Reassessing changes in diurnal temperature range: Intercomparison and evaluation of existing global data set estimates

P. W. Thorne[1,2], M. G. Donat[3], R. J. H. Dunn[4], C. N. Williams[5], L. V. Alexander[3], J. Caesar[4], I. Durre[5], I. Harris[6], Z. Hausfather[7], P. D. Jones[6,8], M. J. Menne[5], R. Rohde[9], R. S. Vose[5], R. Davy[2], A. M. G. Klein-Tank[10], J. H. Lawrimore[5], T. C. Peterson[11], and J. J. Rennie[12]

[1]Department of Geography, Maynooth University, Maynooth, Ireland, [2]Nansen Environmental and Remote Sensing Center/Bjerknes Centre for Climate Change, Bergen, Norway, [3]Climate Change Research Centre and ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, New South Wales, Australia, [4]Met Office Hadley Centre, Exeter, UK, [5]NOAA's National Centers for Environmental Information, Asheville, North Carolina, USA, [6]Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK, [7]Energy and Resources Group, University of California, Berkeley, California, USA, [8]Center of Excellence for Climate Change Research / Dept of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia, [9]Berkeley Earth, Berkeley, California, USA, [10]KNMI, De Bilt, Netherlands, [11]Unaffiliated, Asheville, North Carolina, USA, [12]Cooperative Institute for Climate and Satellites, Asheville, North Carolina, USA

Abstract Changes in diurnal temperature range (DTR) over global land areas are compared from a broad range of independent data sets. All data sets agree that global-mean DTR has decreased significantly since 1950, with most of that decrease occurring over 1960–1980. The since-1979 trends are not significant, with inter-data set disagreement even over the sign of global changes. Inter-data set spread becomes greater regionally and in particular at the grid box level. Despite this, there is general agreement that DTR decreased in North America, Europe, and Australia since 1951, with this decrease being partially reversed over Australia and Europe since the early 1980s. There is substantive disagreement between data sets prior to the middle of the twentieth century, particularly over Europe, which precludes making any meaningful conclusions about DTR changes prior to 1950, either globally or regionally. Several variants that undertake a broad range of approaches to postprocessing steps of gridding and interpolation were analyzed for two of the data sets. These choices have a substantial influence in data sparse regions or periods. The potential of further insights is therefore inextricably linked with the efficacy of data rescue and digitization for maximum and minimum temperature series prior to 1950 everywhere and in data sparse regions throughout the period of record. Over North America, station selection and homogeneity assessment is the primary determinant. Over Europe, where the basic station data are similar, the postprocessing choices are dominant. We assess that globally averaged DTR has decreased since the middle twentieth century but that this decrease has not been linear.

1. Introduction

Diurnal temperature range (DTR) is a measure of the difference between daily maximum and minimum temperatures at a given location. Typically over land (which is considered herein), this is measured at 1.5 to 2 m above the surface using either manual or, more recently, automated measurement techniques. DTR has strong seasonal and geographical variations. Trends and variability in DTR have broad potential applicability, as illustrated by the following nonexhaustive set of examples. How DTR is changing has implications for our understanding of the deeper boundary layer dynamics and near-surface stability [e.g., *Christy et al.*, 2009; *Pielke and Matsui*, 2005; *Zhou and Ren*, 2011; *Parker*, 2006; *Steeneveld et al.*, 2011; *McNider et al.*, 2012]. Changes in DTR can be used as a potential mechanism to differentiate between forcings that have different short-wave and long-wave radiative fingerprints, but may otherwise be similar, and hence lead to similar expressions in mean temperature changes [e.g., *Jackson and Forster*, 2013; *Wang and Dickinson*, 2013]. DTR changes are also important for human health [*Paaijmans et al.*, 2009], ecology [*Peng et al.*, 2013; *Vasseur et al.*, 2014], and agriculture [*Battisti and Naylor*, 2009] among others.

Many of the studies referenced above used either a single homogenized DTR series estimate or some collection arising from one or more of the available disparate data archives directly, without a consideration of their homogeneity. As shown in *Thorne et al.* [2016], there is substantial evidence that the "raw" basic data contain inhomogeneities. Further, there is inevitable uncertainty in all aspects of the homogenization, gridding, and

averaging processes [*Thorne et al.*, 2005a], which means it cannot be certain that a given assessment of homogeneity and resulting data set estimate is entirely successful. It is therefore an open question to what extent conclusions of present and historical analyses using apparent observed changes in DTR are robust to observational uncertainties. The principal aim of the current paper is to assess the extent to which changes in observed DTR are well characterized by (i) analyzing available estimates and then (ii) making use of available additional experiments to ascertain sensitivity of estimates to several possible gridding and analysis processing steps.

Changes in diurnal temperature range (DTR) were recognized as a key uncertainty in the Technical Summary of the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [*Solomon et al.*, 2007]. Surprisingly, there were no subsequent dedicated studies looking solely at large-scale changes in DTR between the Fourth and Fifth Assessment Reports, although a number of data sets and analyses touched upon it. Working Group 1 of the Fifth Assessment Report concluded that there was only medium confidence in our knowledge of DTR changes [*Hartmann et al.*, 2013] given both the lack of recent studies and a number of papers which highlighted potentially distinctly divergent effects of inhomogeneities (or biases) on maximum (T_x) and minimum (T_n) temperatures [e.g., *Fall et al.*, 2011; *Williams et al.*, 2012a]. As discussed in *Thorne et al.* [2016] and many other analyses of surface records [e.g., *Lawrimore et al.*, 2011, and references therein], there are many potential reasons for inhomogeneities in station records including changes in instrumentation, siting, or recording practices.

Despite the lack of dedicated studies of DTR trends, there actually exist several recently produced or updated products, which either include DTR explicitly or from which DTR can be calculated directly:

1. DTR is one of 27 indices defined by the Commission for Climatology/World Climate Research Program/Joint Technical Commission for Oceanography and Marine Meteorology Expert Team on Climate Change Detection and Indices (ETCCDI) [*Peterson et al.*, 2001; *Zhang et al.*, 2011]. Two data sets exist that include the temperature ETCCDI indices elements (including DTR): *GHCNDEX* [*Donat et al.*, 2013a] and *HadEX2* [*Donat et al.*, 2013b]. In addition, *Donat et al.* [2013a] calculated these indices from HadGHCND [*Caesar et al.*, 2006], a data set of gridded daily maximum and minimum temperatures.
2. The high-resolution gridded product CRU TS from the Climatic Research Unit at the University of East Anglia includes among its products DTR [*Harris et al.*, 2014].
3. Changes in DTR have also been derived from the Berkeley Earth Surface Temperature product [*Rohde et al.*, 2012, 2013].
4. The companion analysis [*Thorne et al.*, 2016] describes a new DTR data set, building off the recent first release of the International Surface Temperature Initiative's databank [*Rennie et al.*, 2014] and the National Centers for Environmental Information's (NCEI) Pairwise Homogenization Algorithm [*Menne and Williams*, 2009].
5. Finally, for completeness, we also consider the analysis of changes in DTR used to inform the Intergovernmental Panel on Climate Change's (IPCC) AR4 report findings [*Vose et al.*, 2005a]. This series finishes in 2004, whereas other estimates are available through at least 2010.

The current analysis is principally intended to perform an updated assessment of the observational evidence basis for changes in DTR from the AR4 and AR5 findings. Section 2 describes the construction of each data set to the extent required to interpret subsequent intercomparisons. Differences exist in source data, homogeneity assessment, station inclusion criteria, gridding, and interpolation, all of which may be important. Section 3 describes the necessary preprocessing steps to enable an intercomparison, given the distinct grid resolutions and the availability of some products as absolute values, while others are available only as anomaly products. The intercomparisons are carried out in section 4. An assessment of potential sensitivity to gridding and interpolation choices is given in section 5. Discussion of the results and their implications is given in section 6. Section 7 concludes with an updated assessment of the current state of understanding in DTR changes based upon the results of the comparisons undertaken herein.

2. Summary of Data Sets Considered and Their Methods of Calculation

For most of the products analyzed, DTR is one of the broad suites of temperature-derived (and sometimes other meteorological element) parameters considered. For many applications, a consideration of estimates of changes in multiple meteorological elements constructed in a consistent manner is required. This was

the driving rationale behind the development of many of these products and in large part explains the lack of recent dedicated analyses of DTR changes alone. The following subsections outline in brief the methodologies used, while Table 1 provides an overview. For more technical details, readers are referred to the referenced paper(s) describing each data set. Data set names, when used, are always italicized in subsequent text to make clear that a data set is being referred to. Section 2.8 closes with a summary.

2.1. ISTI+PHA

The data set outlined in *Thorne et al.* [2016] uses the T_x and T_n monthly means series derived from the International Surface Temperature Initiative's Global land surface databank release 1.0.0 [*Rennie et al.*, 2014]. Monthly means are generated from daily series where available allowing for up to three missing days in each month. The basic holdings are just over 32,000 stations of which most have T_x and T_n for at least part of the series but only of the order of 10,000 of which are sufficiently complete to calculate a 1971–2000 climatology. In this data set, and all other data sets considered herein, the stations are far from evenly spread across the globe (see *Thorne et al.* [2016]) Figure 7—this is indicative of the density of input data across all data sets considered herein). From these monthly T_x and T_n, DTR values are calculated as monthly T_x − monthly T_n. These DTR series are then passed through NOAA NCEI's Pairwise Homogenization Procedure (PHA) [*Menne and Williams*, 2009] in the configuration used in Global Historical Climatology Network (GHCN) v3.2.0 [*Lawrimore et al.*, 2011; *Williams et al.*, 2012b]. Here we make use solely of the directly adjusted series referred to in Thorne et al. Resulting homogenized series are gridded as anomalies relative to 1971–2000 using 5° resolution bin averages with no interpolation. Prior to gridding monthly station series, DTR outliers are removed through a 5 standard deviation quality control check based upon trimmed series to avoid the effects of outliers in determining the standard deviation.

2.2. CRU TS 3.22

The *CRU TS* data set includes ten variables, one of which is DTR. Observations of T_x and T_n are updated from CLIMAT (monthly summaries), MCDW (Monthly Climatic Data for the World) and BoM (Bureau of Meteorology). DTR is derived in the usual way (T_x-T_n). The gridded products for T_x and T_n are then derived from DTR and Tm (mean temperatures). Stations with <75% of monthly values between 1961 and 1990 are excluded from interpolation, as station normals are calculated over this period. The station observations are then anomalized using the normals. Because the *CRU TS* data set must be complete over land masses (excluding Antarctica), sparse areas are infilled with the published 1961–1990 climatology (CRU CL v1.0). This happens for grid boxes and time steps where there are no stations within about 850 km of the box. Interpolation to a 0.5° grid of land surfaces then takes place using Delaunay Triangulation. Finally the gridded data are converted to actuals by the addition of the gridded climatology.

2.3. Berkeley Earth Surface Temperature

The *Berkeley* Earth data set merges temperature observations from 12 sources using metadata and data to identify duplicate time series; however, only 8 sources provide the T_x and T_n temperature observations necessary for DTR consideration. The largest source is NCEI's Global Historical Climatology Network-Daily (GHCN-Daily), which provides more than half of all stations. Individual time series are filtered by quality control algorithms that eliminate problems such as severe outliers and repeated/fill values, as well as observations flagged as suspect in the originating data set. Daily time series are averaged into monthly time series allowing up to nine missing days and resulting in about 37,000 distinct monthly station time series that include both T_x and T_n observations. Monthly DTR time series are then calculated for each station in the typical way ($T_x - T_n$). These local DTR time series are again subjected to quality control screening to eliminate severe outliers (typically greater than 4 standard deviations from seasonal expectation). As with any quality control (QC) procedure there is a risk of false positives where the underlying distribution is highly skewed, but the impact on large scales of interest herein would be minimal. The resulting time series are then homogenized and used to produce DTR fields by applying a Kriging process [*Rohde et al.*, 2013].

The homogenization approach uses a pairwise breakpoint detection algorithm, but rather than estimating and applying an adjustment at each inhomogeneity, the time series are ultimately broken into separate fragments at the detected breakpoints. These homogeneous fragments are then directly combined into spatial fields using a Kriging-based approach with integrated bias and outlier detection [*Rohde et al.*, 2013].

Table 1. Summary of Methodological Details of the Various Data Sets Considered in the Present Analysis Highlighting Similarities and Differences Between Them[a]

Data Set	Station Input Source	DTR From Dailies or Monthlies	Quality Control	Homogenization	Gridding/Area Averaging	References
ISTI+PHA	ISTI Global Land Surface Databank (expanded from GHCN-Daily)	Monthlies (but most monthly files from daily source)	NCEI QC plus 5 sigma check on station series	NCEI Pairwise Homogenization Algorithm [Menne and Williams, 2009]	Simple grid box averages of station anomalies over 5° bins, no infilling	Thorne et al. [2016]
CRU TS 3.22	CLIMAT, MCDW, and various national contributions	Monthlies	3 sigma check on station values	At source with homogenized sources preferred	Distance weighted infilling of anomalies to 0.5° resolution with climatology where no constraint and gridding of anomalies is by triangulation	Harris et al. [2014]
Berkeley Earth Surface Temperature	GHCN-Daily and seven additional sources	Monthlies (but most monthly data derived from daily sources)	Various station level checks for anomalous data, and integrated network consistency checks among neighboring data	Pairwise homogenization to detect apparent breakpoints. Series are then broken into numerous homogeneous segments whose relative biases are determined during a global Kriging process	Kriging of anomalies from station locations with a spherical correlation function that falls to zero over finite distance	Rohde et al. [2013]
HadEX2	National Met Services, GHCN-Daily, ECAD, LACAD, SACAD, HadEX, and authors	Dailies	ETCCDI recommended procedures	ETCCDI recommended procedures (f test and t test family tests)	Absolute values gridded. Grid box values are calculated as a weighted average of stations within a given radius, considering the spatial correlation structure of the point data	Donat et al. [2013b]
GHCNDEX	GHCN-Daily	Dailies	GHCN-Daily operational QC [Menne et al., 2012] and ETCCDI recommended procedures	None	As HadEX2	Donat et al. [2013a]
Vose et al. [2005a]	GHCN-Mv2 augmented with other monthly sources	Monthlies	GHCN-Monthly version 2 operational QC [Peterson et al., 1998a]	Reference series method of Menne and Williams [2005]	Simple grid box binning at 5° scale	Vose et al. [2005a]
HadGHCND	GHCN-Daily	Dailies	GHCN-Daily operational QC	None	Angular distance weighting of daily anomalies to 2.5° × 3.75° grid.	Caesar et al. [2006] and Donat et al. [2013a]

[a]Only a subset of the characteristics felt to be most pertinent have been retained here.

By design, this approach is tolerant of short time series and does not require that any specific coverage interval be populated, which permits a large portion of all observations to be used. The correlation matrix applied for Kriging is assumed to obey a spherical correlation function with empirically estimated parameters, such that the correlation falls to zero over finite distance. The expected correlation in DTR is approximately 0.5 at a station separation distance around 950 km. The initial Kriging field is sampled on an equal-area grid with approximately 1.25° resolution at the equator, and products with alternative resolutions were then derived from it.

An alternative DTR field was also produced that interpolated and homogenized T_x and T_n separately and computed their difference. This alternative approach produced very similar results but was found to have somewhat larger calculated uncertainties, probably because calculating DTR at a station level allows for the partial cancelation of homogenization biases that affect both T_x and T_n with the same sign. This alternative field was not used in the present analysis.

2.4. HadEX2

HadEX2 contains global (land-based) grids of the ETCCDI climate indices including DTR. It is based on daily station observations from several different data sources, including freely accessible data archives, ETCCDI workshops held in targeted regions in order to calculate the climate indices when daily data are not shared, and also personal contacts in some countries [*Donat et al.*, 2013b]. Within *HadEX2*, monthly and annual DTR, calculated as the average of daily $(T_x - T_n)$ DTR values, is available from more than 7000 stations globally. The indices are first calculated for all station time series before the actual DTR values are interpolated on a global regular 3.75° × 2.5° (longitude × latitude) grid.

For calculating the gridded fields a modified version of Shepard's angular distance weighting (ADW) algorithm [*Shepard*, 1968] is employed. ADW considers the spatial correlation structure of the data in order to define a search radius within which stations contribute to the calculation of the grid box values. To this end, correlation values between all station pairs within a 2000 km radius are calculated if their data overlap for at least 30 years. The interstation correlations are then averaged into 100 km bins, and a second-order polynomial is fitted to the data points, constrained to a correlation equal to 1 at zero distance. The decorrelation length scale (DLS) is then defined as the distance at which the fitted correlation function falls below $1/\exp(1)$. This DLS is used as a search radius in which station data are considered for the calculation of grid values.

In a following step, grid box values are calculated as a weighted average of all stations within the DLS (measured from the grid box center), with weights decaying exponentially with increasing distance. Grid box values are only calculated if at least three stations are located within the DLS, otherwise, a missing value is assigned.

2.5. GHCNDEX

GHCNDEX, similar to *HadEX2*, provides gridded fields of the temperature and precipitation indices recommended by the ETCCDI. Largely the same methods as for *HadEX2* are used, with the only differences being that a different station data base is used, and fields are interpolated on a different grid resolution. *GHCNDEX* is based solely on stations that are available in the GHCN-Daily archive [*Menne et al.*, 2012], which enables regular updates and makes *GHCNDEX* a suitable tool not only for investigating long-term changes but also for climate monitoring purposes [*Donat et al.*, 2013a]. For the analyses presented in this study, we use the *GHCNDEX* version as of August 2013.

In *GHCNDEX*, monthly mean DTR time series are first generated at the station level. A month's mean DTR is computed as the average of daily DTR values which, in turn, are calculated as the difference between a day's maximum and minimum temperatures. If a day's maximum or minimum temperature is either missing or flagged by the GHCN-Daily quality control process [*Durre et al.*, 2010], that day is not included in the corresponding monthly mean of DTR. *GHCNDEX* therefore, like *HadEX2*, represents a product in which the differencing between daily maximum and minimum temperatures is performed as early in the processing as possible, namely, at the station level and on a daily timescale.

To minimize effects from variations in the station network, only GHCN-Daily stations that provide at least 40 years of near-complete data after 1951 are used to calculate the *GHCNDEX* gridded fields. For DTR, about

4750 stations fulfill this criterion. The same ADW gridding method is used as for *HadEX2*, but *GHCNDEX* gridded fields are calculated on a 2.5° × 2.5° resolution.

2.6. HadGHCND

DTR fields were also calculated from the *HadGHCND* gridded data set of maximum and minimum daily temperatures [*Caesar et al.*, 2006; *Donat et al.*, 2013a]. As with *GHCNDEX* it uses the daily station observations from GHCN-Daily, but in this case the temperature anomalies are gridded first then DTR is calculated from the gridded values. The GHCN-Daily quality control flags [*Durre et al.*, 2010] are used to remove potentially spurious values from the interpolation process. Approximately 7600 stations were selected for which there are adequate available data to calculate a 1961–1990 daily climatology. A maximum of 10 stations are used in each grid box, so the final station count may be lower due to thinning in data rich regions.

An angular distance weighting approach is used to interpolate the station data onto a 3.75° by 2.5° grid. Initially, the daily station anomalies (relative to 1961–1990 and based upon a 5 day window) are gridded. Absolute temperature grids were created by gridding the 1961–1990 normals for each day of the year and adding these to the relevant daily anomaly grids. Finally, the temperature-related ETCCDI indices, including DTR, were calculated from the gridded absolute temperature values of *HadGHCND* [*Donat et al.*, 2013a].

2.7. Vose et al. 2005 Data Set

The analysis of *Vose et al.* [2005a] employed station data from 20 source data sets, including the Global Historical Climatology Network version 2 [*Peterson and Vose*, 1997], World Weather Records, and several national collections. Station data were quality assured as in *Peterson et al.* [1998a] then adjusted for historical changes in station location, temperature instrumentation, and observing practice via the reference series approach of *Menne and Williams* [2005]. The DTR time series for each station was computed by subtracting the adjusted minimum series from the adjusted maximum series. Grid box time series were calculated by applying the climate anomaly method to an input network of approximately 4200 stations that had sufficiently complete base-period (1961–1990) data.

2.8. Summary

The different products considered herein incorporate different sets of stations, derive DTR in different ways (from daily readings or monthly averages), perform gridding/averaging in distinct manners, and undertake varying degrees of quality control and/or homogenization upon the basic station series. These methodological differences permit an assessment of the structural uncertainty arising from the range of possible methodological pathways [*Thorne et al.*, 2005a] inherent in DTR estimates. Such an assessment is predicated upon assumptions regarding both the efficacy of each method and that the range of estimates adequately samples the broad range of potential approaches to all processing steps from station selection through to gridding and interpolation.

If one choice was dominant, then it would be expected that the estimates clustered based upon that choice. Hence, analysis in section 4 should be considered in the context of the summary in Table 1 that highlights the choices at each major processing step by each group. For example, if the choice of source data was important, then there are several products that either rely exclusively upon GHCN-D (*GHCNDEX*, *HadGHCND*) or for which it is the major source component globally (*ISTI+PHA*, *Berkeley*) or regionally (*HadEX2*) (Table 1, column 2, earlier subsections). If the choice of source station data was the dominant determinant of resulting estimates, then it would be expected that at least for those four data sets for which it is the primary global source, resulting estimates would cluster. If, instead, differences arise from the suite of processing steps with no single, dominant, choice, then no single processing step choice (column in Table 1) shall obviously dominate.

3. Data Preprocessing to Enable Intercomparisons

The various data sets are available at resolutions varying from 0.5° to 5° with some of the HadEX family of products (*HadEX2*, *HadGHCND*) being on a grid that has unequal grid spacings in latitude and longitude. Furthermore, those data sets using ETCCDI indices are available only as absolute values, whereas the remaining products are available either as anomalies only or anomalies and absolutes. Even when available as anomalies, the climatology period differs between the products. Additionally, a number of products consist of interpolated records when infilling of regions of missing data occurs either solely within a finite distance of actual data values or to create globally complete fields for land areas. To enable an intercomparison of these estimates,

preprocessing steps are therefore required to enable appropriate comparisons, including as close to like-for-like comparisons as practical.

First, all DTR gridded fields have been regridded to the coarsest 5° regular latitude/longitude gridded resolution. Regridding has been performed by simple averaging of the finer-resolution gridded products, using grid box centers of the finer resolution products as a basis, and bin averaging. While this may have some impact on individual grid boxes compared to more complex methods such as bilinear interpolation, we are not here primarily interested in individual grid box series but rather spatial behavior and large regional averages, which will be preserved by any reasonable choice of regridding approach given the spatial scales of DTR anomalies. *Harris et al.* [2014] report a typical correlation scale (where the correlation is >1/e) for monthly average DTR of 750 km, and in *GHCNDEX* at annual scales it is between 450 and 1200 km (latitudinally dependent). Further, several interpolated products are already spatially smoothed to some extent as a result of the interpolation (see section 5).

Next, new normals and anomalies were calculated relative to a common climatology period of 1971–2000. The climatology criterion is two thirds of monthly data (at this stage all products are monthly) in the 30 year period taken as a whole and at least one half in any given decade (1971–1980, 1981–1990, and 1991–2000). This criterion has been applied to each calendar month in turn and to the gridded series. There is an inevitable trade-off, particularly, for the products that do not undertake infilling between climatological period completeness and eventual coverage. All data sets using in situ measurements and a climate anomaly method approach undertake some similar criteria that retain data based upon completeness across the period and each subperiod. These criteria are similar to those employed in, e.g., CRUTEM4 [*Jones et al.*, 2012] or HadAT [*Thorne et al.*, 2005b].

To enable comparisons that account for the effect of differences arising from infilling/interpolation/data completeness, two versions of each regridded and renormalized data set have been created:

1. The data set at its native data completeness, henceforth, is referred to as "native data set coverage."
2. The data set sampled only at those locations where all data sets that are available at a given time step have a data value. Over 1951–2004, all data products are available; outside of this period, the available data sets compose solely a subset of products (Table 1). Henceforth, this is referenced as "common data set coverage."

Finally, a number of regional averages have been determined. The global average is simply the grid box area weighted average for all available grid boxes in any given month and product. Given the large changes in observational data availability, particularly in products that do not perform interpolation, and prior to 1960, this should not necessarily be interpreted as a true global mean. Three additional areas have been defined. These areas were determined in *Thorne et al.* [2016] based upon data availability since the early twentieth century in the ISTI databank release. The regions considered herein are defined 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). Given the availability of average temperatures in many regions when DTR is not available, T_x and T_n data may also have been measured, but either has not been digitized or has not been shared [*Rennie et al.*, 2014]. If these data can be rescued, then future assessments of DTR could consider additional regions. It is important to note that the reporting of T_x and T_n has been common in English-speaking countries but less common elsewhere (see discussion in *Thorne et al.* [2016]). Elsewhere, and also in the very earliest records in English-speaking countries, available reports may instead be observations taken at regular times of day. Although T_x and T_n may be able to be estimated from such measurements, they would bring additional uncertainties.

Wherever trends have been calculated in sections 4 and 5, these have used the method employed in *Hartmann et al.* [2013]. Trends are calculated using ordinary least squares regression accounting for AR(1) autocorrelation effects (AR(1) estimated from the longest continuous segment for discontinuous series) in assessing significance. Trends are only calculated where data are at least 70% complete and contains sufficient data (30%) in the first and last deciles of the series. Trends are always calculated on the annual mean series, and each season must be represented to calculate an annual mean.

4. Data Sets Intercomparison

First, an assessment of structural uncertainty (uncertainty arising due to choices of methodology) in spatial trend patterns is undertaken in section 4.1. To fully sample the solution space, this comparison of spatial trends is limited to the common period of record for all seven data sets (1951 to 2004) and the subset of these
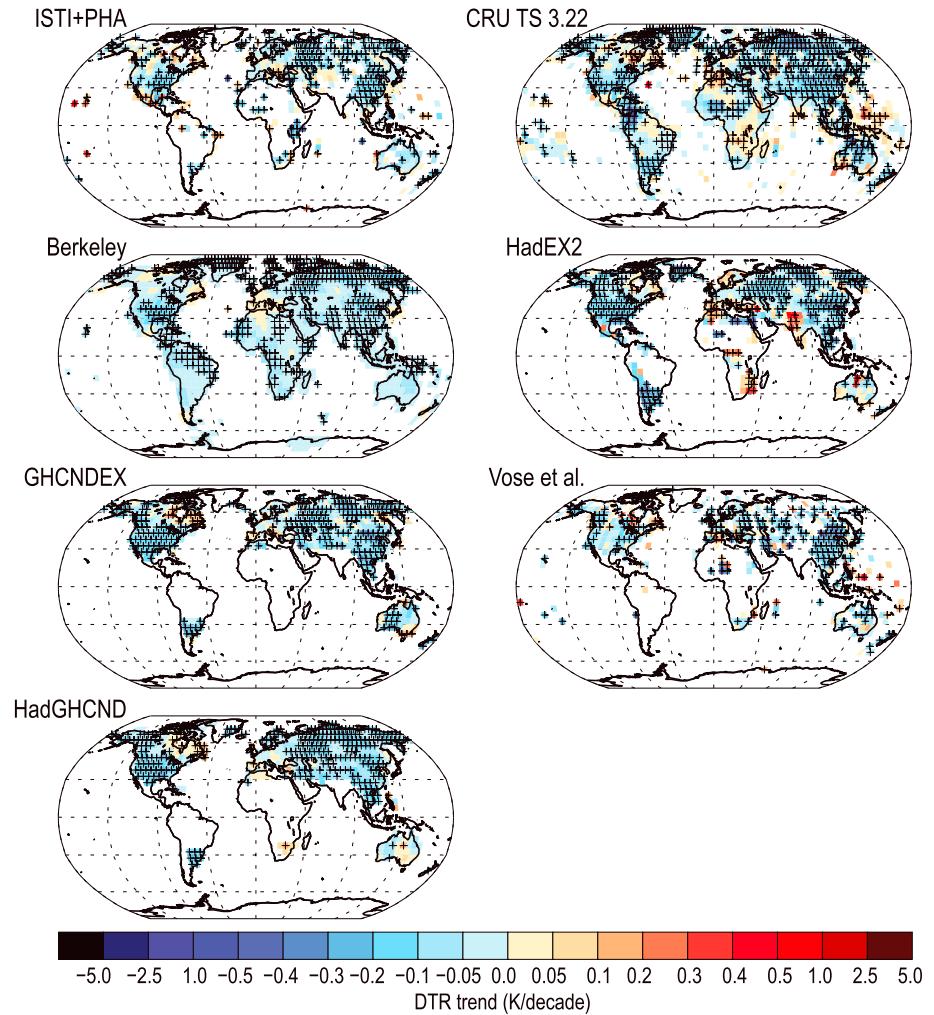
Figure 1. Grid box linear trends in diurnal temperature range estimated using ordinary least squares (OLS) with AR(1) degrees of freedom (d.f.) correction from 1951 to 2004 for the seven different data sets at their native coverage. Only grid boxes with >70% data availability within the period and some data in the first and last deciles of the period are included. Trends which are significantly different from zero are denoted by a cross.

records since 1979 (the advent of satellite records and hence a subperiod often used in assessments) [e.g., *Hartmann et al.*, 2013]. Hence, all data sets cover the same period. Using their full periods would result in different data sets ending in different years and hence conflate data set differences with real physical effects arising from consideration of quasi-distinct periods. In the main text, plots considering geographical trends at the native data set coverage are considered. Analogous plots for common data set coverage are available in the supporting information. Global and regional series for the entire period of record for each respective data set are next shown and discussed in section 4.2.

4.1. Spatial Trends

Over the full common period of record starting in 1951, the different data sets are in substantive agreement that DTR has decreased over most of the global land surface (Figure 1). Despite unanimity in the overall global tendency, differences exist between the products with regard to the spatial patterns, the rate, and the geographical extent over which the negative trends are statistically significant, relative to a null hypothesis of no trend in DTR.

Differences in station selection and the degree, if any, to which interpolation is undertaken (section 2, Table 1), lead to substantive differences in resulting coverage. The two products which are not based on interpolated data (*ISTI+PHA* and *Vose et al.*) have the lowest coverage. However, it would be incorrect to infer that the other
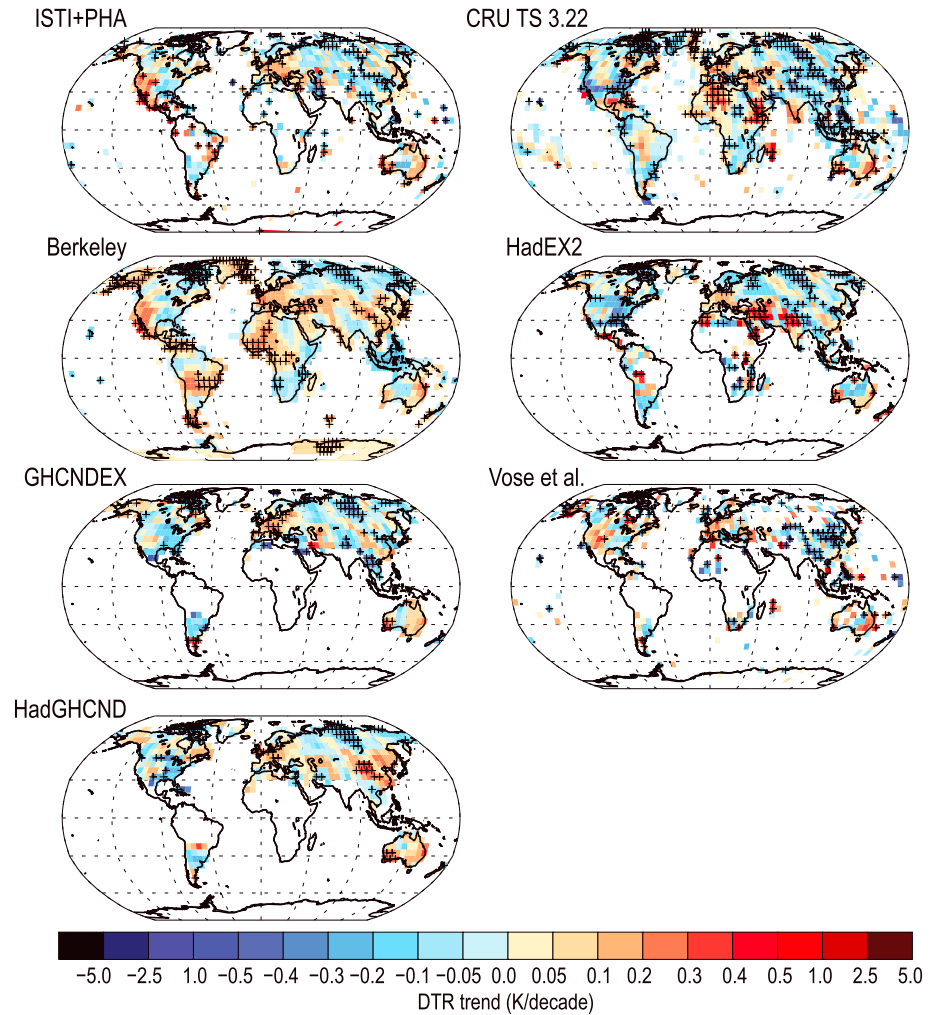
Figure 2. As in Figure 1 but for trends calculated over the period 1979–2004.

products grow from a similar coverage. Specifically, *HadEX2* benefits from data shared on a basis of use to derive indices and through a sustained series of regional workshops under the auspices of the ETCCDI [e.g., *Peterson and Manton*, 2008]; *CRU TS* benefits from a number of National Meteorological Service (NMS) holdings in addition to those available at NCEI, and *Berkeley* uses many short period records not used in *ISTI+PHA* or *Vose et al.* As *ISTI+PHA* is based upon a databank that builds from GHCN-Daily, the two exclusively GHCN-Daily-based products use a subset of the *ISTI+PHA* station basis.

Trends are much more similar if common data set coverage (section 3) is considered (Figure S1 in the supporting information, cf. Figure 1; see section 4.2 for regional series and trend analysis under native data set coverage and common data set coverage). The different products are in general agreement where there are many stations available to construct gridded estimates (North America, Eurasia, and parts of Australia). Over Africa, South America, India, Antarctica, and Greenland (for those data products that produce estimates for these regions at all), the products are not generally in agreement, in particular, with regard to the spatial details. Over much of Africa and India, even the sign of the changes is uncertain for that subset of products that provides estimates.

Over the period starting in 1979 trends are far more spatially heterogeneous (Figure 2), as would be expected from a consideration of signal to noise properties [*Santer et al.*, 2011]. There is no longer a clear preponderance toward a decrease in DTR, with most data sets showing roughly as many regions exhibiting increases in DTR as exhibiting decreases. The implication from this is that over much of the globe the reduction in DTR experienced since the middle twentieth century was substantively over by the early 1980s, as concluded in IPCC AR4 [*Trenberth et al.*, 2007] (see also section 4.2).
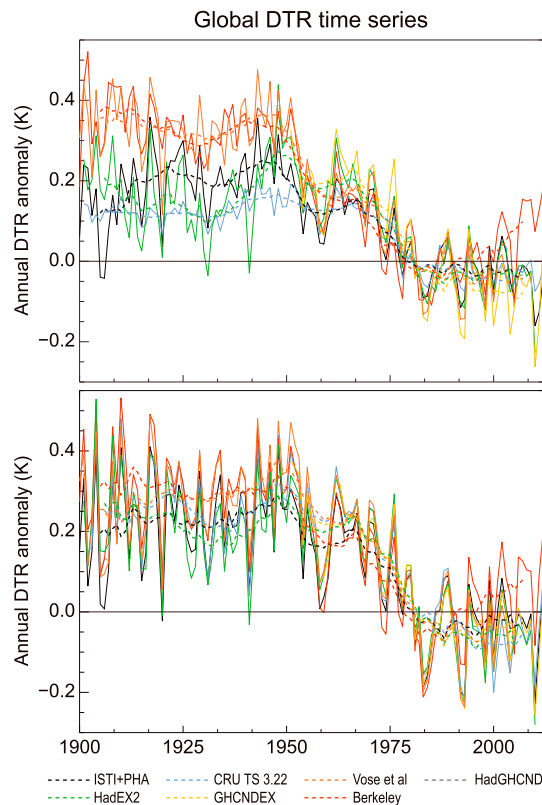
Figure 3. Global anomaly time series for each data set relative to a common climatology period of 1971–2000. Global averages have been calculated from a cos(lat) weighted average of reporting grid boxes at each time step. Averages have been calculated at the native coverage for each data set in the top panel and at common coverage in the bottom panel. Thin traces denote annual mean series while thick dashed lines denote moving decadal averages.

Again, there are substantial differences between data sets over this shorter period, and these differences are once more greatest in regions that are generally poorly sampled. Considering common data set coverage (Figure S2) highlights substantive differences over much of Central Asia and the western half of North America (also apparent in Figure 2). Over central Asia, *HadGHCND*, *Berkeley*, and to a lesser extent *GHCNDEX* show increases in DTR, whereas the remaining products show decreases. This is particularly acute over northern China where, depending upon the choice of data product, the DTR could be found to be either significantly increasing or significantly decreasing. There are regional products over China [*Li et al.*, 2009] and India making use of many more stations, but data restrictions preclude their use in improving the global products, which are the focus herein. Over western North America there is a clear split with *ISTI+PHA*, *Vose et al.*, and *Berkeley* all suggesting increasing DTR and the remaining products suggesting little change to a slight decrease (very few grid boxes show significant changes in any of the data sets). Conversely, the data sets all agree that DTR has been increasing over much of Europe and Eastern Australia since the early 1980s.

4.2. Areal Average Trends and Time Series Behavior

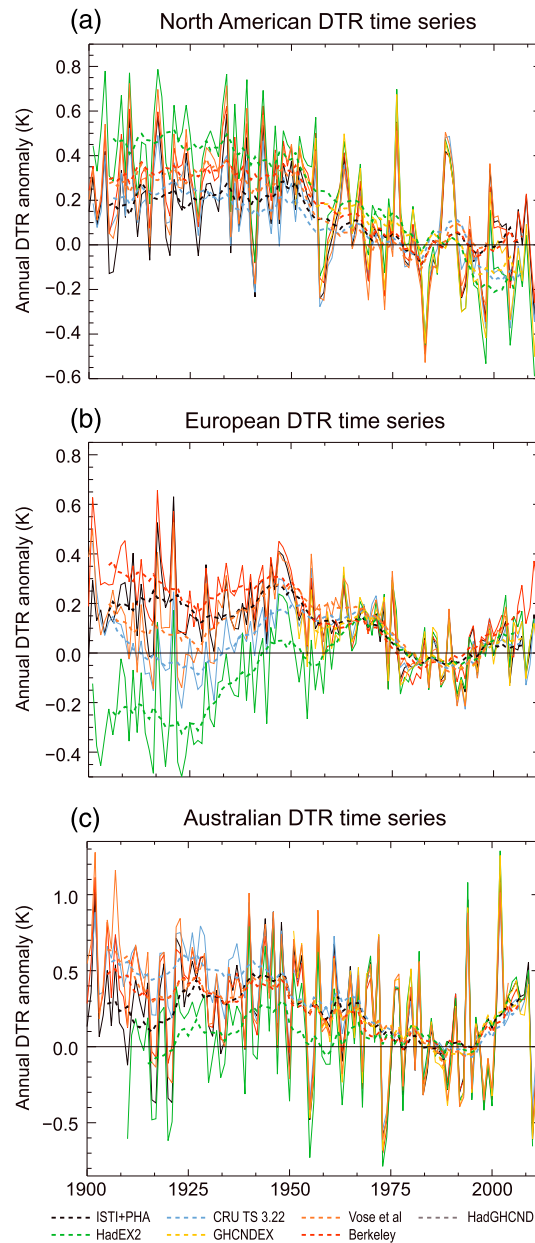At native data set coverage, the data sets all agree that over the period spanning roughly 1960 to 1980 the globally averaged DTR decreased by of the order 0.15–0.25 K (Figure 3, top). In the period following this, most data sets imply little change in DTR, although *GHCNDEX* exhibits a continued slow decline and *Berkeley* suggests a marked rebound in the last 15 years. Estimates prior to 1960 exhibit a broad spread, and several data sets exhibit distinct decadal timescale variations in this early period. We note that similarity over 1971–2000, in this and all similar figures (Figures 4, 5, 8, 9, and S3), will to some extent result from the use of this period as a common climatology in the preprocessing described in section 3 [*Hawkins and Sutton*, 2016].

Restricting to common data set coverage yields substantially improved inter-data set coherence, particularly prior to 1960 (Figure 3, bottom). The largest change is to *CRU TS*. This product is globally complete but, as clearly stated in its data set descriptor, infills with climatology when there is no observational constraint sufficiently close enough [*Harris et al.*, 2014] (Table 1). As would be expected, this choice significantly impacts both the mean (toward zero anomaly) and the variability (reduced) using native data set coverage compared to a common data set coverage analysis (compare top and bottom of Figure 3). The *CRU TS* includes a data flag, and if analysis is restricted to those grid boxes where a data constraint is indicated in the netcdf data file, then the results are almost identical to applying the common data set coverage analyses herein (not shown for brevity). The effect therefore arises from infilling with climatology at points distal from a data constraint.

The common data set coverage version of *CRU TS* is in far better agreement with the behavior evident in the remaining estimates. However, while *CRU TS* has the most substantive change, it is noticeable that all the interpolated products exhibit pronounced changes in time series smoothness and even decadal timescale

Figure 4. As in Figure 3 but for (a) North America, (b) Europe, and (c) Australia. See section 3 for derivation of region bounds. Time series are at the native spatial completeness for each product and constitute the area weighted average of contributing grid boxes within the region at each time step. Temperature axes ranges differ by region but are identical between Figures 4 and 5 for ease of comparison.

characteristics with masking to common coverage. This yields a question as to the efficacy of estimates in interpolated regions whether the interpolation be statistically or climatologically based, as well as the suitability of such grown fields when making global series estimates. Whether this finding extends to similar interpolation efforts for average temperatures is out of scope for the present study, but logically it must to at least some extent. It is also, however, the case that restricting *ISTI+PHA* to the common mask (effectively *Vose et al.*) has large impacts prior to 1950. The effects of common data set coverage are hence hard to uniquely disentangle.

When subsampled to common data set coverage, data sets agree on a further period of decreasing DTR from the late 1940s for about a decade and that prior to this global DTR may have increased slightly. There remains disagreement between estimates in post-1980s behavior. This lack of cancelation of differences in the most recent period likely arises largely because most data products are better sampled in this period, so enforcing subsampling has much less effect on late series behavior than early series behavior when proportionately more of the field consists of infilled estimates rather than primary data. The post-1980 disagreements between data sets most plausibly relate to station inclusion criteria or homogeneity assessment choices (many stations have changed from Stevenson screens to (semi-)automated approaches).

Regional comparisons at native data set coverage reveal even greater spread than global comparisons (Figure 4, cf. Figure 3, top). The spread is particularly marked for the European region prior to 1960, with disagreement on even the sign of the DTR anomalies before circa 1940. The *HadEX2* product exhibits the greatest degree of divergence from remaining estimates prior to 1960 in all three regions. In North America, it is consistently greater and in the other two regions consistently lower. Agreement between data sets is best post-1960, although in North America there is a clear divergence between *ISTI+PHA*
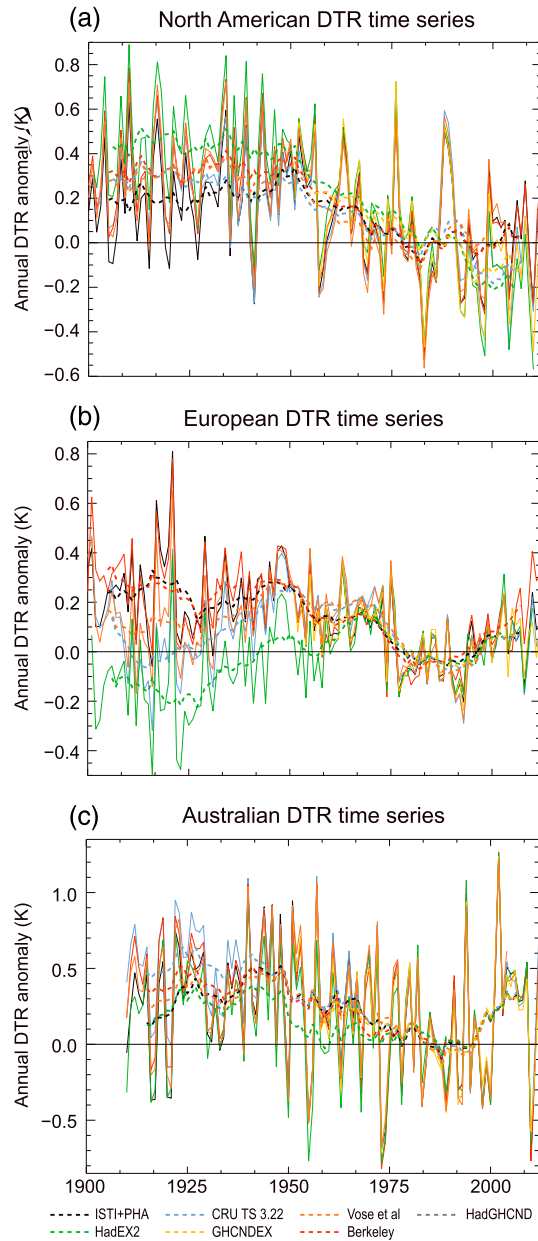
and *Berkeley* compared to the remaining estimates post-1980, with the first two suggesting little change, and the remainder exhibiting a continuing decline. The behavior of the first two would be consistent in timing and sign with the adjustment required to account for the widespread migration to Maximum Minimum Temperature Sensors (MMTS) in the United States Cooperative Observer Program (COOP) network, which various studies have shown to impart a spurious narrowing in DTR of the order 0.2 K [*Fall et al.*, 2011; *Williams et al.*, 2012a, and references therein]. Both data sets would have explicitly accounted for this in their homogenization assessment, whereas it is less clear that the remaining products considered would have done so (section 2, Table 1).

Figure 5. As in Figure 4 but masked to common data availability.

These three regions were chosen because there is reasonably complete coverage in the International Surface Temperature Initiative's databank for them back to the early twentieth century [*Thorne et al.*, 2016]. Other data holdings that form the basis for the remaining products share this station data mask characteristic, except that *HadEX2* does not consider Australasian data prior to 1910. Therefore, subsetting to common data set coverage has more limited impact on the apparent regional average series in these well-sampled region than is the case for a global average where the effects of interpolation and gridding choices can be disproportionately larger (Figure 5 cf. Figure 4 and compare to the two panels in Figure 3).

Trends for the global and regional averages over 1951–2004 and 1979–2004 across all data sets are summarized in Table 2. Reduction to common data set coverage has little effect except for the *CRU TS* product, where the decreasing trends globally become substantially stronger. As discussed previously, this likely relates to the choice of use of infilling with climatological values when there is no data constraint. This serves to damp both trends and variability if used naively, without due consideration of the flagging of availability of a data constraint. Regardless of whether the products are considered at native data set coverage or common data set coverage, all data sets agree that there has been a significant reduction in DTR globally over 1951–2004. The *ISTI+PHA* and *Berkeley* data sets exhibit substantially more muted reductions in DTR than the remaining data products since 1951. These products have both been homogenized using automated techniques under neighbor-based comparisons, but given other potential confounding distinctions, it is not clear whether this is the likely cause except, perhaps, over North America where the sign and timing of divergence is strongly correlated with the known artifact arising from MMTS transition. Over the shorter period starting 1979, trends are not significant in any data set and are of mixed sign.

Regionally, there is less coherence between the different data sets regarding the trends in DTR and their significance (Table 2). In none of the three regions or two periods considered are the data sets consistently indicating significant trends. The closest is over Europe from 1951 to 2004, when all except *HadEX2* indicate significant reductions in DTR. Over this domain, *HadEX2* diverges from the remaining estimates prior to the early 1960s (Figures 4 and 5). There is an agreement on the sign, but not the statistical significance, of DTR trends at the regional level on the following aspects: a decrease in DTR over North America over 1951–2004 and an increase in DTR over 1979–2004 in Europe and Australia. For all other cases there is a disagreement in even the sign of the regional trends between the various data sets.

Table 2. Trends in K/Decade Returned for the Globe and Different Regions Defined in Section 3 for Different Data Sets and for Two Common Periods 1951–2004 and 1979–2004[a]

Domain	Period	ISTI+PHA	CRUTS3.22	Berkeley	HadEX2	GHCNDEX	Vose et al.	HadGHCND
				Common data set coverage				
Global	1951–2004	**−0.058**	**−0.087**	**−0.052**	**−0.076**	**−0.079**	**−0.088**	**−0.072**
	1979–2004	0.004	−0.050	0.052	−0.027	−0.016	−0.019	0.035
North America	1951–2004	**−0.054**	**−0.069**	−0.063	**−0.107**	**−0.085**	−0.041	**−0.102**
	1979–2004	0.033	−0.091	0.028	−0.116	−0.062	0.056	−0.078
Europe	1951–2004	**−0.035**	**−0.052**	**−0.030**	−0.006	**−0.039**	**−0.052**	**−0.036**
	1979–2004	0.059	0.055	**0.084**	0.054	0.034	0.062	0.056
Australia	1951–2004	−0.053	−0.052	−0.040	0.004	−0.043	−0.058	−0.002
	1979–2004	0.093	0.095	0.071	0.044	0.097	0.107	0.125
				Native data set coverage				
Global	1951–2004	**−0.042**	**−0.048**	**−0.046**	**−0.061**	**−0.075**	**−0.066**	**−0.067**
	1979–2004	0.001	−0.020	0.028	−0.005	−0.035	−0.012	0.015
North America	1951–2004	−0.034	−0.043	−0.053	**−0.100**	**−0.072**	−0.034	**−0.083**
	1979–2004	0.042	−0.068	0.028	−0.112	−0.055	0.035	−0.055
Europe	1951–2004	**−0.036**	**−0.042**	**−0.026**	−0.000	**−0.034**	**−0.049**	**−0.036**
	1979–2004	0.033	0.035	**0.080**	0.048	0.026	0.058	0.046
Australia	1951–2004	−0.045	**−0.068**	−0.040	0.010	−0.044	−0.049	−0.018
	1979–2004	0.086	0.022	0.039	0.031	0.080	0.076	0.105

[a]The top set is when the data sets have been spatially matched. The bottom set is estimates at the data set native data completeness. Trends have been calculated using ordinary least squares regression. Confidence has been evaluated after correcting for first-order autocorrelation effects after *Santer et al.* [2008]. Trends that are significantly different to zero at the two-tailed 90% confidence interval are given in bold.

To examine the practical importance of differences between the data sets, linear trends and significance have been further assessed based upon inter-data set difference series (differences between pairs of global-mean series) (Table 3). As is visually obvious from Figures 3 to 5, the series exhibit strong interannual coherence in their variations. Taking the difference series removes this covariability resulting in a series that is more linear in nature with greatly reduced standard deviation and hence permits a stronger assessment of potential consistency than comparing their individual trends and confidence intervals. Assessed in this manner, 5 comparisons at common data set coverage and 11 comparisons at native data set coverage (in both cases out of 21 potential comparisons) indicate significant inter-data set global-mean time series trend differences. The data sets appear to be in two broad groupings: *ISTI+PHA* and *Berkeley* in one group and *HadEX2*, *GHCNDEX*, *Vose et al.*, and *HadGHCND* in the other. When considering common data set coverage *CRU TS* is similar to the latter group, whereas at native data set coverage it is more similar to the former grouping. Again, the simplest explanation of this grouping would be to attribute it to the fact that *Berkeley* and *ISTI+PHA* apply neighbor-based automated homogenization methods. However, the potential confounding effects, and that *CRU TS* agrees so well when considering common data set coverage, complicate such a clean analysis.

5. Assessment of the Potential Roles of Station Selection and Reasonable Gridding and Interpolation Methodology Differences

As highlighted in sections 4.1 and 4.2, there exists substantial spread in published DTR data set estimates, particularly acutely prior to 1960, both globally and regionally. Even after 1951, there are in many cases statistically significant differences between the data sets. Because the data set producers make very varied choices for all aspects of data selection, processing, and postprocessing (Table 1), a clean interpretation of the causes of any differences is not possible, as was noted repeatedly within section 4.

To assess the potential sensitivity to postprocessing choices on the station series alone (i.e., ruling out changes in station basis and homogenization as causes) and to station selection and homogenization, further investigation of *HadEX2* and *ISTI+PHA* is undertaken here using a number of variants (Table 4). The same postprocessing to a common resolution and climatology as described in section 3 has been applied to all these variants to enable exactly the same comparisons as were undertaken in section 4.

We note that a similar analysis to that undertaken within this section was undertaken for global-mean surface temperature data sets by *Vose et al.* [2005b]. This earlier study served to highlight that these postprocessing

Table 3. Linear Trend Estimates in K/Decade From the Time Series of Differences Between Global Annual Means for Pairs of Data Sets (Data Set Named in Row Minus That Named in Column) Over 1951–2004—The Longest Common Period of Record[a]

Data Set	ISTI+PHA	CRUTS3.22	Berkeley	HadEX2	GHCNDEX	Vose	HadGHCND
			Common data set coverage				
ISTI+PHA		**0.029**	−0.006	**0.018**	**0.021**	**0.030**	0.014
CRUTS3.22	**−0.029**		−0.035	−0.011	−0.008	0.001	−0.016
Berkeley	0.006	0.035		0.024	0.027	**0.036**	0.020
HadEX2	**−0.018**	0.011	−0.024		0.003	0.012	−0.005
GHCNDEX	**−0.021**	0.008	−0.027	−0.003		0.009	−0.008
Vose	**−0.030**	−0.001	**−0.036**	−0.012	−0.009		−0.016
HadGHCND	−0.014	0.016	−0.020	0.005	0.008	0.016	
			Native data set coverage				
ISTI+PHA		0.006	0.004	**0.019**	**0.033**	**0.024**	**0.024**
CRUTS3.22	−0.006		−0.002	**0.014**	**0.027**	**0.018**	0.019
Berkeley	−0.004	0.002		0.016	**0.029**	**0.020**	**0.021**
HadEX2	**−0.019**	**−0.014**	−0.016		**0.014**	0.005	0.005
GHCNDEX	**−0.033**	**−0.027**	**−0.029**	**−0.014**		−0.009	−0.008
Vose	**−0.024**	**−0.018**	**−0.020**	−0.005	0.009		0.001
HadGHCND	**−0.024**	−0.019	**−0.021**	−0.005	0.008	−0.001	

[a]Trends have been calculated using OLS and are given in bold where significant at 90% confidence interval. Top set is for common data set coverage, and lower set is for native data set coverage. In each set the values are asymmetric (flipped in sign but otherwise identical) about the diagonal. Values along the diagonal would be zero by construction but are omitted for clarity.

choices, rather than the underlying station data and their adjustment, were the likely dominant causes of differences between the then current versions of CRUTEM, GHCN, and GISS products for T_m in recent decades. Given the increased similarity between data sets when common data set coverage is applied in section 4, this would also appear to potentially be the case for DTR.

The station set underlying *ISTI+PHA*, which uses a climate anomaly method (CAM) and no infilling, was also gridded using the *HadEX2* method of angular distance weighting (termed *ISTI+PHA-HadEX*). This grids absolute values and performs limited interpolation into regions without data and smoothing of fields, by calculating grid box values as a weighted average of stations within the search radius (section 2, Table 1). For *HadEX2*, the substantive uncertainty analysis of *Dunn et al.* [2014] provides a broad range of variant gridded DTR products, all starting from the *HadEX2* input stations that make broad-ranging decisions regarding station inclusion, anomaly calculation method, gridding method, and interpolation approach for all 27 ETCCDI indices. Here use is made of a relevant subset of these realizations for DTR. In addition to the default *HadEX2* version the following variants are considered (see Table 4 or subsequent text for a further description of methods): (1) a version gridded solely from stations with a period of record exceeding 80 years

Table 4. Summary of Several Facets of the Eight Variants Considered in Section 5 Prior to the Preprocessing Steps Described in Section 3 Being Applied[a]

Variant	Station Basis	Gridding Method	Anomalies or Absolutes
ISTI+PHA	ISTI databank v1 stations with sufficient data 1971–2000	Bin average	Anomalies relative to 1971–2000 climatology
ISTI+PHA-HadEX	ISTI databank v1 all stations	Angular distance weighting	Absolutes
HadEX2	HadEX2 all stations	Angular distance weighting	Absolutes
CAM	HadEX2 stations with 25 years data during 1961–1990	Bin average	Anomalies relative to 1961–1990 climatology
FDM	HadEX2 all stations	Bin average	Anomalies calculated from cumulative sums forward in time (relative to 1901)
FDM_rev	HadEX2 all stations	Bin average	Anomalies calculated from cumulative sums back in time (relative to 2010)
Longstations	HadEX2 stations >80 years duration	Angular distance weighting	Absolutes
RSM	HadEX2 all stations	Infilling, using a decorrelation length scale as determined from the angular distance weighting	Absolutes

[a]For further details of the 6 HadEX2 variants (last 6 rows) see *Dunn et al.* [2014].
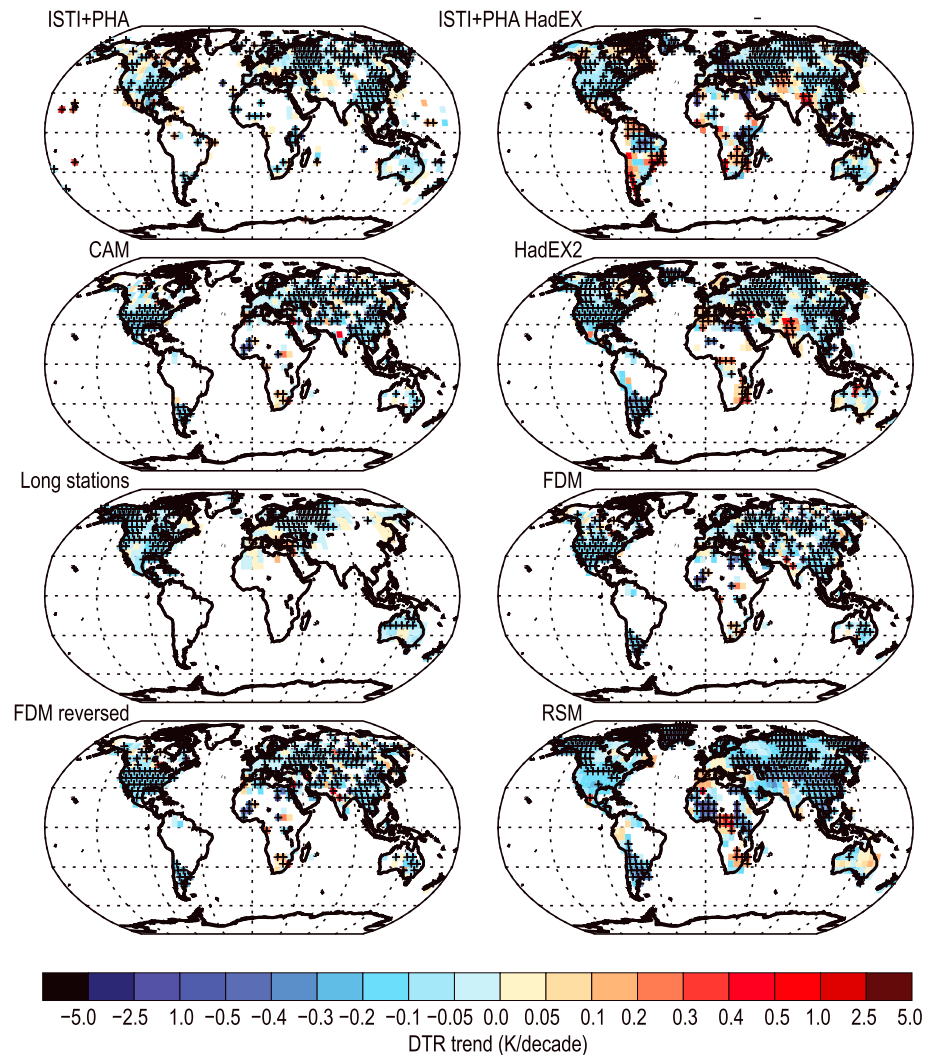
Figure 6. Spatial trends and significance assessment over 1951–2004 calculated as in Figure 1 for two flavors of the ISTI +PHA data set ((top left) STI+PHA with CAM and (top right) ISTI+PHA-HadEX with HadEX processing) and six flavors of HadEX2 (with CAM, standard HadEX2 processing, Longstations, FDM, FDM_rev, RSM). See Figure 1 caption for further details of plotting methodology.

(*longstations*), (2) a version gridded using *CAM* (which can be compared to *ISTI+PHA*), (3) a version gridded using the cumulative summation First Difference Method (*FDM*) and also a version using the same method in reverse (*FDM_rev*), and (4) a version using a Reference Station Method (*RSM*) with limited linear interpolation.

There are therefore a total of six *HadEX2* variants where the station basis is the same but the postprocessing choices that lead to station selection, gridding, and any interpolation vary substantively (Table 4). It should be stressed that these variants include methodological approaches not employed in the data sets considered within section 4, and hence, this analysis is complementary to that undertaken therein. Specifically, none of the data sets considered in section 4 employ *FDM* or *RSM*, and restricting to 80 years is a more restrictive station selection criteria than employed by any group in constructing a data set in section 2. However, they were all analyzed in *Dunn et al.* [2014] and, therefore, are deemed plausible choices to the gridding and postprocessing steps herein.

HadEX2 and the *longstations* variant differ only in the station period of record criteria for inclusion, with the *longstations* variant requiring 80 years record length for a station to be considered and *HadEX2* considering all stations in the underlying database irrespective of station length. The remaining four Dunn et al.-based methods all build the gridded product in a different manner from the *HadEX2* station database. The *CAM*,

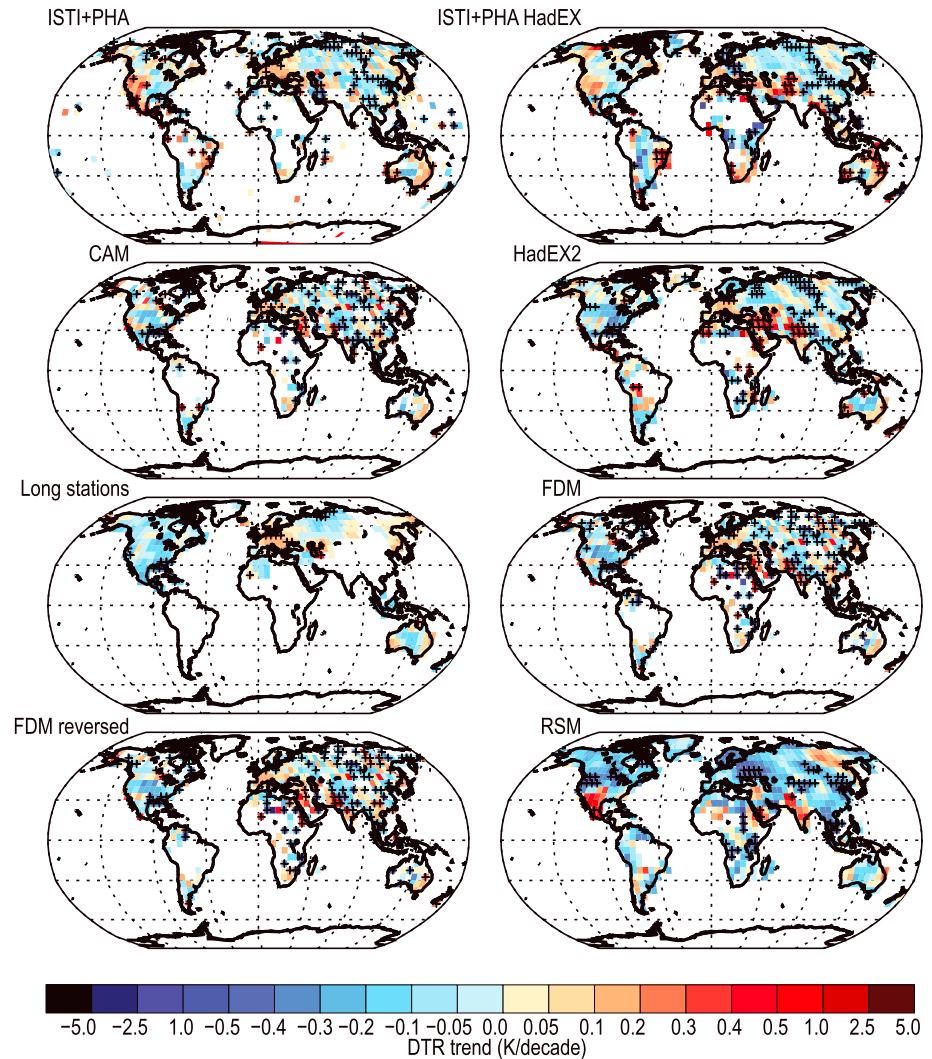**Figure 7.** As in Figure 6 but for the subperiod 1979–2004.

FDM, and *FDM_rev* techniques undertake no infilling or interpolation and therefore have more restrictive coverage. The *CAM* data set version requires 25 out of 30 years during 1961–1990 and grids resulting anomaly series using simple bin averaging. The *FDM* and *FDM_rev* techniques use first differences and a cumulative summation approach rather than anomalies to build grid box series [*Peterson et al.*, 1998a]. Errors arising from random and uncorrected systematic errors grow either forward (*FDM*) or backward (*FDM_rev*) in time. Finally, the *RSM* defines a set of long-term basis station series and grows these by merging in other series adjusting to be similar to the chosen reference series within a radius of influence [*Hansen and Lebedeff*, 1987]. Stations with at least 20 years of record are required. The method is critically dependent upon the homogeneity of the chosen reference series from which to grow the gridded series.

Furthermore, there are two variants of *ISTI+PHA* and *HadEX2* that can be utilized to directly ascertain sensitivity to choice of station basis and their homogenization. The pairs "*ISTI+PHA* versus *CAM*" and "*ISTI+PHA-HadEX* versus *HadEX2*" have postprocessing that is (almost in the first pair) identical but differ in input data and its homogenization. Differences between each pair will therefore principally arise due to the underlying station basis and the choices of homogeneity adjustments.

Gridded trends diverge markedly between the variants of the *ISTI+PHA* and *HadEX2* products over both 1951–2004 (Figure 6) and 1979–2004 (Figure 7). Over both periods, the effects of the gridding and interpolation postprocessing (*ISTI+PHA* versus *CAM* compared to *ISTI+PHA-HadEX* versus *HadEX2*) are very substantial.
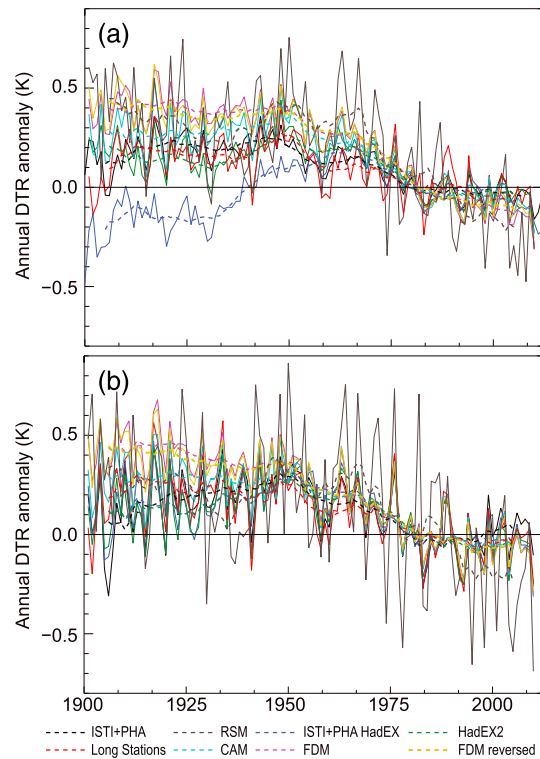
Figure 8. Global-mean time series of anomalies relative to 1971–2000 for the set of gridded variants. (a) Native coverage for each variant. (b) Subsampled only to common coverage at each time step. See Figure 3 for further details of plot construction and to compare to inter-data set differences for the same metric (although note that temperature axes ranges are extended compared to those in Figure 3 to account for interannual variance in *RSM* series).

The use of the HadEX gridding methodology (angular distance weighting) [*Shepard*, 1968] substantially grows the coverage, regardless of the choice of station basis set. However, over much of the expanded domain, the trends are either much larger (both increases and decreases) than in the well-sampled regions or exhibit strong horizontal gradients which are atypical of areas of greater station data coverage. The impacts of gridding method choice are not limited to areas where *HadEX2* grows the data into—there are also changes to both products in areas where both the *CAM* and HadEX-method gridded versions have data coverage. For example, use of the HadEX gridding methodology changes the sign of trends in DTR over eastern Central Europe over 1979–2004 from an increase (*CAM* approach) to a decrease (HadEX approach). The two underlying station sets also impart differences into the geographical patterns. This is particularly marked over North America where the two *ISTI+PHA* data set variants suggest an increase in DTR since 1979 west of the Rockies, which is not greatly evident in the two *HadEX2* gridded data set variants.

A comparison of the six *HadEX2* variants highlights substantial spread in many regions and individual grid boxes. The *longstations*, *CAM*, and two FDM-based estimates have lower coverage than either *RSM* or *HadEX2*. For the *CAM* and the two FDM-based estimates there are several grid boxes that are, on a visual inspec-

tion, potential outliers. There is some possibility that *HadEX2* and, in particular, *RSM* may be susceptible to growing these erroneous values regionally through their interpolation techniques. The *RSM* estimate has substantially stronger regional trends of both signs than the remaining estimators over both periods. This likely relates to the gridding method that merges stations matching to the chosen reference, which will tend to retain the original local-scale variance/trend at the seed stations disproportionately.

Global and regional series have been constructed as in section 4.2. Figure 8 shows global anomalies under the two approaches. By comparison to Figure 3, which illustrates inter-data set differences, the spread in *ISTI+PHA* and *HadEX2* variants is at least as large, if not larger, than the inter-data set spread. As is the case for data set comparisons, reducing to common coverage substantially reduces the apparent spread. The choice as to whether to apply a *CAM* or *HadEX2* gridding approach has extremely marked effects when using the *ISTI+PHA* station set, changing entirely the nature of the early twentieth century changes in global-mean DTR. The effects of gridding choices are much less marked when instead using the *HadEX2* station basis, but the spread is still comparable to the inter-data set spread, suggesting that the apparent spread in data sets could plausibly be explained entirely by the different choices to gridding and postprocessing. The *RSM* method shows greater interannual and even decadal variability than other variants. Again, this likely relates to the noncancelation, indeed growing, of local effects from the seeded reference stations inherent in the approach.

The inflation of variance for the *RSM* approach is far more pronounced at regional scales (Figure 9 cf. Figure 4). Over North America the principal driver of distinctions between estimates appears to be the underlying station network and its homogenization, with a clear distinction in multidecadal behavior between the *HadEX2* station set-based estimates and the *ISTI+PHA* station-based estimates. The two *ISTI+PHA* estimates exhibit markedly reduced long-term trends and are flat rather than rapidly decreasing post-1980. This is consistent with *ISTI+PHA* more aggressively addressing the MMTS transition data homogeneity issue (section 4.1). Over the contiguous
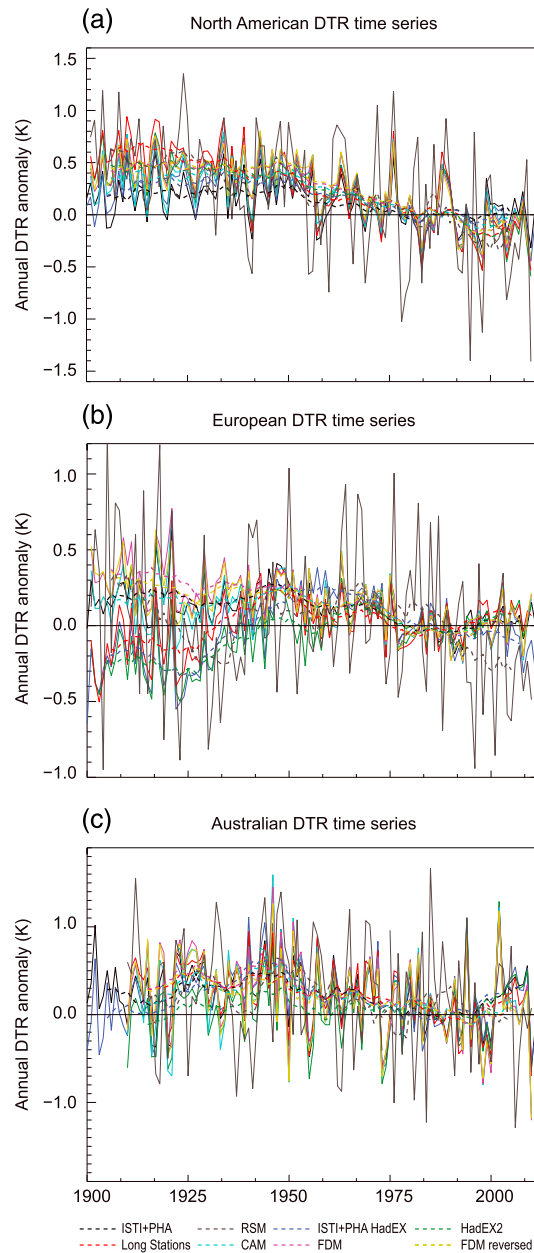
Figure 9. As in Figure 4 but for the variants of the *ISTI+PHA* and *HadEX2* products. Note that *y* axis scales differ between Figures 4 and 9.

states, which constitute much of this region, *HadEX2* takes in solely a subset of GHCN-Daily without apparent breaks. By contrast the *ISTI+PHA* station basis set is homogenized monthly mean series for all stations across this domain.

In marked contrast to the situation over North America and Europe, the principal determinant of differences is the choice of gridding and interpolation methodology. Those methods that interpolate show distinct multidecadal behavior compared to the four estimates that do not. Further, *RSM* is distinct from the three HadEX gridding method versions. Over the European domain both the *HadEX2* station basis and the *ISTI+PHA* station basis arise largely from the European Climate Assessment and Data Set effort [*Klok and Klein Tank*, 2009].

Unlike over North America and Europe, there is no clear single cause of the differences in behavior over Australia. In the last 15 to 20 years the two *ISTI+PHA* estimates do show a slight tendency to greater increases in DTR than the various *HadEX2* variants, but it is not as clear as in other regions considered that station basis and homogenization matter. This is, in part, because there is much greater interannual variability in this region (see *Thorne et al.* [2016] and references therein for a more in-depth discussion).

Similar plots to Figure 9 for common coverage are shown in Figure S3. As is the case in the data sets comparisons in section 4.2 (Figure 5 cf. Figure 4), restricting to common data set coverage serves to improve congruence between estimates but does not yield an exact match.

The spread in global and regional trends for the *ISTI+PHA* and *HadEX2* variants at their native coverage (Table 5) is at least as large as that between data sets considered in section 4 (Table 2). All the variants of *HadEX2* and *ISTI+PHA* agree that over 1951–2004, there has been a significant decrease in globally averaged DTR, although they vary over a factor of 3 in the estimated magnitude (*ISTI+PHA* gives −0.042 K/ decade and *RSM* −0.119 K/decade). There is also an agreement upon a decrease for the same period over North America, but not its significance. For all other regions and periods estimates are of mixed sign and significance. The families of estimates identified in Figure 9 and the discussion thereof (see preceding text) for North America are reflected in the long-term trend estimates. The sensitivities identified in the prior discussion over Europe, however, precede the period over which Table 5 trends are calculated.

6. Discussion

In sections 4 and 5, an assessment first through intercomparison of seven data sets, and second for two of these data sets of sensitivity to gridding and interpolation choices, has served to highlight substantial

Table 5. As in Table 2 but for the Variants on ISTI+PHA and HadEX2 Considered in Section 4.2 and Only at the Native Data Set Coverage

Domain	Period	ISTI+PHA	ISTI+PHA-HadEX	HADEX2	CAM	FDM	FDM Reversed	Longstations	RSM
Global	1951–2004	**−0.042**	**−0.047**	**−0.061**	**−0.085**	**−0.102**	**−0.100**	**−0.047**	**−0.119**
	1979–2004	0.001	0.000	−0.005	−0.020	−0.059	−0.045	−0.031	**−0.127**
North America	1951–2004	−0.034	**−0.066**	**−0.100**	**−0.090**	**−0.106**	**−0.113**	**−0.086**	−0.095
	1979–2004	0.042	0.012	−0.112	−0.051	−0.079	−0.080	−0.103	−0.131
Europe	1951–2004	**−0.036**	**−0.075**	−0.000	**−0.047**	**−0.050**	**−0.047**	−0.013	−0.061
	1979–2004	0.033	−0.031	0.048	0.028	0.022	0.026	0.075	−0.233
Australia	1951–2004	−0.045	−0.043	0.010	−0.061	−0.073	−0.056	−0.059	−0.034
	1979–2004	0.086	0.129	0.031	−0.055	−0.087	−0.065	0.006	−0.041

uncertainty in many aspects of global and regional DTR behaviors. Despite the spread in estimates, a statistically significant decrease in globally averaged DTR since the middle twentieth century is found to be robust to both inter-data set differences and sensitivity to reasonable gridding and interpolation processing choices. Changes at regional scales or shorter timescales are either not consistently statistically significant or of mixed sign. Changes are much more uncertain in data sparse regions and epochs.

Because the data sets so comprehensively span structural uncertainty in all aspects of DTR data set construction (Table 1), it is hard to pinpoint unambiguously the causes of differences. Despite this, it is clear that much of the spread in data sets arises from whether interpolation is performed or not. When all data sets are restricted to the common grid box availability, they are in much better agreement. Often, however, users request globally complete or quasi-complete estimates, unaware of the substantial challenges inherent in making global estimates from a sparse and unevenly distributed network of local point observations. The data sets and variants considered here span four approaches to this problem in total, and the results imply great sensitivity to these choices.

Users who require interpolated fields should clearly consider a range of such estimates and understand them sufficiently to avoid using versions (or estimated data) that are inappropriate for their application area. The analysis highlighted particular potential issues with long-term trends and variance for products using both HadEX and *CRU TS* approaches to interpolation. The *HadEX2* gridding method changes substantively early twentieth century behavior over Europe in comparison to *CAM* or *FDM* techniques, especially so for the *ISTI+PHA* station set. The *CRU TS* method tends to damp both trend and variance, if no restriction is applied using the data flags provided to ignore grid boxes far from a data constraint that are solely a repeating climatology of DTR.

The inherent problems associated with interpolation highlighted herein will apply more broadly to other both analyses and reanalyses. Regardless of whether using statistical or data assimilation-based techniques, the uncertainty shall always be greatest where the observational constraint is either weak or entirely absent. While it is undoubtedly true that better techniques may be possible, there will always be a limit unless sufficient new data can be recovered to provide an observationally based constraint in data sparse or data absent regions. This highlights the need to rescue additional data.

Station selection and homogeneity assessment also matter greatly. Results using variants of the *ISTI+PHA* and *HadEX2* products imply an intuitive interplay, although this may just be fortuitous. The data sets are built from fairly similar sources over most regions except North America, where the station basis and, in particular, their homogenization for the MMTS transition in the 1980s differ substantively. Over North America, there is a clear split driven by this choice, whereas over Europe the split is primarily determined by the gridding and interpolation choices. Results globally and over Australia are indeterminate.

There are other potential confounding issues that have not been considered explicitly herein. Most prominently, roughly half of the data sets (*HadEX2*, *GHCNDEX*, and *HadGHCND*) are built from daily series of maximum and minimum temperatures, while the remainder is built from monthly averages of maximum and minimum temperatures or a mixture of daily and monthly averages. For a perfectly complete record this would not matter. However, many stations are not complete and the more missing days or flagged days exist in a month the greater the disparity between the calculated monthly mean and the true monthly mean of dailies may become. It is not possible, given the available data, to fully ascertain the probable impact of this choice.

A common thread running through the various findings and caveats detailed in this discussion is that robust analyses of DTR globally prior to 1950, and even in some cases regionally thereafter, are limited by underlying observational data availability. All techniques would likely work well if plentiful data were available throughout the period of record and across the globe. Differences become greater, the fewer data are available or the more the data availability changes with time. Therefore, the ability to improve understanding of changes in DTR prior to the middle twentieth century is inextricably linked to efforts to rescue and digitize additional maximum and minimum temperature series. This will also help with other surface temperature variables but is an acute need if DTR analyses are going to be extended back prior to the middle twentieth century with any degree of confidence. Data rescue alone may not entirely reduce uncertainties in DTR changes. There are still discrepancies between available DTR estimates post-1980, when data availability is less of an issue, but these differences are smaller than those prior to 1950.

7. Conclusions

The driving rationale behind this work was the lack of explicit progress in the literature in assessing DTR changes between the fourth and fifth assessment reports of the IPCC. Based upon the findings herein, where a new assessment to be performed by IPCC of the observational DTR record at this time the text might read as follows (use of IPCC carefully calibrated uncertainty language and italicization [*Mastrandrea et al.*, 2010] is intended).

It is *virtually certain* that globally averaged DTR has significantly decreased since 1950. This reduction in DTR is robust to both choice of data set and to reasonable variations in station selection and gridding methodology. However, differences between available estimates mean that there is only *medium confidence* in the magnitude of the DTR reductions. It is *likely* that most of the global-mean decrease occurred between 1960 and 1980 and that since then globally averaged DTR has exhibited little change. Because of current data sparsity in the digitized records, there is *low confidence* in trends and multidecadal variability in DTR prior to the middle twentieth century. It is *likely* that considerable pre-1950 data exist that could be shared and/or rescued and used in future analyses. All assessed estimates of global DTR changes are substantially smaller than the concurrently observed increases in mean and maximum and minimum temperatures (*high confidence, virtually certain*).

References

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

Caesar, J., L. Alexander, and R. Vose (2006), Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set, *J. Geophys. Res.*, *111*, D05101, doi:10.1029/2005JD006280.

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

Donat, M. G., L. V. Alexander, H. Yang, I. Durre, R. Vose, and J. Caesar (2013a), Global land-based datasets for monitoring climatic extremes, *Bull. Am. Meteorol. Soc.*, *94*, 997–1006, doi:10.1175/BAMS-D-12-00109.1.

Donat, M. G., et al. (2013b), Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset, *J. Geophys. Res. Atmos.*, *118*, 2098–2118, doi:10.1002/jgrd.50150.

Dunn, R. J. H., M. G. Donat, and L. V. Alexander (2014), Investigating uncertainties in global gridded datasets of climate extremes, *Clim. Past*, *10*, 2171–2199, doi:10.5194/cp-10-2171-2014.

Durre, I., M. J. Menne, B. E. Gleason, T. G. Houston, and R. S. Vose (2010), Comprehensive automated quality assurance of daily surface observations, *J. Appl. Meteorol. Climatol.*, *8*, 1615–1633.

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

Hansen, J. E., and S. Lebedeff (1987), Global trends of measured surface air temperature, *J. Geophys. Res.*, *92*, 13,345–13,372, doi:10.1029/JD092iD11p13345.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister (2014), Updated high-resolution monthly grids of monthly climatic observations: The CRU TS 3.10 dataset, *Int. J. Climatol.*, *34*, 623–642, doi:10.1002/joc.3711.

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 University Press, Cambridge, UK and New York, doi:10.1017/CBO9781107415324.008.

Hawkins, E., and R. Sutton (2016), Connecting climate model projections of global temperature change with the real world, *Bull. Am. Meteorol. Soc.*, 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.

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.

Klok, E. J., and A. M. G. Klein Tank (2009), Updated and extended European dataset of daily climate observations, *Int. J. Climatol.*, *29*, 1182, doi:10.1002/joc.1779.

Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Renni (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., H. Zhang, X. Liu, J. Chen, W. Li, and P. Jones (2009), A mainland China homogenized historical temperature dataset of 1951–2004, *Bull. Am. Meteorol. Soc.*, *90*, 1062–1065.

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., G. J. Steeneveld, A. A. M. Holtslag, R. A. Pielke Sr., S. Mackaro, A. Pour-Biazar, J. Walters, U. Nair, and J. Christy (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 Jr. (2005), Detection of undocumented changepoints using multiple test statistics and composite reference series, *J. Clim.*, *18*, 4271–4286.

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

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, *JAOT*, *29*, 897–910, doi:10.1175/JTECH-D-11-00103.1.

Paaijmans, K. P., S. Blanford, A. S. Bell, J. I. Blanford, A. F. Read, and M. B. Thomas (2009), 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., and M. J. Manton (2008), Monitoring changes in climate extremes: A tale of international collaboration, *Bull. Am. Meteorol. Soc.*, *89*, 1266–1271.

Peterson, T. C., and R. S. Vose (1997), An overview of the Global Historical Climatology Network Temperature Database, *Bull. Am. Meteorol. Soc.*, *78*, 2837–2849, doi:10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2.

Peterson, T. C., T. R. Karl, P. F. Jamason, R. Knight, and D. R. Easterling (1998a), First difference method: Maximizing station density for the calculation of long-term global temperature change, *J. Geophys. Res.*, *103*(D20), 25,967–25,974, doi:10.1029/98JD01168.

Peterson, T. C., R. Vose, R. Schmoyer, and V. Razuvaëv (1998b), Global historical climatology network (GHCN) quality control of monthly temperature data, *Int. J. Climatol.*, *18*, 1169–1179, doi:10.1002/(SICI)1097-0088(199809)18:11<1169::AID-JOC309>3.0.CO;2-U.

Peterson, T. C., et al. (2001), Report on the activities of the working group on climate change detection and related rapporteurs 1998–2001, WMO, Rep. WCDMP-47, WMO-TD 1071, Geneve, Switzerland, 143 pp.

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.*, *1*, doi:10.4172/gigs.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.*, *1*, doi:10.4172/gigs.1000103.

Santer, B. D., et al. (2008), Consistency of modelled and observed temperature trends in the tropical troposphere, *IJOC*, *28*, 1703–1722, doi:10.1002/joc.1756.

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.

Shepard, D. (1968), A two-dimensional interpolation function for irregularly spaced data, paper presented at 23rd National Conference, Assoc. for Comput. Mach, New York.

Solomon, S., et al. (2007), Technical summary, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., Cambridge University Press, Cambridge, U. K., and New York.

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 (2005a), Uncertainties in climate trends—Lessons from upper-air temperature records, *Bull. Am. Meteorol. Soc.*, *86*, 1437–1442.

Thorne, P. W., et al. (2005b), Revisiting radiosonde upper air temperatures from 1958 to 2002, *J. Geophys. Res.*, *110*, D18105, doi:10.1029/2004JD005753.

Thorne, P. W., et al. (2016), Reassessing changes in Diurnal Temperature Range: A new dataset and characterization of data biases, *J. Geophys. Res. Atmos.*, *121*, doi:10.1002/2015JD024583.

Trenberth, K. E., et al. (2007), Observations: Surface and atmospheric climate change, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., Cambridge University Press, Cambridge, U. K., and New York.

Vasseur, D. A., J. P. DeLong, B. Gilbert, H. S. Greig, C. D. G. Harley, K. S. McCann, V. Savage, T. D. Tunney, and M. I. O'Connor (2014), Increased temperature variation poses a greater risk to species than climate warming. *Proc. R. Soc. B.*, *281*, 20132612.

Vose, R. S., D. R. Easterling, and B. Gleason (2005a), Maximum and minimum temperature trends for the globe: An update through 2004, *Geophys. Res. Lett.*, *32*, L23822, doi:10.1029/2005GL024379.

Vose, R. S., D. Wuertz, T. C. Peterson, and P. D. Jones (2005b), An intercomparison of trends in surface air temperature analyses at the global, hemispheric and gridbox scale, *Geophys. Res. Lett.*, *32*, L18718, doi:10.1029/2005GL023502.

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.

Williams, C. N., M. J. Menne, and P. W. Thorne (2012a), Benchmarking the performance of pairwise homogenization of surface temperatures in the United States, *J. Geophys. Res.*, *117*, D05116, doi:10.1029/2011JD016761.

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 Tech. Rep., NCDC 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.]

Zhang, X., et al. (2011), Indices for monitoring changes in extremes based on daily temperature and precipitation data, *WIREs Clim. Change*, doi:10.1002/wcc.147.

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