly marked over much of Asia and North America. Application of adjustments substantively changes the trend behavior over North America, where trends are reduced with a sign change in many grid boxes west of the Rockies to an increasing DTR, and very few North American grid box series trends remain statistically 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

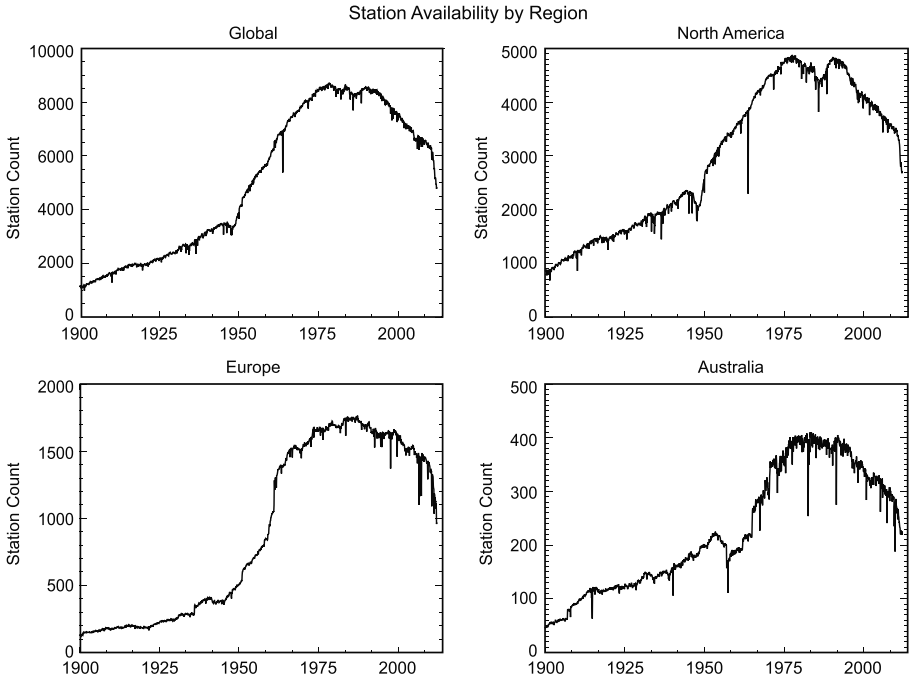

Figure 8. Station count over time for the globe and North America (45°W–135°W, 25–60°N), Europe (10°W–60°E, 25–60°N), and Australia (110°E–155°E, 10°S–45°S). Note that the *y* axis range in each panel differs substantially. Stations are included only if a climatology over 1971–2000 can be calculated.

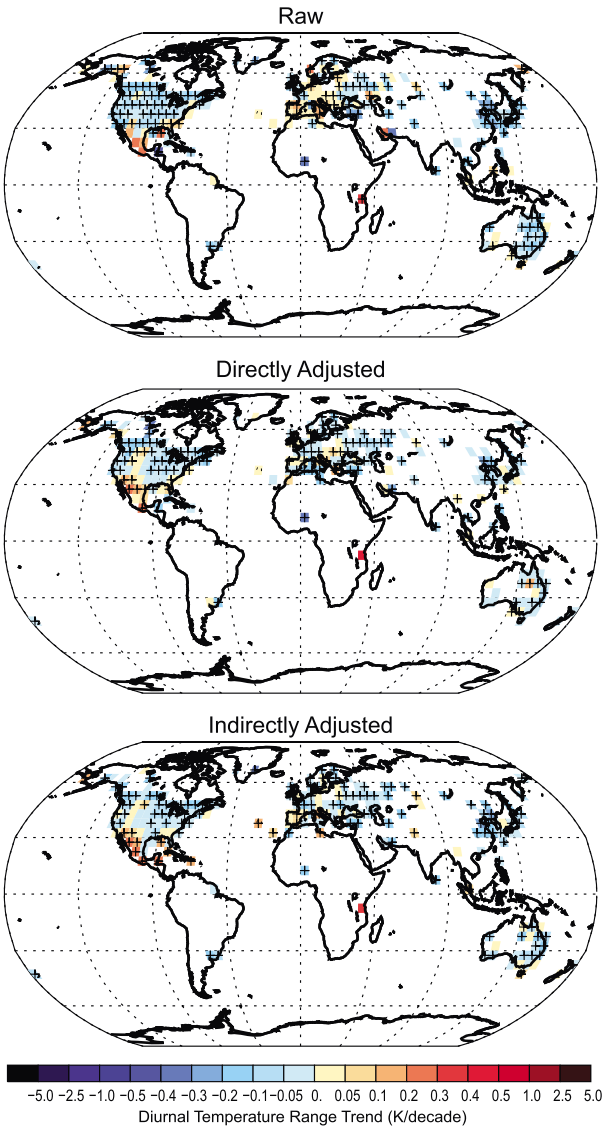

Figure 9. Grid box linear trends in DTR estimated using OLS with AR(1) dof correction from 1901 to 2012. Only grid boxes with >70% data availability within the period, and sufficient data in the first and last decile of the period, are included. Trends that are significantly different from zero are denoted by a cross. (top) Trends from the raw databank holdings, (middle) after running PHA directly upon these DTR series, and (bottom) after inferring returned adjustments from those returned from running on *T*m and *T*n.

indirectly to DTR records. However, use of a spatial smoothness criteria alone to ascertain efficacy of adjustment approaches may be misleading [*Sherwood et al.*, 2009].

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 the satellite era, 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 grid boxes. The adjusted DTR increases are statistically significant in several grid boxes in the southwestern states. This adjustment is consistent with understanding of the effects 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 over 1984–1987 (although some such transitions occur even recently). In this change both the instrument and its shielding were changed substantively, often associated with a change in measurement location necessitated by the need for an electric power source. This change affected roughly 70% of the Cooperative Observer Program (COOP) network, which is the backbone of the U.S. records. Field-based studies and statistical analyses have consistently concluded that the CRS to MMTS transition led to a positive bias in *T*n and a negative bias in *T*x, 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 apparent 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

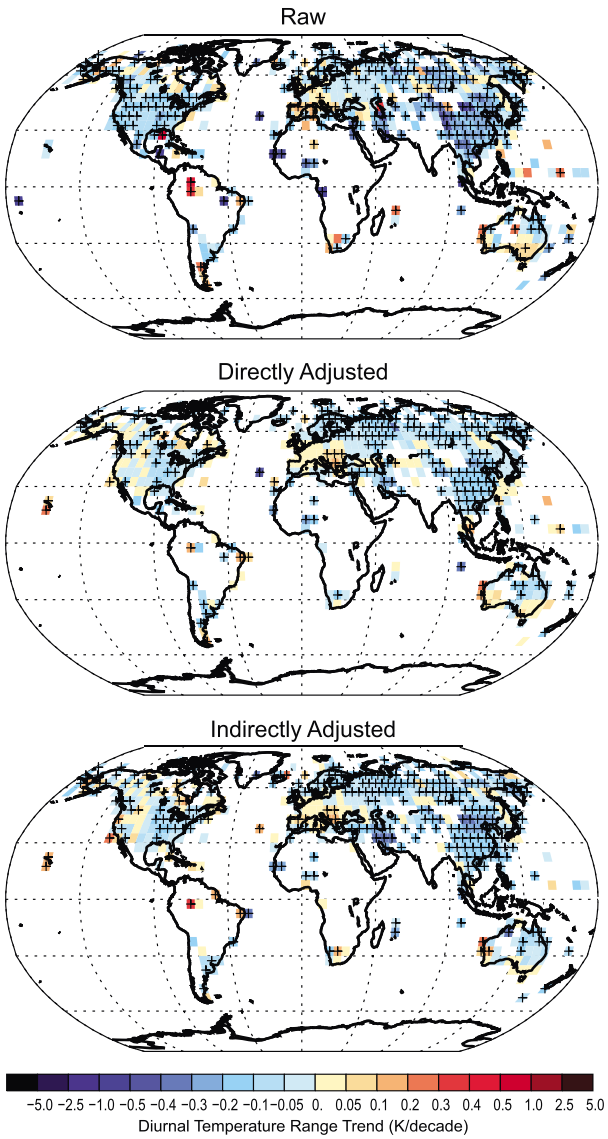

Figure 10. As in Figure 9 but for the period 1951 to 2012.

is quantitatively consistent with understanding from various side-by-side field comparisons, under the assumption that approximately 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 grid box 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 Time Series and Trends

As is visually obvious from Figures 9–11, linear trend estimates do not describe all facets of the time series behavior globally or regionally, which is trivially true by definition. Time series for global (Figure 12) and regional (Figure 13) DTR averages serve to highlight the presence of substantial interannual to multidecadal variability in DTR, even globally. In all cases, these time series have been derived from averaging all available gridded data at each time step using cos(lat) area weighting. As noted earlier, care should be taken in interpretation in particular of behavior across 1960, when coverage and station count both increase substantively. The effects of different completeness inclusion criteria for this step are further discussed and analyzed in the accompanying paper [*Thorne et al.*, 2016]. Furthermore, in all cases series shall be similar by construction over the chosen climatology period and diverge away from this, so care is required in interpretation [*Hawkins and Sutton*, 2015].

Following adjustments, it is estimated that globally averaged DTR was elevated relative to present day until the late 1950s, declined of the order 0.2°C 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. The apparent quiescent effect of adjustments in 1970–2000 may in large part be an artifact of the choice of this period as the climatological basis period as noted above.

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 2). Nonetheless, trends toward globally reduced DTR are statistically significant

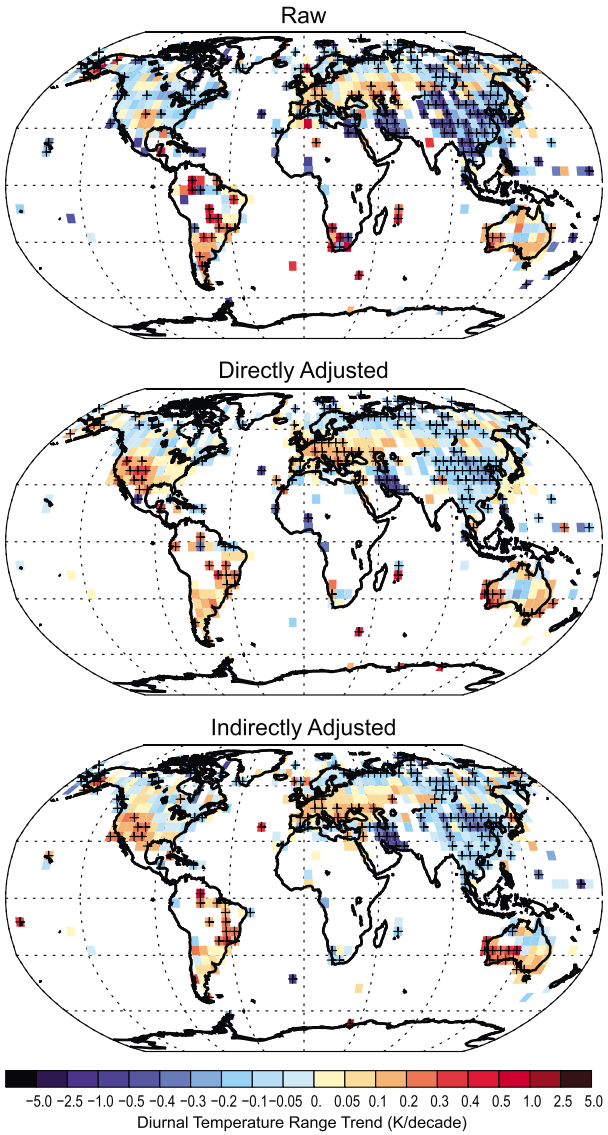

Figure 11. As in Figure 9 except for the 1979–2012 period.

over the period 1901 to 2012, and the shorter subperiod 1951 to 2012, for the raw series and remain so for the adjusted series. Over the period 1979 to 2012, the global mean trend changes from a significant reduction in the raw data, to only a very slight reduction in both of the adjusted series, neither of which is statistically significant (cf. 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). As discussed previously, post-1980 changes are consistent with understanding of the effects of transition from CRS to MMTS across roughly 70% of the U.S. 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 2). 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). This yields a marked change in multidecadal 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 2). Over 1951–2012, all three trend 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 North 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, bottom). Variability appears to be highly correlated with continental scale aridity/rainfall (and by extension El Niño–Southern Oscillation (ENSO)) [*Karoly and Braganza*, 2005]. For example, the very wet year of 2010/2011 is associated with a marked negative DTR anomaly, consistent with basic theoretical understanding of partitioning of fluxes [*Peterson et al.*, 2011]. Unlike Australian averaged precipitation which is dominated by ENSO, changes in aridity/wetness tend to cancel more on European and North American domains and hence

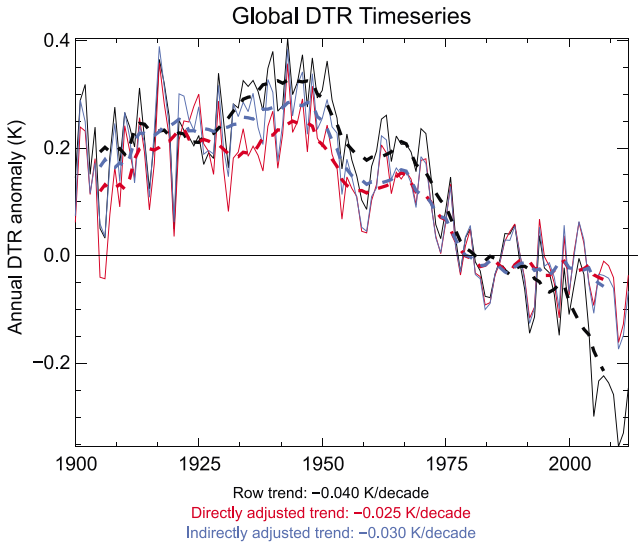

Figure 12. Globally averaged DTR behavior. Each monthly DTR average has been defined by a simple cos(lat) weighted average of all grid boxes reporting data in the given month. Then annual means (solid) and running decadal means (dashed) calculated. Data are globally incomplete (Figure 6) and vary substantially in density and trend behavior regionally over time (Figures 7–10). Black is raw data, red is homogenized directly, and blue is homogenized indirectly. Trend estimates from OLS are given below the figure panel for the whole period of record. See Table 2 for their 5–95% confidence intervals (CIs) and a comparison to those for other periods and the regions in Figure 13. Close agreement during 1971–2000 is in part an artifact of the data shown being anomalies from this base period.

be somewhat less extreme at the continental average scale even when they occur. The effect of the adjustments is more muted for Australia, with slight increases in DTR in the midtwentieth century and reductions in the early twentieth century. Trends over the three periods considered are not significant in the adjusted series, with the exception of the indirectly adjusted series for 1901–2012 (Table 2), and confidence intervals are larger than for other regions considered, reflecting the much greater year-to-year variability in the series. Interestingly, *Fawcett et al.* [2012] find that the reduction in DTR essentially disappears if the rainfall effects are accounted for by regression to remove this component. Over this region, there is less obvious concordance between the two 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 and statistically significant warming in the vast majority of the grid boxes 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 U.S. biases arising from the CRS to MMTS transition. Cooling in *T*x in Southern China is also reduced, and several obviously erroneous grid box series in the raw data look more similar to surrounding series after homogenization. Adjustments to *T*n also serve to adjust several obviously erroneous grid box trends and increase slightly the apparent warming in eastern North America but otherwise have little obvious effect at the grid box scale.

Global average time series 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 3). 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 *T*m, 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 is whether there are more optimal approaches to homogenization than analyzing *T*m, as is commonly the case for global centennial

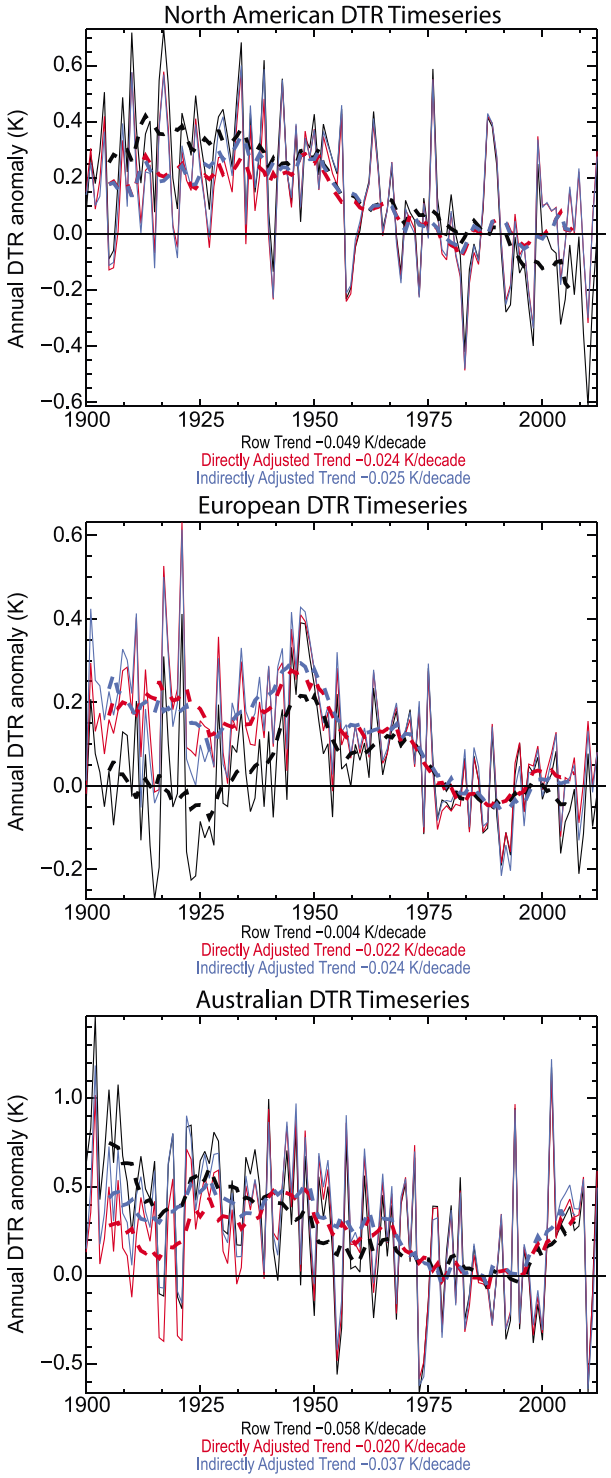

Figure 13. As in Figure 12 but for the three regional domains defined in Figure 7. Note that the *y* axis range varies by region and in each case is distinct from that in Figure 12.

timescale Land Surface Air Temperature reconstructions to date, and second is why, metrologically, the overarching 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 independent and identically distributed white noise arising from random measurement errors and real physical intersite variability. This 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 difference series, regardless of the choice of test statistic. 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 efficacy, while simultaneously minimizing false alarm rates. All breakpoint algorithms return bimodal adjustment distributions (cf. Figure 3) that in reality are the two wings of the true Gaussian distribution of real-world breaks, with breaks around zero (which are very likely the largest subpopulation) 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 toward 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,

Table 2. Decadal Trend Estimates in DTR (K/Decade) for the Globe and Several Subregions[a]

Period	Adjustments Applied	Global	North America	Europe	Australia
1901–2012	Raw	**−0.042** ± 0.022	**−0.052** ± 0.016	−0.005 ± 0.011	**−0.054** ± 0.022
	Directly adjusted	**−0.026** ± 0.010	**−0.025** ± 0.015	**−0.023** ± 0.008	−0.021 ± 0.022
	Indirectly adjusted	**−0.032** ± 0.011	**−0.026** ± 0.016	**−0.026** ± 0.009	**−0.037** ± 0.021
1951–2012	Raw	**−0.088** ± 0.034	**−0.071** ± 0.045	**−0.032** ± 0.014	−0.006 ± 0.059
	Directly adjusted	**−0.044** ± 0.017	−0.026 ± 0.045	**−0.029** ± 0.016	−0.025 ± 0.060
	Indirectly adjusted	**−0.051** ± 0.018	−0.031 ± 0.045	**−0.038** ± 0.015	−0.029 ± 0.066
1979–2012	Raw	**−0.092** ± 0.068	−0.079 ± 0.121	−0.007 ± 0.032	0.061 ± 0.140
	Directly adjusted	−0.016 ± 0.031	0.036 ± 0.120	0.030 ± 0.030	0.086 ± 0.135
	Indirectly adjusted	−0.022 ± 0.033	0.032 ± 0.121	0.016 ± 0.034	0.083 ± 0.139

[a]All trends have been calculated using OLS accounting for AR(1) effects from annual means after *Santer et al.* [2008]. Ranges are two-tailed 90% confidence intervals. Trends for which zero is not encompassed within this range are highlighted in bold. Region definitions are as in Figure 6.

as the dominant direction of breaks becomes increasingly orthogonal to *T*m (Figure 3). Section 3 strongly implies breakpoints are at least 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

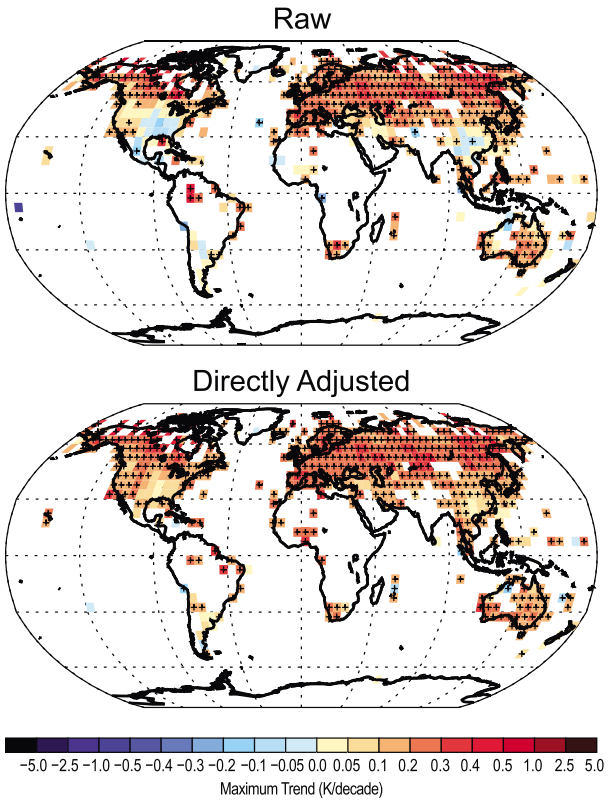

Figure 14. As in Figure 9 but for trends in maximum temperatures over 1951–2012. For maximum temperatures only direct adjustments have been made. Compare to Figure 10 for trends in DTR and Figure 15 for trends in minimum temperatures over the same period calculated using the same method.

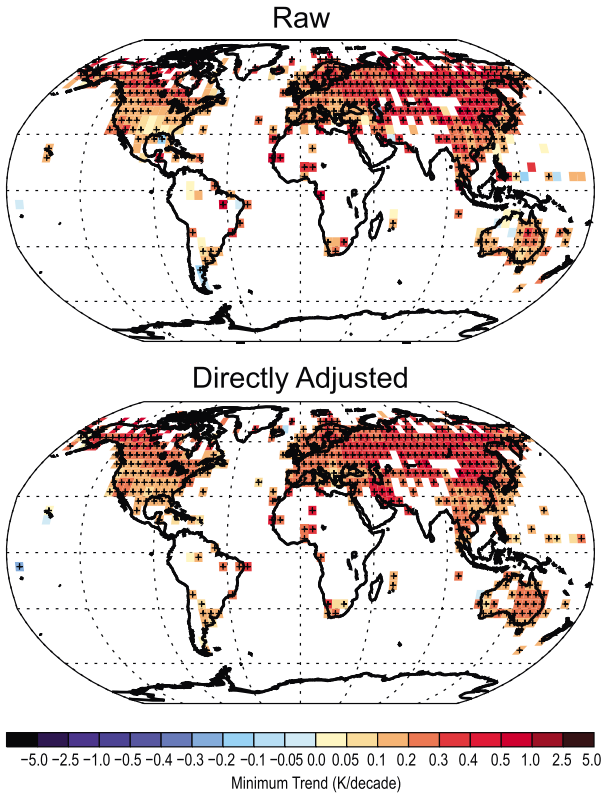

Figure 15. As in Figure 9 but for minimum temperatures over 1951–2012. Compare to Figure 10 for trends in DTR and Figure 14 for trends in maximum temperatures over the same period calculated using the same method.

set of four adjustment estimates (or better still conditional density functions of these 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.

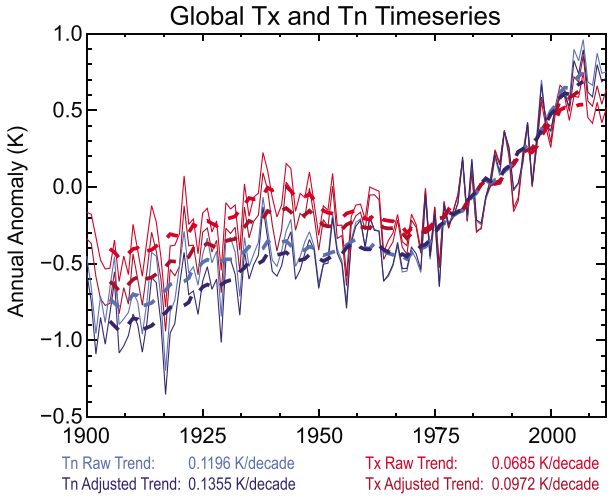

Figure 16. Time series plot for globally averaged maximum (*T*x, red shaded lines) and minimum (*T*n, blue shaded lines) temperatures using the same method as for DTR in Figure 11. Note that the *y* axis range in this figure is 2.5 times that in Figure 11. Only direct adjustments are available for these elements.

Table 3. Linear Trend Fits and 90% CIs Calculated As in Table 2 but for Global Averages and for DTR, *T*x, and *T*n in K/Decade

	Raw	Directly Adjusted	Indirectly Adjusted
		1901–2012	
DTR	**−0.042** ± 0.022	**−0.026** ± 0.010	**−0.032** ± 0.011
Maximum	**0.070** ± 0.024	**0.099** ± 0.024	
Minimum	**0.122** ± 0.031	**0.138** ± 0.023	
		1951–2012	
DTR	**−0.088** ± 0.034	**−0.044** ± 0.017	**−0.051** ± 0.018
Maximum	**0.154** ± 0.056	**0.186** ± 0.060	
Minimum	**0.242** ± 0.067	**0.236** ± 0.055	
		1979–2012	
DTR	**−0.092** ± 0.068	−0.016 ± 0.031	−0.022 ± 0.033
Maximum	**0.236** ± 0.079	**0.277** ± 0.078	
Minimum	**0.327** ± 0.064	**0.300** ± 0.058	

[a]Values in bold are 90% CI using OLS corrected for dof.

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, bottom). 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 those 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 reference standards involved. In these calibration approaches, such as climatic chambers or liquid baths, the sensors are kept in adiabatic conditions with the environment, to guarantee temperature stability and uniformity. The calibration uncertainty budget is hence composed of a number of components related to the characteristics of the calibration system and can only ever represent a portion of the total uncertainties associated with field measurements.

Far from being in thermal adiabatic condition, when operating in the field, a thermometer used to measure air temperature actually measures the instantaneous mix of convective, radiative, and contact heat transfers. All of these thermodynamic effects are difficult to be corrected, with a quantified 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 such 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 typical 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) [*Grykałowska et al.*, 2015]. 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 related to the measurement environment.

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 natural or, more recently aspirated, 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. For long-term stations typically several changes in configuration shall have occurred over the station lifetime.

In principle, three physical covariates shall influence the temperature measurements: radiation, wind speed, and humidity. In days with wind blowing and limited solar 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 sensitivities to 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 nonaspirated. Artificially aspirated measurements exhibit substantially lower sensitivity to prevailing meteorological conditions, so long as they are 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. Aspirated techniques have only become common with automation in the past 30 years. Nonaspirated measures will exhibit substantially greater sensitivity to prevailing meteorological conditions and the details of shielding and instrument configuration. 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 microenvironment. The change from CRS to MMTS (both nonaspirated but very distinct) had differential effects on *T*x and *T*n, with *T*x decreasing and *T*n increasing. Manual-to-Automatic Weather Station transitions also have a potential influence on DTR via the response time of instruments—if an instrument has a faster response time then it would be expected to have a larger DTR, through better sampling of short-period peaks/troughs in temperature.

Changes in the site microenvironment can change one or both of the immediate radiative environment or the local atmospheric dynamics near the surface. A change in surface heat capacity and/or albedo will impact the local thermal environment, with the effect being most marked in synoptically calm conditions when mixing is minimized. The effect on *T*x and *T*n shall depend upon the actual nature of the microenvironment change. Because *T*x is dominated by shortwave radiative responses and *T*n by longwave responses, it is likely that the responses shall, on average, differ from one another sufficiently to be at least random if not conditionally independent.

For nonaspirated instruments, microenvironment induced changes in the natural atmospheric dynamics that mediate the instrument ventilation shall lead to a conditionally independent response. If the ventilation is increased, then *T*x shall decrease as the effects of instrument contact heating owing to shortwave heating of the structure, stand, and instrument are reduced. At the same time *T*n shall increase owing to reduced contact cooling with the instrument under a longwave cooling regime. So, when natural ventilation increases owing to microclimate effects, the DTR shall decrease. The opposite effect holds true for decreased natural ventilation.

Macroscale changes such as urbanization or afforestation affect one or more of the thermal environment, radiative environment, or atmospheric ventilation over broader scales around the site. Similarly to microsite changes, the effects of such changes are likely to be sufficiently distinct in *T*x and *T*n to not be conditionally dependent. For example, urbanization effects have a greater effect on *T*n than *T*x in most cases. There remains an open question as to at what spatial scales such land use land cover effects change from being a nonclimatic artifact that should be removed to a real regional forcing, the response to which should be retained in the record.

Time of observation biases have been shown to affect primarily *T*n or *T*x individually depending upon whether the old or changed time of measurement is close to the typical minimum (dawn) or maximum (middle to late afternoon) in the diurnal cycle, respectively. Readings made around dawn tend to double-count extreme minima across 2 days. Similarly, readings made around middle to late afternoon tend to double-count extreme maxima across 2 days. In the event that the change in time of observation is from dawn to dusk both minima and maxima would be expected to increase and vice versa. Hence, time of observation changes shall either be random or if the switch is between dawn and late afternoon or vice versa conditionally dependent such that DTR changes are, on average, smallest.

In terms of common causes of nonclimatic artifacts, that leaves the effects of site relocations (here treated as moves on a scale of hundreds of meters or more) and single instrument replacements. It is not a priori obvious that such effects would yield random, conditionally dependent, or conditionally independent responses. It is likely that the nature of the response shall vary on a case-by-case basis depending upon the particulars of the change and local/regional climate regime.

Overall, there are theoretical reasons to expect at least a substantial subset of the most common issues encountered in maintaining long-term station records—instrument changes, automation, and macroclimate and microclimate changes—to impart either random or conditionally independent break structures into the records. This does not, of course, mean that all breaks conform to this expectation. Indeed, time of observation biases, for example, are likely to be conditionally dependent.

5.3. Caveat Pertaining to Use of Current Data Products

For analyses of DTR using the data set constructed herein, the effects of the changing station availability through time are potentially an insidious effect. The primary effects are twofold. First, changing the neighbor constraint substantively through time will affect the efficacy of any homogenization algorithm, and PHA is not immune to this. Second, the changing data mask may confound a clean interpretation of global and regional trends even if the data were perfect (which they are not). Use of an anomaly method mitigates, but does not completely remove, any such effects. Care should therefore be taken in interpreting pre-1960 records when the station mix changes substantively both globally and regionally.

6. Data Set Availability

The data set is made available through https://www.maynoothuniversity.ie/icarus. The following series are made available: (1) adjusted station series as CF-compliant netcdf files (one per station) containing several time series fields and (2) gridded raw and adjusted series for *T*x, *T*n, and DTR (including indirectly adjusted) as CF-compliant netcdf files (a total of seven files).

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

7. Conclusions

The present analysis has reexamined changes in DTR globally and regionally using improved holdings and National Centers for Environmental Information's (NCEI's) PHA algorithm. Adjustments to the basic raw data have a nonnegligible impact upon the resulting series behavior on multidecadal 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 midtwentieth century, but the adjustments reduce, by about 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 approximately 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. Searching for and adjusting breaks in average temperatures is likely to be suboptimal, as the breakpoint 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 nonclimatic artifacts in the records.

Acknowledgments

We thank NOAA NCEI internal reviewers and two anonymous peer reviewers for their valuable insights which served to improve the paper. Fabio Bertiglia provided useful input to section 5. The work of NERSC was partially supported by the Belmont Forum grant HIARC 247468 and the Norwegian Centre for Climate Dynamics grant BASIC. Early work was carried out while P.W.T. was an employee of CICS-NC, a cooperative venture between NCSU and NOAA NCEI. NCEI Graphics team are gratefully acknowledged for their help in polishing the figures to be publication ready. 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, *J. Climatol.*, *6*, 661–675.

Allan, R. J., et al. (2011), The International Atmospheric Circulation Reconstructions over Earth (ACRE) initiative, *Bull. Am. Meteorol. 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 (2013), Global land-based datasets for monitoring climatic extremes, *Bull. Am. Meteorol. Soc.*, *94*, 997–1006.

Dunn, R. J. H., K. M. Willett, C. P. Morice, and D. E. Parker (2014), Pairwise homogeneity assessment of HadISD, *Clim. Past*, *10*, 1501–1522, doi:10.5194/cp-10-1501-2014.

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.*, *114*, D04102, doi:10.1029/2008JD010519.

Fall, S., A. Watts, J. Nielsen-Gammon, E. Jones, D. Niyogi, J. R. Christy, and R. A. Pielke (2011), Analysis of the impacts of station exposure on the US Historical Climatology Network temperatures and temperature trends, *J. Geophys. Res.*, *116*, D14120, doi:10.1029/2010JD015146.

Fawcett, R. J. B., B. Trewin, K. Braganza, R. J. Smalley, B. Jovanovic, and D. A. Jones (2012), On the sensitivity of Australian temperature trends and variability to analysis methods and observation networks, *CAWCR Technical Report No. 050*. [Available at http://www.cawcr.gov.au/technical-reports/CTR_050.pdf.]

Grykałowska, A., A. Kowal, and A. Szmyrka-Grzebyk (2015), The basics of calibration procedure and estimation of uncertainty budget for meteorological temperature sensors, *Met. Apps*, *22*, 867–872, doi:10.1002/met.1527.

Hartmann, D. L., et al. (2013), Observations: Atmosphere and Surface, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker et al., pp. 159–254, Cambridge Univ. Press, Cambridge, U. K., and New York, doi:10.1017/CBO9781107415324.008.

Hawkins, E., and P. D. Jones (2013), On increasing global temperatures: 75 years after Callendar, *Q. J. R. Meteorol. Soc.*, *139*, 1961–1963.

Hawkins, E., and R. Sutton (2015), Connecting climate model projections of global temperature change with the real world, *BAMS*, doi:10.1175/BAMS-D-14-00154.1.
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. Atmos.*, *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.*, *117*, D05127, doi:10.1029/2011JD017139
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. Meteorol.*, *25*, 145–160.
Karl, T. R., et al. (1993), A new perspective on recent global warming: Asymmetric trends of daily maximum and minimum temperature, *Bull. Am. Meteorol. Soc.*, *14*, 1007–1023.
Karoly, D. J., and K. Braganza (2005), A new approach to detection and anthropogenic temperature changes in the Australian region, *Meteorol. Atmos. Phys.*, *89*, 57–67.
Lawrimore, J. H., et al. (2011), An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3, *J. Geophys. Res.*, *116*, D19121, doi:10.1029/2011JD016187.
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), *Int. J. Thermodyn.*, doi: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., et al. (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, doi:10.1029/2012JD017578.
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.*, *115*, D11108, 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, *J. Atmos. Oceanic Technol.*, *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. (2014), Metrology activities in NY-Ålesund (Svalbard), submitted to Meteorological Application as MMC2014 Proceedings
Paaijmans, K. P., S. Blanford, A. S. Bell, J. I. Blanford, A. F. Read, and M. B. Thomas (2010), Influence of climate on malaria transmission depends on daily temperature variation, *Proc. Natl. Acad. Sci. U.S.A.*, *107*, 15,135–15,139.
Parker, D. E. (2006), A demonstration that large-scale warming is not urban, *J. Clim.*, *19*, 2882–2895.
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, doi:10.1029/2011GL048442.
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, doi:10.1029/2005GL024407.
Rennie, J. J., et al. (2014), The International Surface Temperature Initiative global land surface databank: Monthly temperature data release description and methods, *Geosci. Data J.*, doi:10.1002/gdj3.8.
Rohde, R., R. A. Muller, R. Jacobsen, E. Muller, S. Perlmutter, A. Rosenfeld, J. Wurtele, D. Groom, and C. Wickham (2012), A new estimate of the average Earth surface land temperature spanning 1753 to 2011, *Geoinform. Geostat. An Overview*, *1*, doi:10.4172/2327-4581.1000101.
Rohde, R., R. Muller, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Wurtele, J. Curry, C. Wickham, and S. Mosher (2013), Berkeley Earth temperature averaging process, *Geoinform. Geostat. An Overview*, *1*, doi:10.4172/2327-4581.1000103.
Santer, B., et al. (2008), Consistency of modelled and observed temperature trends in the tropical troposphere, *Int. J. Climatol.*, *28*, 1703–1722.
Santer, B. D., et al. (2011), Separating signal and noise in atmospheric temperature changes: The importance of timescale, *J. Geophys. Res.*, *116*, D22105, doi:10.1029/2011JD016263.
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, doi:10.1029/2004GL022201.
Sherwood, S. C., H. A. Titchner, P. Thorne, and M. P. McCarthy (2009), How do we tell which estimates of past climate change are correct, *Int. J. Climatol.*, *29*(10), 1520–1523, doi:10.1002/joc.1825.
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.*, *116*, D02122, doi:10.1029/2010JD014612.
Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears (2005), Uncertainties in climate trends: Lessons from upper-air temperature records, *BAMS*, *86*(10), 1437–1442.
Thorne, P. W., et al. (2011), Guiding the creation of a comprehensive surface temperature resource for 21st century climate science, *Bull. Am. Meteorol. Soc.*, doi:10.1175/2011BAMS3124.1.
Thorne, P. W., et al. (2016), Reassessing changes in diurnal temperature range: Intercomparison and evaluation of existing global data set estimates, *J. Geophys. Res. Atmos.*, *121*, doi:10.1002/2015JD024584.
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. (2010), Exposure, instrumentation and observing practice effects on land temperature measurements, *Wiley Interdiscip. Rev. Climate Chang.*, *1*, 490–506, doi:10.1002/wcc.46.
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. R. 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, *Nat. Geosci.*, *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.*, *117*, D18110, doi:10.1029/2012JD017859.

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, doi:10.1029/2005GL024379.

Wang, G., and M. E. Dillon (2014), Recent geographic convergence in diurnal and annual temperature cycling flattens global thermal profiles, *Nat. Clim. Change*, doi:10.1038/NCLIMATE2378.

Wang, K., and R. E. Dickinson (2013), Contribution of solar radiation to decadal temperature variability over land, *Proc. Natl. Acad. Sci. U. S. A.*, doi:10.1073/pnas.1311433110.

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. Climatol.*, *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, [Available at 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. [Available at 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.*, *117*, D05116, doi:10.1029/2011JD016761.

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