The Global Historical Climatology Network Monthly Temperature Dataset, Version 4

MATTHEW J. MENNE, CLAUDE N. WILLIAMS, AND BYRON E. GLEASON

NOAA/National Centers for Environmental Information/Dataset Section, Asheville, North Carolina

J. JARED RENNIE

Cooperative Institute for Climate and Satellites, North Carolina State University, Asheville, North Carolina

JAY H. LAWRIMORE

NOAA/National Centers for Environmental Information/Dataset Section, Asheville, North Carolina

(Manuscript received 28 February 2018, in final form 26 July 2018)

ABSTRACT

We describe a fourth version of the Global Historical Climatology Network (GHCN)-monthly (GHCNm) temperature dataset. Version 4 (v4) fulfills the goal of aligning GHCNm temperature values with the GHCN-daily dataset and makes use of data from previous versions of GHCNm as well as data collated under the auspices of the International Surface Temperature Initiative. GHCNm v4 has many thousands of additional stations compared to version 3 (v3) both historically and with short time-delay updates. The greater number of stations as well as the use of records with incomplete data during the base period provides for greater global coverage throughout the record compared to earlier versions. Like v3, the monthly averages are screened for random errors and homogenized to address systematic errors. New to v4, uncertainties are calculated for each station series, and regional uncertainties scale directly from the station uncertainties. Correlated errors in the station series are quantified by running the homogenization algorithm as an ensemble. Additional uncertainties associated with incomplete homogenization and use of anomalies are then incorporated into the station ensemble. Further uncertainties are quantified at the regional level, the most important of which is for incomplete spatial coverage. Overall, homogenization has a smaller impact on the v4 global trend compared to v3, though adjustments lead to much greater consistency than between the unadjusted versions. The adjusted v3 global mean therefore falls within the range of uncertainty for v4 adjusted data. Likewise, annual anomaly uncertainties for the other major independent land surface air temperature datasets overlap with GHCNm v4 uncertainties.

1. Introduction

The Global Historical Climatology Network (GHCN)-monthly (GHCNm) dataset was originally developed at a time of limited access to digital climate data from global land stations. Like the earliest versions of the University of East Anglia Climate Research Unit Temperature (CRUTEM) data (Jones et al. 1986a,b), the GHCNm weather station records were collated from numerous different sources through formal and informal exchanges in an effort to build a reasonably comprehensive set of land station data that could be used to provide historic perspective on surface air temperature and precipitation fields across global land areas. Since GHCNm, version 1, was released (Vose et al. 1992), the dataset has undergone a number of revisions to incorporate additional station data as well as innovations to data processing (Peterson and Vose 1997; Lawrimore et al. 2011). Starting with version 2, short time-delay updates have been part of the processing suite, which has allowed the GHCNm station values to be used as the land component in global surface temperature (GST) monitoring (Hansen et al. 2010; Vose et al. 2012).

In this paper, we provide an overview of a fourth major version of the temperature component of GHCNm. The motivation for producing this version was to align

Denotes content that is immediately available upon publication as open access.

Corresponding author: Matthew Menne, matthew.menne@noaa.gov

DOI: 10.1175/JCLI-D-18-0094.1

the GHCNm dataset with GHCN-daily (GHCNd; Menne et al. 2012). Up to this point, GHCNm and GHCNd have developed along largely independent pathways. With the release of GHCNm, version 4 (v4), GHCNd and GHCNm now share common identifiers, and monthly values are computed directly from GHCNd whenever possible. The other motivation was to compute more comprehensive uncertainty estimates for land surface air temperatures over global land areas. The paper is organized as follows: In section 2, we provide a brief history of the dataset. In section 3, we describe the process of combining different source archives to build a more complete monthly temperature database as well as the mechanisms for updating the dataset. The quality-control process is described in section 4, and the data trends and impact of homogenization is covered in section 5. Section 6 provides an assessment of the uncertainty components of the land surface air temperature data, and comparisons to other datasets are provided in section 7.

2. History of the dataset

The first version of GHCNm (Vose et al. 1992) was assembled using early efforts in global data collection like the World Weather Records program (Clayton 1927), which endures today, with other collections like the World Monthly Surface Station Climatology managed at the National Center for Atmospheric Research. Combining these with other sources led to a version 1 dataset that provided monthly mean temperatures from some 6000 land surface stations worldwide. Additional efforts to acquire data sources through professional contacts and bilateral agreements continued through the mid-1990s and led to a new GHCNm version release in 1997 (Peterson and Vose 1997). Version 2 was made up of data from 31 sources and was operationalized to provide routine quality control and short time-delay updates. These routine updates facilitated use of GHCN as the land component for GST records developed at the National Aeronautics and Space Administration (NASA; Hansen et al. 1999, 2010) and the National Oceanic and Atmospheric Administration (NOAA; Quayle et al. 1999; Smith et al. 2008). Version 2 processing remained essentially unchanged until the release of version 3 in 2011, which included improvements to quality-control and homogenization methods (Lawrimore et al. 2011). These efforts to construct GHCNm have occurred broadly in parallel with efforts to create the CRUTEM datasets (Bradley et al. 1985; Jones et al. 1986a,b; Jones 1994; Jones and Moberg 2003; Jones et al. 2012). More recently, a third group, Berkeley Earth, produced a global land surface air temperature dataset (Rohde et al. 2013) using many of the same sources as GHCN, but with a greatly expanded station network relative to earlier efforts. Both CRUTEM and Berkeley Earth provide uncertainty estimates, and GHCNm v4 is compared to the latest versions of these datasets in section 7. Most recently, a fourth group (Xu et al. 2018) associated with the China Meteorological Agency (CMA) has combined data from all three of these global land datasets (GHCN, CRUTEM, and Berkeley Earth) with some national datasets to produce another land surface air temperature (LSAT) dataset known as CMA-LSAT. The CMA-LSAT dataset was not publicly accessible at the time of writing.

3. The v4 monthly database

GHCNm v4 monthly land surface air temperatures are an amalgamation of many different datasets archived at NOAA/NCEI, which have been collated under the auspices of the International Surface Temperature Initiative (ISTI; Thorne et al. 2011; Rennie et al. 2014). More specifically, GHCNm v4 temperature values are based on version 1.1.0 of the ISTI monthly temperature databank. The ISTI monthly data were produced by merging monthly temperatures calculated from the GHCNd dataset (Menne et al. 2012) with over 50 additional smaller source archives to create an integrated set of monthly land station temperature records using GHCNd as the foundational source. The aim for ISTI, like previous versions of GHCNm, was to build the most complete set of monthly station records possible while also reconciling the redundancies among the separate data sources (Rennie et al. 2014). Version 1.1.0 contains monthly temperature series for about 32 000 separate stations. However, GHCNm v4 uses only those station records that span a minimum of 10 years, which reduces the total number of stations to about 26 000.

The ISTI databank working group (Thorne et al. 2011) advocates for the provision of data from all processing stages beginning with the version that is as close to the original raw measurements as possible. These basic sources are then reformatted centrally at NOAA/NCEI and combined into the ISTI integrated monthly dataset. The relevant process stages defined by the ISTI project are as follows (Thorne et al. 2011):

- Stage 0: Original data—image or information regarding primary document
- Stage 1: Original native digital format used by the provider
- Stage 2: Data converted into a common format and with provenance and version control information appended

- Stage 3: Collation of all stage-2 data including merging/mingling of records for stations represented across multiple stage-2 sources

In essence, the 50-plus stage-1 data sources (see Table 1 from Rennie et al. 2014) were reformatted into the common stage-2 format. The stage-2 data were then merged to build the integrated set of station records that comprise stage 3. In the stage-2 processing, sources with daily data were converted to monthly averages. The methodology for creating a monthly average follows the recommendation of the World Meteorological Organization (WMO 2017), which specifies that a monthly average can have at most five missing days but never more than three consecutive missing days. When these criteria are not met, the monthly value is set to missing from that source. In the case of redundancy among sources, monthly data computed directly from ISTI daily data sources are given priority over sources where only monthly values are provided. This prioritization helps to ensure that monthly averages can be traced back to and are consistent with the daily data whenever possible. When monthly mean maximum and minimum temperatures were available, these sources were also given higher priority over sources having only a single monthly mean. This also helps to improve traceability in how the monthly averages were derived, and an effort was made to specifically exclude stage-1 sources that were already adjusted (homogenized) prior to being submitted to the databank.

The merge step that leads from separate stage-2 sources to an integrated stage-3 database is performed in a manner that compares the ‘‘master dataset’’ with each successive candidate source that requires integration into the master set. The master dataset starts with the highest-ranking data source in the stage-2 data and then grows as each new source is added during the merging process in a predetermined order agreed upon by the ISTI databank working group. During the merge process, two types of comparisons are conducted between station records in the master dataset and those in the source to be added: a metadata comparison and a data comparison.

In the metadata comparison, geographical attributes (i.e., latitude, longitude, elevation, name, and record start date) for each station in the new source are compared to all stations in the master set. A source station that meets the threshold for similarity with a station in the master set is identified for further consideration via data comparisons using the index of agreement (Willmott et al. 1985). In the case of a sufficiently high data similarity [see Rennie et al. (2014) for details], the station record is merged with the existing master station record. If the overlapping data appear to be two distinct station records, based on lower values of the index of agreement, the source station is added as a new station to the master set despite the metadata similarity. Likewise, if no stations in the master set are sufficiently similar based on data or metadata, the source station is considered to be new to the master set, and its data are added to the master database under a new station identifier. To conduct the comparisons, the merge algorithm makes use of two sets of empirical probabilities created to quantify the likelihood of a match versus the likelihood of being distinct. When the match or distinctness probabilities are inconclusive (i.e., largely overlapping), then the source station is withheld from the master set entirely.

The algorithm is run automatically, and there are some merge errors whereby a source station is added as new to the master dataset when it is in reality already represented in the master set from a different source, and conversely, cases when a source record is matched to a master station record when it is in fact a distinct, separate station record. The merge algorithm is particularly challenged in cases where a data source contains records in which two nearby but distinct station records were previously combined prior to being provided to ISTI, while they also exist as separate station records in a higher-ranking source. In such cases, the master set to which the source is being added may already treat the two station records as coming from separate locations. The algorithm is then faced with a decision to merge the composited two-station record of the lower-ranking source with one of the two separate master records, add it as a new distinct location, or withhold the station records from the merged data. Because the first two options are problematic, the best outcome is to withhold the station; however, withholding is not guaranteed in the framework of the empirical probabilities used to quantify the likelihood of uniqueness and similarity between sets of climate records. For this reason, the CRUTEM source data, which are known to contain numerous composited station records, were not used to build version 1.1.0 (Rennie 2015) though they were used in the first release (v1.0.0; Rennie et al. 2014).

GHCNd is itself made up of 30 unique sources (Menne et al. 2012) so the ISTI, stage 3, version 1.1.0, actually contains data from 87 unique sources. Each monthly value is accompanied by a source flag (flag 3). In terms of data volume, about 75% of all monthly values originate from GHCNd. The next highest contributions come from sources that have monthly values only. These include ‘‘ghcnsource’’ [the data collected and used for GHCNm, version 2 (v2); Peterson and Vose 1997] at 5.6%, 4.4% from the European Climate Assessment and Data (Klein Tank et al. 2002), and 2.5%

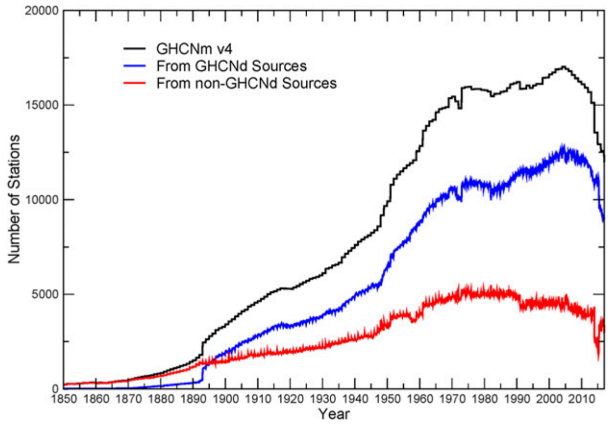

FIG. 1. Number of stations with data by month for GHCNm v4 (black)—based on the ISTI, stage 3, dataset (version 1.1.0)—and the subsets coming from GHCNd (blue) and non-GHCNd (red) sources.

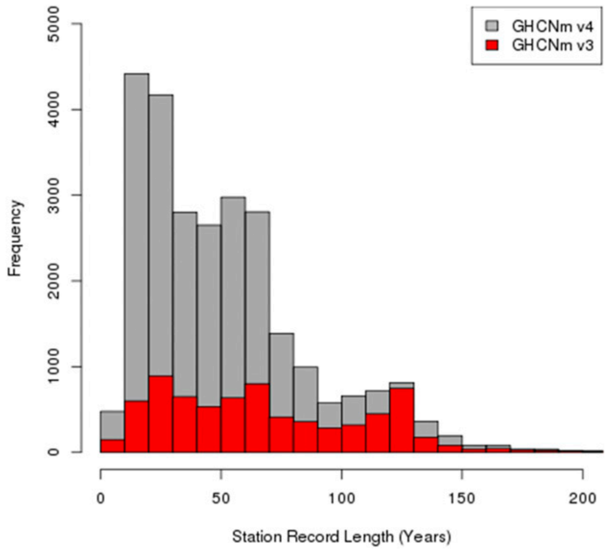

FIG. 2. Histogram of station temperature record length.

from ‘‘russsourceghcn’’ (a collection of numerous monthly source archives from NCEI). As shown in Fig. 1, the number of stations reporting values in v4 is much higher than in version 3 (v3) throughout the period of record, with 10 000 or more stations having average land surface temperature values after about 1950. The majority of stations in v4 have record lengths of about 50 years or less, but v4 also roughly doubles the number of stations with very long records of a century or more, as shown in Fig. 2. Improvements in spatial coverage are discussed in section 6.

The ISTI, version 1.1.0, stage 3, data are updated monthly for use in GHCMm v4 using two data streams: new monthly values calculated from GHCNd and the CLIMAT monthly climate summary messages transmitted over the WMO’s Global Telecommunications System (GTS) that have been used in previous versions of GHCNm. The use of GHCNd as an update source provides many thousands more stations with short time-delay updates compared to GHCNm v3, which only provided monthly updates from CLIMAT data for stations outside of the United States.

4. QC checks

Following data collection and integration, quality control (QC) is performed to flag likely errors in the monthly temperature values. The QC consists of basic integrity, outlier, and spatial consistency checks, described briefly in Table 1. The same checks described in Lawrimore et al. (2011) for GHCNm v3 are applied to the v4 station data. In the v4 process, four additional checks have been included: 1) an interstation duplicate check similar to what is applied in GHCNd (Menne et al. 2012), 2) a spatial “*z* score” check, 3) a streak check, and 4) a world record extremes check. These new checks are also described in Table 1. The process begins with four basic integrity checks followed by an outlier check and then applies the spatial consistency checks. Once an observation fails any quality-control check, the value is excluded from subsequent checks.

The full period of record is quality controlled each time the GHCNm processing system is updated to ensure the QC assessment can capitalize on the full period of record as new observations are added. If an observation fails a quality-control check, the value is accompanied by QC flag (flag 2) that indicates what check was failed. A value with no quality-control flag indicates that the datum passed all checks applied or was unable to be tested because of insufficient data. As described in Lawrimore et al. (2011), the quality-control process was designed so that each check has a low false-positive rate, that is, a low probability that valid observations are flagged as errors according to the evaluation approach outlined in Durre et al. (2008). Nevertheless, a QC flag can be overridden if later found by expert assessment to be inaccurate. Any overrides, and the associated justification for making the correction, are documented in NCEI’s Datzilla system (NOAA 2007).

5. Homogeneity testing, trends, and comparison to v3

Nearly all weather stations, at some point during their history, undergo changes in the circumstances under which measurements are taken (Trewin 2010). For example, thermometers require periodic replacement or recalibration, and measurement technology has evolved over time. Temperature recording protocols have also

TABLE 1. List of QC checks applied to monthly temperatures. The * indicates checks that are new to GHCNm v4.

Type of error	Description of check
Interstation duplicate check*	Identifies a station's annual data that are duplicated in any year of another station's data (annual data must have at least three or more nonmissing years of data and at least 12 values (less missing values) within 0.015°C. (E flag)
Series duplication	Identifies data duplication between years within a station (must have 12 exact values, based on integer-to-integer value comparison). (D flag)
World record extremes check*	Identifies temperatures that fall outside the range of the highest and lowest monthly mean maximum and minimum temperature values. (R flag)
Streak*	Identifies runs of the same value (nonmissing) in five or more consecutive months. (K flag)
Consecutive month duplication	Used to identify duplicate retransmission and mislabeling of previous month's temperature for the current month. Occurs in GTS-transmitted CLIMAT bulletins from 2000 to the present. (W flag)
Isolated value	Identifies months that are isolated in time. From one to three consecutive months of nonmissing values are identified and flagged when they are separated from other nonmissing months by 18 or more consecutive months of missing values both before and after the 1–3 months that are isolated. (L flag)
Climatological outlier	Identifies temperatures that exceed their respective climatological means for the corresponding station and calendar month by at least five standard deviations using biweight mean and biweight standard deviation (Lanzante 1996). (O flag)
Spatial inconsistency 1	Flags value when the station z score satisfies any of the following algorithm conditions. Definitions used are neighbor = any station within 500 km of target station; z score = (biweight standard deviation/biweight mean); $S(Z)$ = station's z score; and $N(Z)$ = the set of the "five" closest nonmissing neighbor z scores. (Note: This set may contain fewer than five neighbors, but must have at least one neighbor z score for algorithm execution.) Algorithm: $S(Z) \geq 4.0$ and < 5.0 and "all" $N(Z) < 1.9$ $S(Z) \geq 3.0$ and < 4.0 and "all" $N(Z) < 1.8$ $S(Z) \geq 2.75$ and < 3.0 and "all" $N(Z) < 1.7$ $S(Z) \geq 2.5$ and < 2.75 and "all" $N(Z) < 1.6$ $S(Z) \leq -4.0$ and > -5.0 and "all" $N(Z) > -1.9$ $S(Z) \leq -3.0$ and > -4.0 and "all" $N(Z) > -1.8$ $S(Z) \leq -2.75$ and > -3.0 and "all" $N(Z) > -1.7$ $S(Z) \leq -2.5$ and > -2.75 and "all" $N(Z) > -1.6$ (S flag)
Spatial inconsistency 2*	Identifies when the temperature z score compared to the inverse distance-weighted z score of all neighbors within 500 km (at least two or more neighbors are required) is greater than or equal to 3.0. (T flag)

changed at many locations from recording temperatures at fixed hours during the day to once-per-day readings of the 24-h maximum and minimum. "Fixed" land stations are sometimes relocated, and even minor temperature equipment moves can change the microclimate exposure of the instruments. In other cases, the land use or land cover in the vicinity of an observing site can change over time; this can impact the local environment that instruments are sampling even when measurement practice is stable. All of the above modifications to the circumstances of recording near-surface air temperature can cause systematic shifts in temperature readings from a station that are unrelated to any real variation in local weather and climate. Moreover, the magnitude of these shifts (or inhomogeneities) can be large relative to true climate variability. Inhomogeneities can therefore lead to large systematic errors in the computation of climate trends and variability not only for individual station records but also in spatial averages.

For this reason, detecting and accounting for artifacts associated with changes in observing practice is an important and necessary endeavor in building climate datasets. In GHCNm v4, shifts in monthly temperature series are detected through automated pairwise comparisons of the station series using the algorithm described in Menne and Williams (2009). This procedure, known as the pairwise homogenization algorithm (PHA), systematically evaluates each time series of monthly average surface air temperature to identify cases in which there is an abrupt shift in one station's temperature series (the "target" series) relative to many other correlated series from other stations in the region (the "reference" series). The algorithm uses the standard normal homogeneity test (Hawkins 1976; Alexandersson 1986) to detect breaks and seeks to resolve the timing of shifts for all station series before computing an adjustment factor to compensate for any one particular shift. These adjustment factors are based on the average change in the magnitude of monthly temperature differences between the target station series with the apparent shift and the reference series with no apparent concurrent shifts. The assumption behind this approach to relative homogenization is that spatially isolated and sustained shifts in the mean level of one station series relative to other highly correlated series from nearby stations are artificial, or, at least, that they are unlikely to have originated from regional variations in weather and climate.

The PHA has undergone extensive evaluation and benchmarking (e.g., Menne and Williams 2009; Venema et al. 2012; Williams et al. 2012) and has been shown to be effective at finding a wide variety of shifts in realistic simulated monthly temperature datasets. The algorithm has also been shown to improve the accuracy of anomalies and trends in real data when compared to observations from the U.S. Climate Reference Network (Hausfather et al. 2016), whose stations meet the highest station siting standards and are managed to prevent changes that cause shifts in the data (Diamond et al. 2013). When applied to GHCNm v4 data, the PHA finds about 70 000 shifts in the nearly 26 000 temperature stations that comprise GHCNm v4. Figure 3 shows the frequency distribution of these shifts. As the distribution indicates, the smallest temperature shifts are not detected. Rather, the distribution is bimodal with peaks for detected shift magnitudes of around ±0.5°C and much lower frequencies of adjustments near zero and at absolute magnitudes greater than 2°C. Similar results have been discussed in assessments of the U.S. land surface air temperature series (e.g., Menne et al. 2009; Williams et al. 2012), for GHCNm v3 data (Lawrimore et al. 2011), and in comparisons of homogenization techniques on simulated data (Venema et al. 2012).

There are about 1800 more negative shifts detected in GHCNm v4 series than positive shifts, leading to an average shift magnitude that is slightly less than zero: −0.023°C. The imbalance means that localized jumps in monthly temperature have more often than not caused surface air temperature trends over land stations to be underestimated. A similar finding was also reported for GHCNm, version 3 (Lawrimore et al. 2011). The corresponding shift adjustments are designed to compensate for these artifacts. The result is an average global land surface air temperature trend that is somewhat higher for the period of record than the unadjusted data, as shown in Fig. 4. In v4, the aggregate impact of adjustments on global mean land surface air temperature trends is an increase of about 0.2°C during the period from 1880 to 2016, whereas in v3, the impact of adjustments on global trends is closer to 0.4°C during the same period.

As shown in Fig. 4, both the GHCNm v4 and v3 shift adjustments have an especially large impact on monthly mean temperatures from the late nineteenth century, when thermometers in many parts of the world were rapidly transitioning from north-facing wall mounts, free-standing open structures, and other nonstandard configurations to the enclosed and more standardized Stevenson screen shelter (also known as the cotton-region shelter). The exposure characteristics of the nineteenth century pre-Stevenson screen installations have been shown to have had a positive bias relative to instruments installed in the Stevenson screens, which provide more consistent protection from direct solar radiation (Chenoweth 1993; Parker 1994). The adjustments for the pre-1900 records in GHCNm v4 are consistent with these assessments. After 1900, adjustments to GHCNm v4 have relatively little impact on the global mean land surface air temperatures until around 1940 when adjustments reduce the magnitude of anomalies during that decade, but then increase them over the unadjusted data by about 0.1°C through 1970 compared to the level circa 1940. The v4 adjustments then have essentially no impact on the global mean trend after 1970 (Fig. 4b). For the GHCNm v3 station data, there is more of a monotonic impact of the shift adjustments, which increase the trend throughout the period of record.

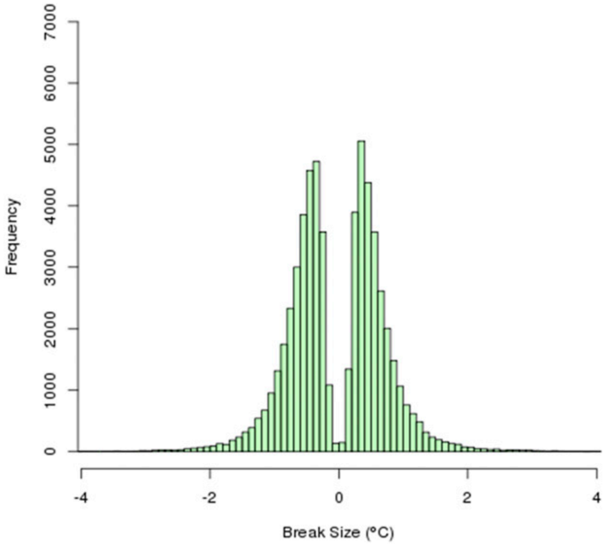

FIG. 3. Histogram of the distribution of shifts identified by the PHA.

There are probably a variety of reasons for why historical shifts in station data have caused the global mean LSAT trends to be underestimated. In the United States, for example, these downward jumps have resulted from a combination of changes in the time of observation for once-per-day temperature measurements (systematic change from afternoon to morning readings). Likewise, beginning in the 1930s, the location of many of the principal weather stations around globe moved from city centers to airports and, in certain cases, from inside to outside urban heat islands. Additional attribution work is nevertheless required to more fully

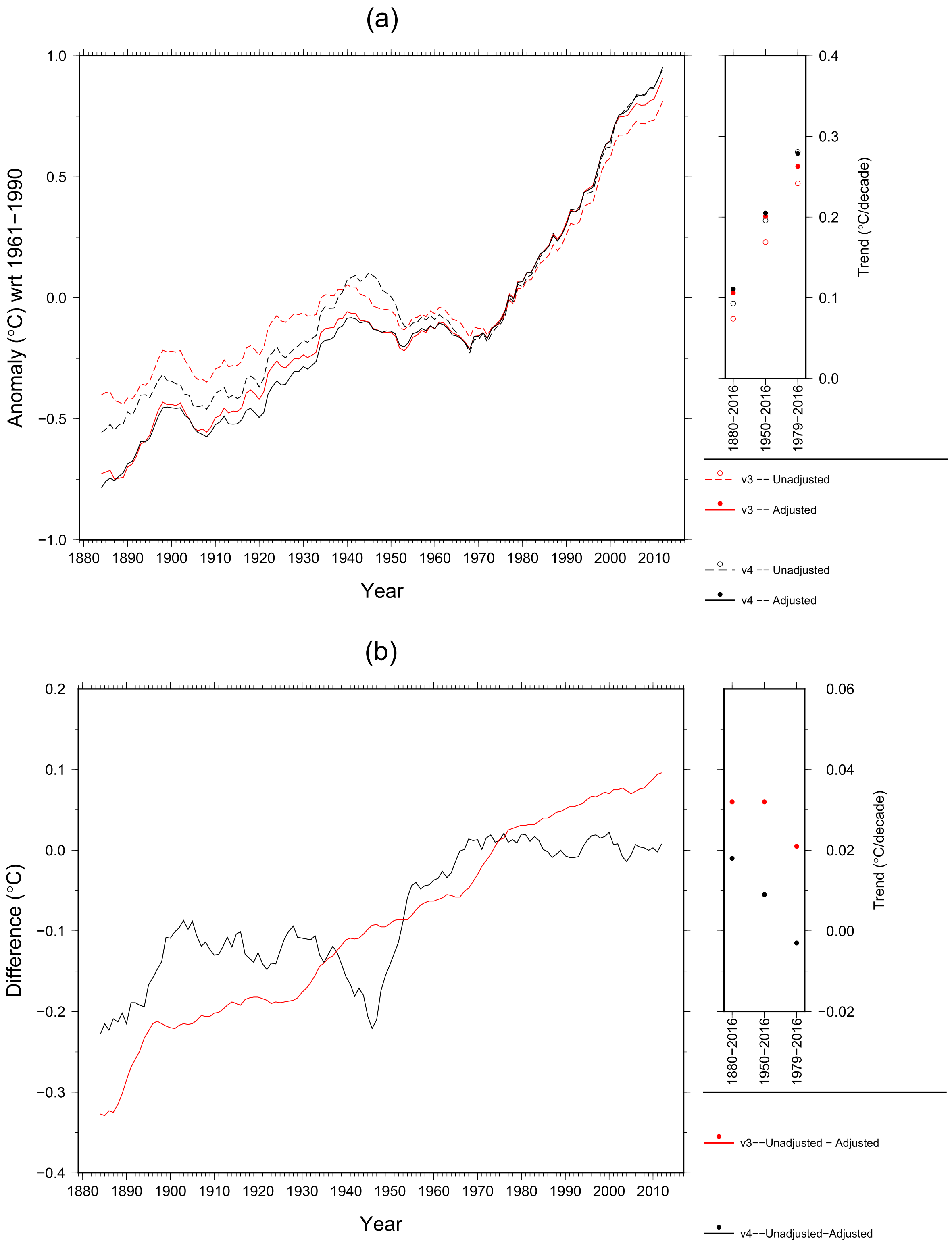

FIG. 4. (a) Average annual global LSAT for GHCNm v3 and v4 based on unadjusted and adjusted station temperature series and (b) mean difference between global average unadjusted and adjusted LSAT computed from GHCNm v3 and v4 stations. Time series in both (a) and (b) are shown as 9-yr running averages.

understand the potential causes of the shifts. Section 7 provides some comparisons to other datasets, which provide independent assessments of the nature and consequences of shifts in the data.

Overall, GHCNm v4 unadjusted data have higher global mean annual trends than the unadjusted v3 for the periods shown in Fig. 4. Notably, the shift adjustments lead to much greater consistency between the adjusted versions in terms of the trend magnitude and in terms of the amplitude of multidecadal variations. As shown in Fig. 5, the greater consistency in the adjusted data occurs across latitude bands for a variety of trend periods with the exception of the most recent two decades when adjustments have relatively little impact on trends. During this time, global mean anomalies diverge somewhat between v3 and v4. This is a period of rapid warming in high latitudes of the Northern Hemisphere and trends diverge more from the global land average than in previous periods (Cowtan and Way 2014). As shown in Fig. 5f, adjustments in v3 actually reduce the trend in the sparsely sampled highest latitudes of the Northern Hemisphere for the period since 2000. This is caused primarily by adjustments compensating for major shifts in anomalies during the 2000s at a few high-Arctic stations located in the Barents and Kara Sea regions where large sea ice loss has occurred (see, e.g., Kintisch 2014). Areas of sea ice loss have been accompanied by unprecedented jumps in temperature anomalies, and these have appeared as artificial discontinuities from a homogeneity perspective at the noted high-latitude stations. In spite of the somewhat higher number of high-latitude stations (north of 60°N), v4 data are not automatically adjusted by the PHA north of 60° because of the rapid changes to anomaly patterns that are altering temperature correlation scales. Rather, any apparent artificial shifts associated with station management changes noted in the future in those areas will require manual intervention.

6. Uncertainty budget for the GHCNm v4 data

Our approach to building an uncertainty budget for GHCNm v4 temperatures broadly follows methods described in Brohan et al. (2006) and Morice et al. (2012) that have been applied to the CRUTEM dataset (Jones et al. 2012). Further, like Morice et al. (2012), we produce an ensemble of the GHCNm v4 dataset as a way to quantify uncertainty that arises from correlated error structures in the data. These errors, correlated in both space and time, originate largely from the artificial shifts in station temperature series discussed in section 5 and do not cancel out when spatially averaged. We describe the major components of the uncertainty budget in the subsections below.

a. *Homogenization uncertainty*

We begin by quantifying the uncertainties associated with shifts in each specific GHCNm v4 station series by running the PHA in ensemble mode. Calculating the homogenization uncertainty at the station level as the first step allows the uncertainty of regional and global averages to be scaled consistently from the station series. The total homogenization uncertainty can be broken down into two parts: 1) uncertainty associated with the use of the PHA itself—termed the parametric uncertainty (e.g., Williams et al. 2012) and 2) uncertainty associated with incomplete homogenization caused by shifts that remain undetected by the algorithm.

To assess the parametric uncertainty of the PHA, we ran the algorithm as an ensemble of 100+ members. The ensemble was generated by varying key parameters as in Williams et al. (2012). These parameters include, for example, the minimum number of neighbors required to be used as a reference series and the minimum length of time between inhomogeneities. In our case, however, the setting combinations used (e.g., minimum number of neighbors required) were determined to have higher skill compared to the purely random set of combinations described in Williams et al. (2012). Skill was defined by the settings having a high hit rate and low false-detection rate in the Williams et al. benchmarks as well as higher skill in recovering the true magnitude of the underlying climate trend in the synthetic benchmarks. The outcome of running the PHA with different parameter settings is a set of detected shifts and adjustments for each of the GHCN series that span the possible outcomes of using the PHA with reasonable settings. Figure 6a provides an example of the range of results for monthly average temperature from Reno, Nevada, a location with a number of location and equipment changes as well as a growing urban heat island signal (Menne et al. 2009). Because of the many shifts caused by station changes, there is a relatively large parametric uncertainty associated with the PHA for the Reno series as each version of the algorithm produces slightly different estimates for the timing and magnitude of the shifts. Most station series are characterized by smaller data shifts and therefore lower parametric uncertainty. By convention, shifts are adjusted to make the earlier data consistent with the latest observation practice so uncertainty increases backward in time, especially when a station record has multiple shifts during its period of record. In the case of Reno, there are multiple shifts during and after the 1961–90 base period. This leads to relatively large parametric uncertainties for the base period mean, which in turn cause uncertainties in the magnitude of anomalies even for recent data, as shown in Fig. 6b. Nevertheless, a great deal of the uncertainty does

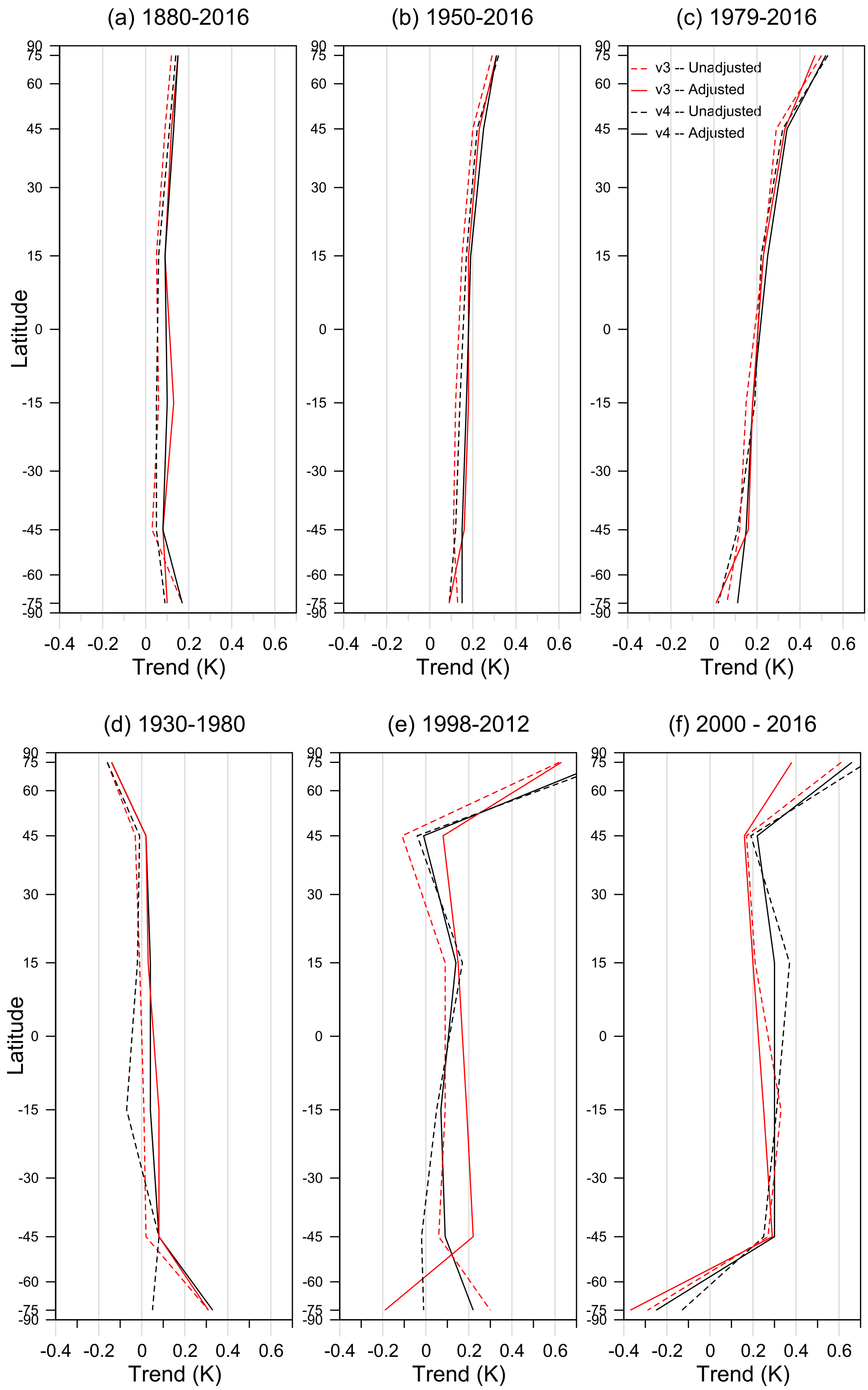

FIG. 5. Trend magnitudes averaged by latitude for six different periods of record.

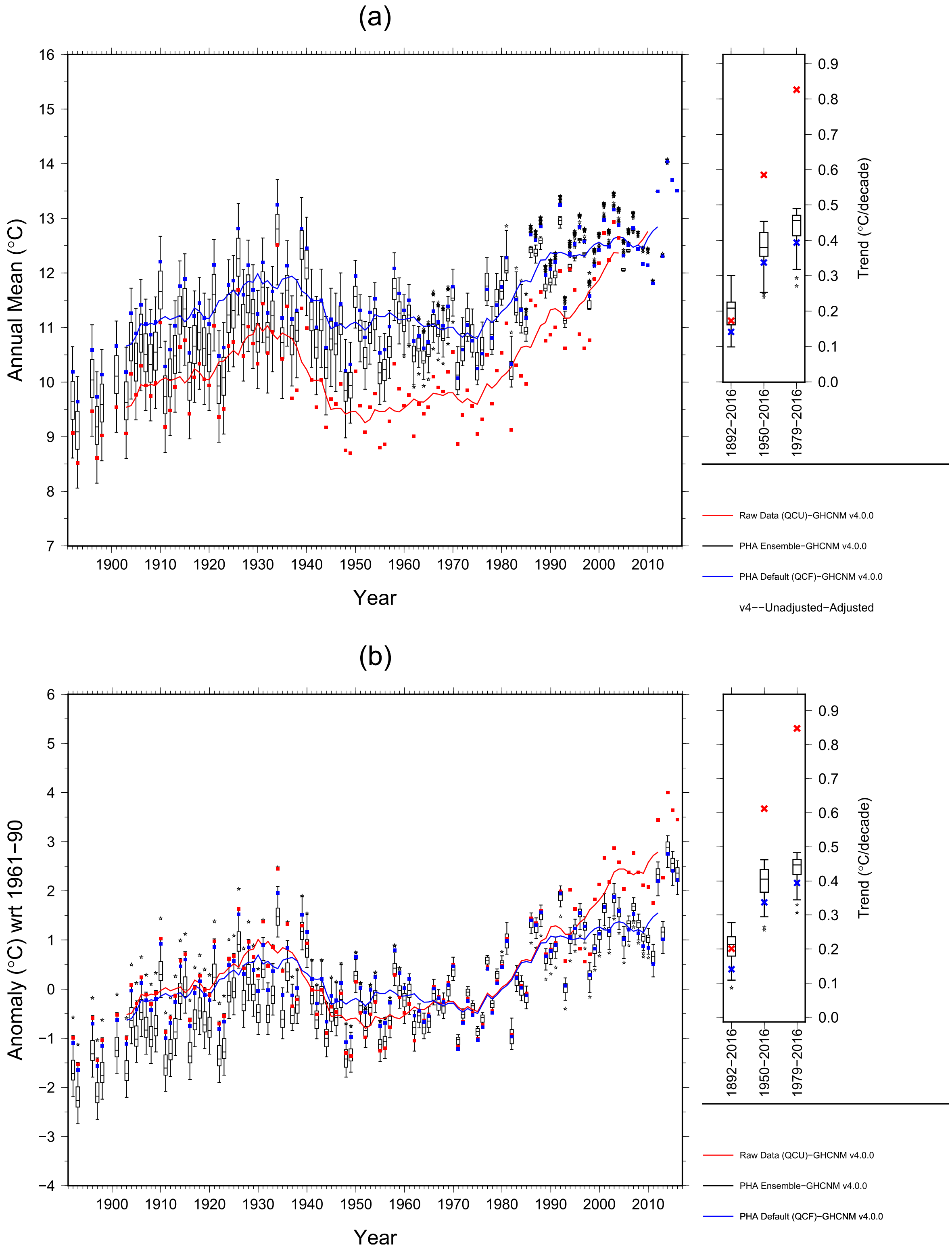

FIG. 6. Parametric uncertainty for (a) mean annual temperature and (b) annual anomalies for Reno. Box-and-whiskers plots depict the range of ensemble values. Red is unadjusted (raw) data; blue is the default PHA result. Unadjusted and default PHA time series shown as 9-yr running averages. As in (a) and (b), but for (c) total homogenization uncertainty and (d) total homogenization plus anomaly uncertainty for annual mean temperatures at Reno.

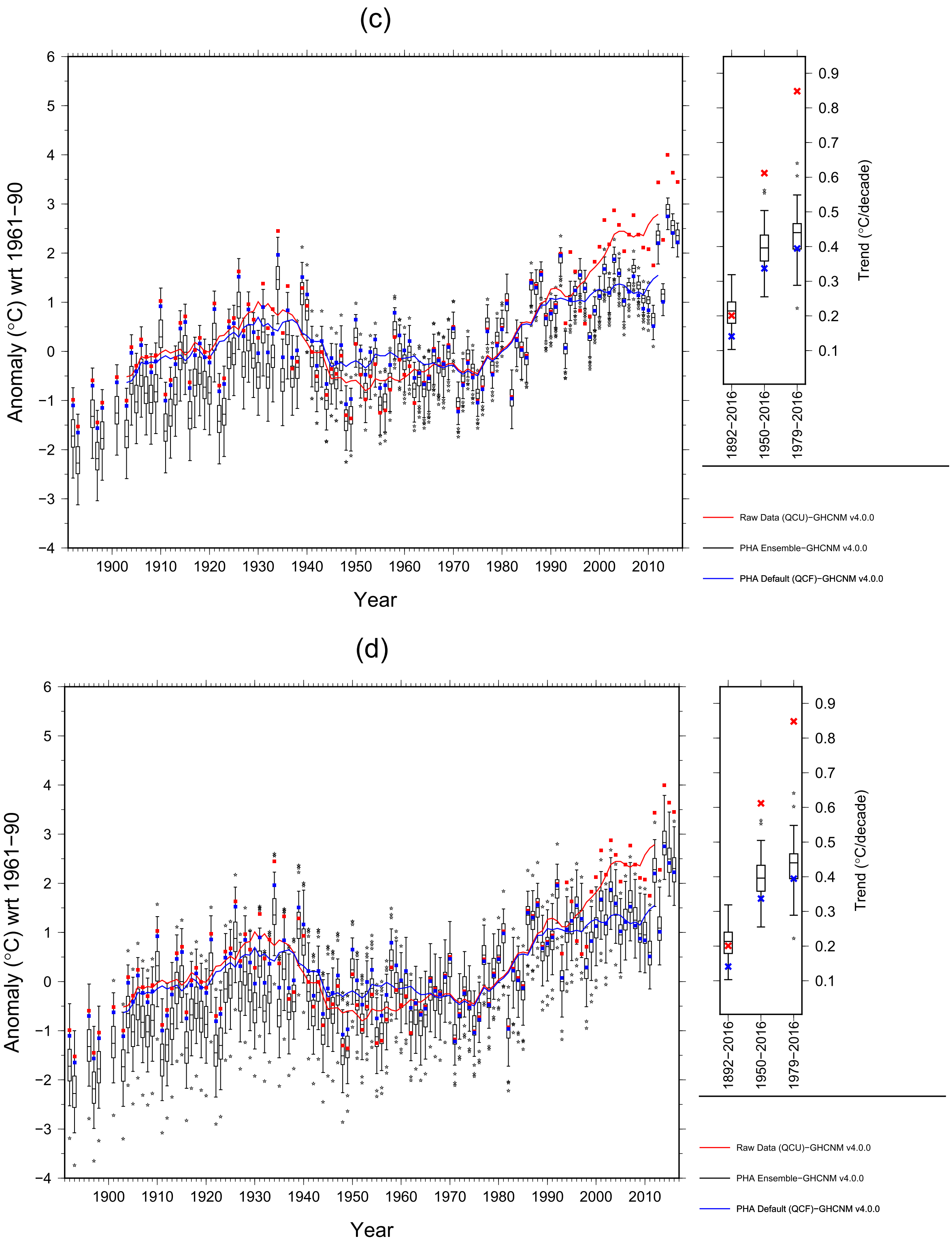

FIG. 6. (*Continued*)

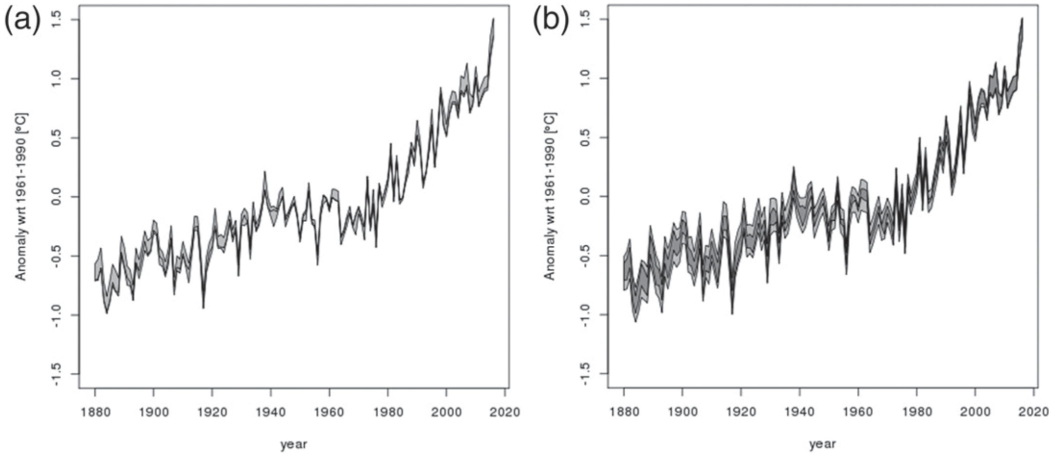

FIG. 7. (a) Parametric uncertainty of the global LSAT and (b) parametric uncertainty (dark gray) plus total homogenization uncertainty (light gray).

cancel out when averaged spatially, as shown in Fig. 7a, which provides the global land annual average parametric uncertainty.

The inability of the PHA (or other algorithms) to resolve small systematic shifts in the temperature series is evident in Fig. 3 and has been discussed in previous analyses (e.g., Brohan et al. 2006; Menne and Williams 2009; Menne et al. 2009). Though small, these undetected shifts comprise an important part of the overall uncertainty budget for every station record. They can also lead to errors in computing regional averages when they do not have an expected mean of zero, which is evidently the case for GHCN-monthly data (Fig. 3; see also Lawrimore et al. 2011). Using some national homogenization efforts as case studies, Brohan et al. (2006) reckoned that undetected shifts in the CRUTEM dataset roughly follow a Gaussian distribution with a mean of zero and a standard deviation of 0.4 and that the average interval between undetected shifts was on the order of 40 years. In the GHCNm v4 data, we have the benefit of a systematic screening for shifts in each temperature series. In particular, we can more directly estimate the size and frequency of the so-called missing middle of the shift distribution (Thorne et al. 2016). Based on the frequency distribution of detected shifts shown in Fig. 3, we estimate that the missed shifts in GHCNm v4 have a mean of about −0.01°C and a standard deviation of 0.2 with an average frequency of occurrence of about 1 in 50 years. The estimated frequency of missed breaks and detected breaks is shown in Fig. 8.

Accordingly, uncertainty from missed adjustments are quantified by adding shift adjustments to stations series using magnitudes and frequency drawn from the missing middle of detected shifts. Specifically, the magnitudes are selected from the normal distribution $N(0.01, 0.2)$ with a rate of occurrence drawn from a Poisson distribution having an average frequency of 1 in 50 years, or 0.0016 month^{-1}, and each month has an equal probability of being the date of the missed shift. This approach uses the global distribution of detected shifts to estimate missed shifts everywhere. Admittedly, there could be future refinements to this approach that include, for example, determining the potential missed shifts regionally since the nature of shifts will vary nationally according to station management policies. In addition, the potential for missed breaks is higher in low-station-

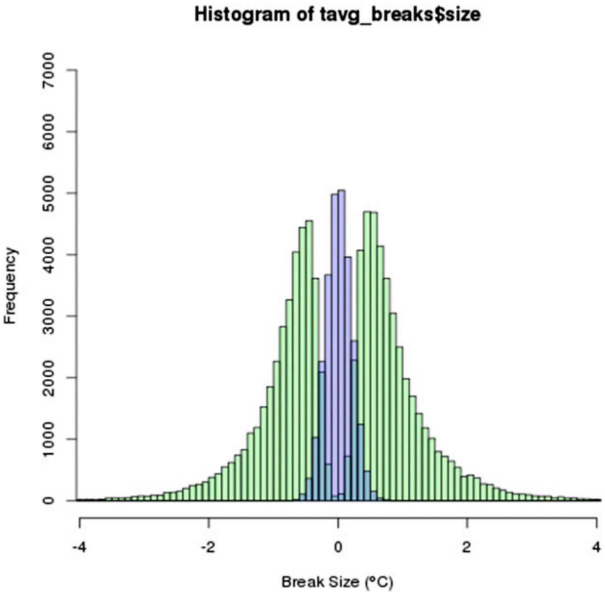

FIG. 8. As in Fig. 3, but with an estimate of the distribution of missed breaks. The green bars are the breaks detected by the PHA. The purple bars show the estimated distribution of the missing middle. The true distribution of breaks in the data is thought to be close to the combination of the two histograms.

density areas where the correlation between neighboring stations may be lower and thus impact the signal-to-noise ratio for shift detection. The total homogeneity uncertainty for the Reno mean annual temperature series uncertainties is shown in Fig. 6c, and the global average land surface air temperature homogeneity error is shown in Fig. 7b.

b. *Station anomaly uncertainty (normals uncertainty)*

To produce grid box and large area averages, monthly mean temperatures from each station are often converted into anomalies (departures from a base period mean) that have large correlation decay scales (Jones et al. 1997). An average anomaly for each, say, 5° × 5° grid box is often calculated as the arithmetic mean of all available station-based anomalies within the bounding box. This approach, sometimes called the climate anomaly method (e.g., Jones et al. 1986a), requires that a station have at least some observations during the base period in question, usually at least for half of the 30-yr period (Morice et al. 2012). The minimum number of values required to compute an anomaly for use in the NOAA GST dataset (NOAAGlobalTemp; Vose et al. 2014) has traditionally been 20 out of 30 values for the base period (1961–90) for each calendar month. Land-area grid boxes without any station anomalies are left missing unless some kind of interpolation is used.

Calculation of station anomalies adds further uncertainty to the data. As Brohan et al. (2006) note, the sampling distribution of the base period mean (climatological normal) has a standard deviation of $\sigma_w/\sqrt{P}$, where σ_w is the standard deviation of the monthly temperatures for a particular calendar month and P is the number of years in the base period. We use this to calculate the sampling uncertainty for GHCN station anomalies ε_N. In addition, in order to maximize the number of stations with anomalies over time, we produce normals estimates for stations with partially or completely missing monthly data during the 1961–90 base period. Estimates are generated as in Menne et al. (2009) using an optimal interpolation technique known informally as ‘‘fill in the network’’ (FILNET). The FILNET procedure iterates to find a set of neighboring correlated series for each station series requiring estimates (the target) that minimizes the confidence limits for the difference between the target values and estimates of these values derived using neighboring values. The difference between the target and neighbor average is used as an offset in the interpolation to account for climatological differences between the target and neighbors.

As shown in Fig. 9, using estimates for base period averages greatly expands the number of stations used to compute global land anomalies. However, when such base period estimates are used, an additional error term is required to account for uncertainty in the monthly estimates. This error term was derived from a resampling exercise in which successive numbers of values were censored for stations with complete data during the 1961–90 base period, with the censored data then being estimated from surrounding stations. The root-mean-square error was then calculated between the observed 30-yr mean and the mean with varying numbers of estimated values during the base period. The normal estimate error is then calculated as a function of latitude and number of missing years during the base period. The two error terms (sampling and estimate) are treated as independent and $\varepsilon_{N+E} = \sqrt{\varepsilon_N^2 + \varepsilon_E^2}$, and ε_E is zero when no estimates are required. The anomaly uncertainty is also considered to be uncorrelated in space. Notably, for stations with partial records during the base period, the use of estimates can reduce the overall anomaly uncertainty in spite of the error associated with the estimates.

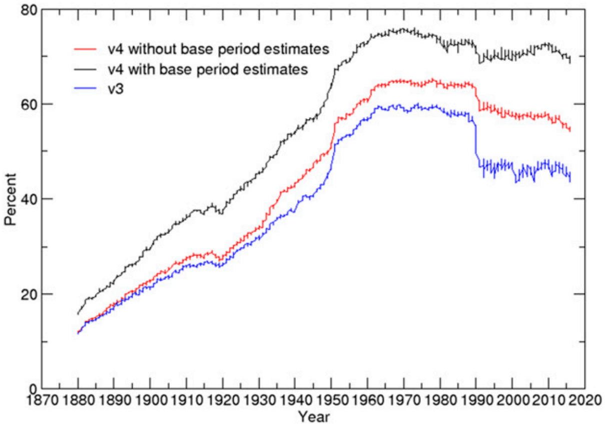

FIG. 9. Time series of the percent of 5° × 5° grid boxes with any land area that contains temperature anomalies from at least one land station.

The combined homogenization and anomaly error for Reno is shown in Fig. 6d. On the whole, Fig. 6 indicates that the unadjusted Reno mean annual temperatures often fall outside the combined range of homogenization and anomaly uncertainty as does the trend since 1950 and since 1979 due in large part to the localized upward shifts (localized jumps) since 1970 thought to be associated with the growing urban heat island (UHI) signal around the airport site. Regarding the detection of UHI, we have evaluated some additional prominent UHI examples: Phoenix, Arizona, and Shenzhen, China. These additional case studies, like Reno, suggest that when homogenization testing is conducted on serial monthly data as we do in GHCNm v4, the target-neighbor difference series are at least as good or better treated as a series of step changes with like signs rather than as a trend. Consequently, at least in these cases, the PHA is able to detect and account for the growth of a local UHI signal. Given that these cases are consistent with the Hausfather et al. (2013) study that the PHA is effectively accounting for local UHI signals in the United States, we do not include a separate uncertainty component for UHI, but rather include UHI as part of the homogenization uncertainty.

c. *Instrument exposure bias from nonstandard screens*

We adopt exactly the same approach to nonstandard screening discussed in Morice et al. (2012), which was also used in Brohan et al. (2006). Their work was in turn based on the exposure model described in Folland et al. (2001), which was itself based on Parker (1994). The model is intended to account for uncertainty stemming from the introduction of new types of thermometer screens, especially during the period of the late nineteenth and early twentieth centuries. Unlike the previous uncertainty components, the exposure uncertainty is applied at the regional level. For each ensemble member, a single random number is drawn from a normal distribution that reflects the 1σ exposure uncertainty for the latitude range. For grid boxes in the 20°S–20°N range, exposure is associated with a 1σ uncertainty of 0.2°C before 1930, which decreases linearly to zero by 1950. Outside of that range, the 1σ uncertainty is 0.1°C prior to 1900, which decreases linearly to zero by 1930.

d. *Gridbox sampling error*

Even though estimates are used, where necessary, to compute a 30-yr normal, the number of stations within any particular 5° × 5° grid box still varies considerably by region. In addition, those anomalies that are available in any particular grid box may deviate from the spatial average that would result from denser spatial sampling. The gridbox sampling error seeks to quantify the uncertainty of gridbox average throughout the period of record. Jones et al. (1997) described a method for calculating this uncertainty, which they determine to be a function of the number of stations with values in the grid box, the variability of the temperature within the grid box, and the correlation between the available stations. More specifically, Jones et al. (1997) used the following equation to calculate the gridbox sampling error:

$$\mathrm{SE}^2 = \frac{\overline{\sigma_i^2}\,\bar{r}(1-\bar{r})}{1+(n-1)\bar{r}}(x), \qquad (1)$$

where $\overline{\sigma_i^2}$ is the mean standard deviation in the grid box, n is the number of stations in the grid box, and $\bar{r}$ is the average interstation correlation. Like Brohan et al. (2006) and Morice et al. (2012), we use this approach to calculate the gridbox sampling error when we calculate gridbox averages. Although the GHCNm v4 ensemble members do have some variation in their anomalies for grid boxes, the global mean error associated with gridbox sampling is essentially the same for each member.

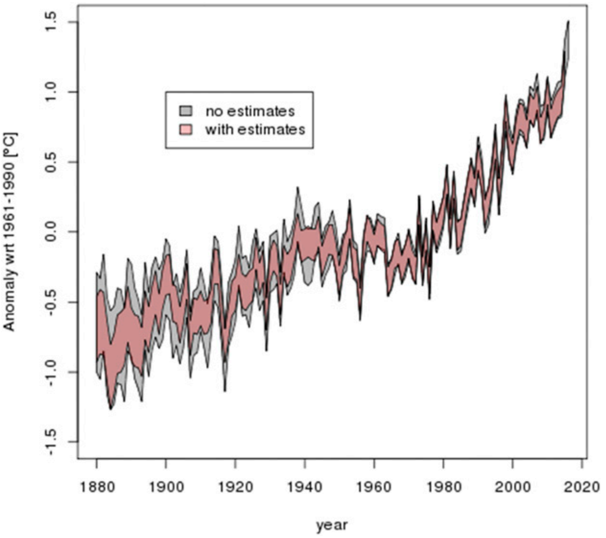

FIG. 10. Spatial coverage uncertainty with and without estimates for missing station anomalies during the 1961–90 normals period.

e. *Spatial coverage uncertainty*

The last major uncertainty that we address is the spatial coverage uncertainty. As shown in Fig. 9, the percentage of land cover with station anomalies is improved with the use of estimated normals but is still lower during the early decades of the period of record. This percentage is relatively constant after about 1950, but parts of the high latitudes, Africa, and South America remain less sampled than other regions. The spatial coverage uncertainty addresses the error that can arise from a global mean land surface air temperature anomaly calculated from spatially incomplete data. Like Brohan et al. (2006), we use the NCEP–NCAR reanalysis (Kalnay et al. 1996) to compute the root-mean-square error between a global land average calculated from spatially complete reanalysis data compared to one using a data mask determined by the available GHCNm v4 monthly temperature anomalies. For each data month, we compute the difference between the complete and incomplete reanalysis means for each of the 30 years from 1986 to 2015. This period was selected to include the period after 1998 when the consequences of excluding parts of the Arctic from the global mean calculation are significant (Cowtan and Way 2014). The spatial uncertainty for mean annual values using the GHCNm v4 coverage is shown in Fig. 10. In Fig. 11, we illustrate the global impact of each of the major uncertainties covered in this section.

7. Comparison to other datasets and methods

Comparison to other datasets is done to provide an assessment of the structural uncertainty of the data,

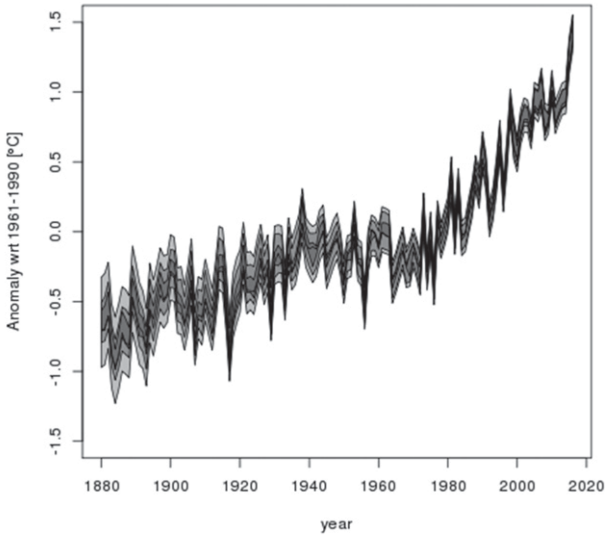

FIG. 11. Total uncertainty for GHCNm v4 mean annual global LSAT anomalies (parametric plus missed breaks plus anomaly plus spatial coverage uncertainty). Darker greys show homogenization uncertainties (parametric and missed breaks) and the lighter greys show anomaly and spatial coverage uncertainties. The uncertainties are displayed as cumulative so the uncertainty bounds depicted in each lighter shade includes the uncertainty of the darker shades.

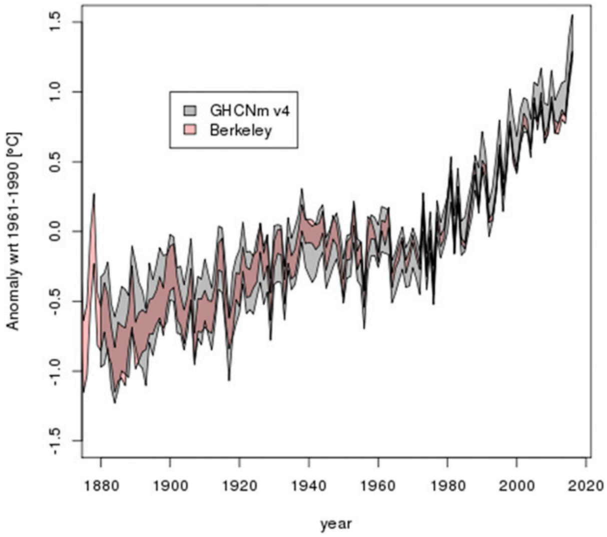

FIG. 12. Mean annual global LSAT uncertainties for GHCNm v4 and Berkeley Earth. Dark shades of pink indicate areas of overlap.

which refers to the uncertainty that is revealed when using independent methods to address the same problem (Thorne et al. 2005). To begin, we compare the annual time series with uncertainty estimates for GHCNm v4 to those provided for CRUTEM, version 4 (CRUTEM4; Jones et al. 2012; Morice et al. 2012) and the Berkeley Earth Surface Temperature dataset (Rohde et al. 2013). The annual average time series and their associated uncertainties are shown in Figs. 12 and 13. The figures indicate that the uncertainties for the three datasets overlap in all years. Not surprisingly, the magnitude of uncertainties between GHCNm v4 and CRUTEM4.5.0.0 are similar since they are based on a more or less common set of factors (though the GHCM v4 uncertainty range is slightly less than CRUTEM). In contrast, the uncertainty estimates for Berkeley Earth are considerably smaller than GHCNm and CRUTEM.

Berkeley Earth has also recently provided an analysis of their raw data spatial averages. These averages are produced without parsing each station series into separate segments that begin and end at times of apparent shifts. Using their approach known as the scalpel method (Rohde et al. 2013), the step of identifying the timing of shifts has the same goal as the PHA. The impact of the shifts can then be accounted for in the averaging approach. As shown in Fig. 14, the Berkeley Earth group finds that the distribution of shifts in data collectively cause a negative cool bias in the early part of the record up to the period around World War II, similar to GHCNm. Their scalpel method yields an increase in the global mean land surface air temperature of about 0.2°C over the raw data from the late nineteenth century to about 1950. After 1950, parsing the data with the scalpel method has relatively little impact on the global land average anomalies but does lead to a slight decrease relative to the raw data. This small decrease associated with their parsing approach may be the reason the Berkeley Earth anomalies are near the low range of the GHCNm v4 uncertainty for the last couple of decades.

In Fig. 15, we also compare the shifts identified by the published configuration of the PHA (Menne and

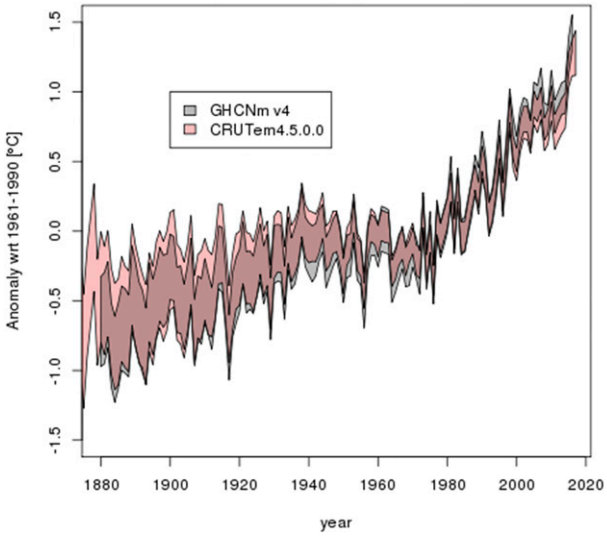

FIG. 13. Mean annual global LSAT uncertainties for GHCNm v4 and CRUTEM4.5.0.0. Dark shades of pink indicate areas of overlap.

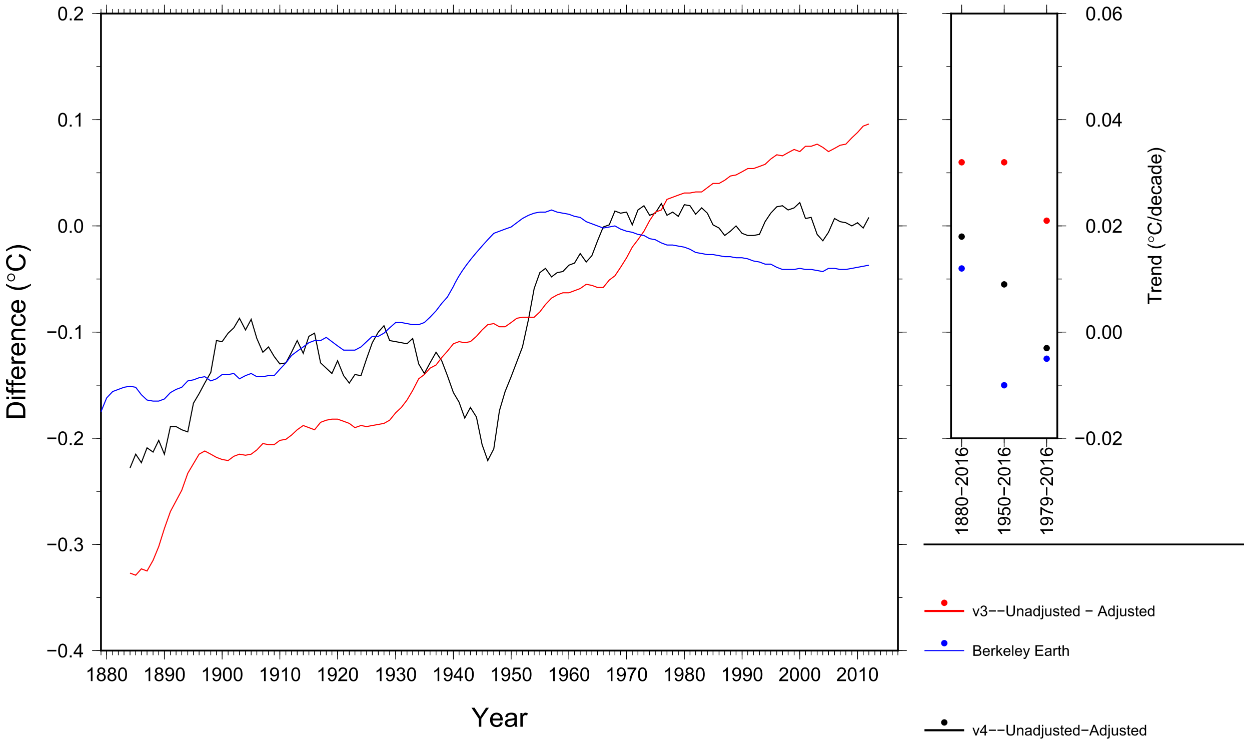

FIG. 14. Global mean difference between adjusted and unadjusted GHCNm v3, v4, and Berkeley Earth datasets.

Williams 2009) to an independent method of identifying shifts called the Bayes factor algorithm (BFA; Zhang et al. 2012). As the histogram shows, the PHA and BFA methods find a similar number of shifts (about 71 000 for the PHA and 74 000 for the BFA) in the GHCNm v4 station series. Moreover, the BFA breaks coincide with PHA breaks about two-thirds of the time, and when coincident, the detected shifts are of the same sign 97% of the time. The mean (median) absolute difference in the estimated magnitude of shifts that both methods identify is about 0.21°C (0.13°C) with the BFA finding fewer smaller breaks than the PHA. This is largely a consequence of the PHA being run with station history information that informs the algorithm of the timing of potential shifts in the United States related to station moves as well as instrument and time of observation changes. The use of station histories allows the PHA to identify smaller changes than is possible in the absence of documented changes.

Finally, in Figs. 16 and 17 we compare GHCNm v4 station data averages to some smaller datasets that have been homogenized and averaged independently at the national level. Two examples are provided: one for Switzerland based on Begert and Frei (2017) and another for Australia based on the Australian Climate Observations Reverence Network–Surface Air Temperature (ACORN-SAT) dataset (Trewin 2013). In the case of Switzerland, the GHCNm adjustment process yields adjusted time series that are in closer agreement with the Begert and Frei series than the GHCNm unadjusted series, even though the station composition differs considerably between v3 and v4. Notably, the GHCNm v3 adjustments increase the long-term trend

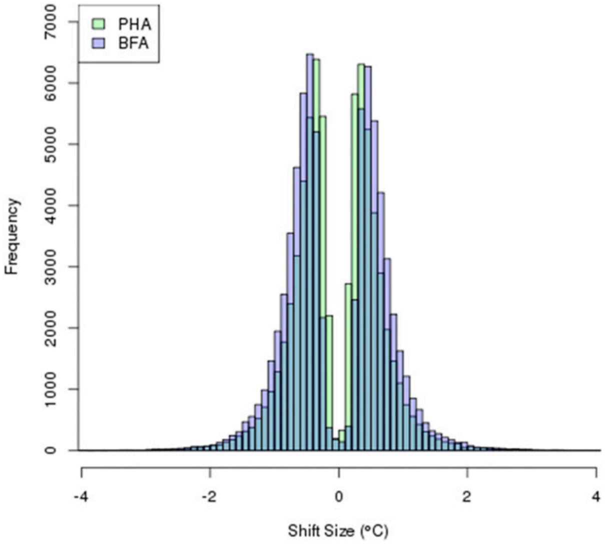

FIG. 15. Histogram of shifts identified by the PHA and BFA homogenization algorithms.

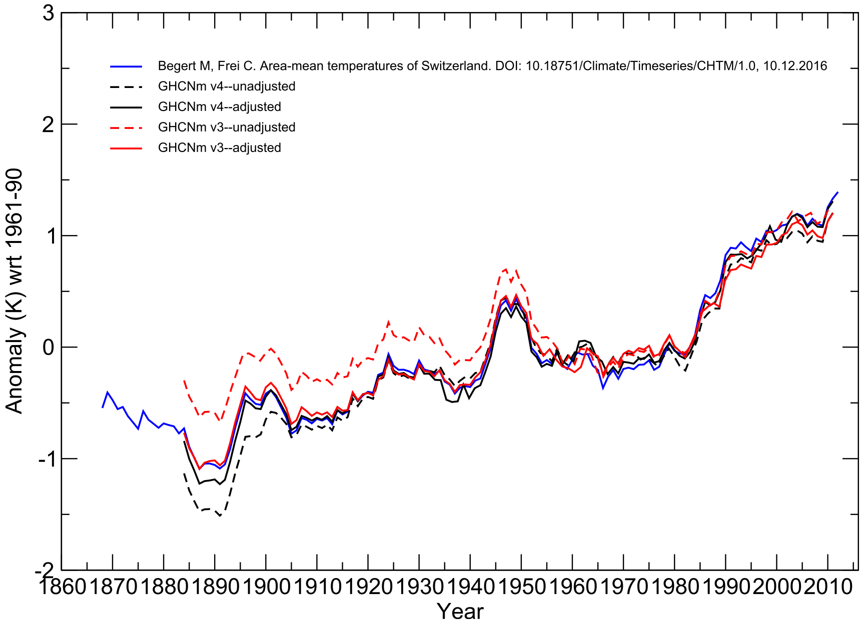

FIG. 16. Mean annual temperature for Switzerland. Anomalies are shown using a 9-yr moving average.

for Switzerland whereas the v4 adjustments decrease the long-term trend, but both lead to increased consistency with the Begert and Frei (2017). For Australia, both the v3 and v4 adjustments reduce the magnitude of anomalies in the early record, but much more so in the case v3. As a result, the v4 adjusted long-term trend more closely resembles the ACORN-SAT trend, but GHCNm shows a somewhat higher trend in the last two decades.

Nationally homogenized station series like these Swiss and Australian examples are used directly in building the CRUTEM and CMA-LSAT datasets. Both Jones et al. (2012) and Xu et al. (2018) advocate for using

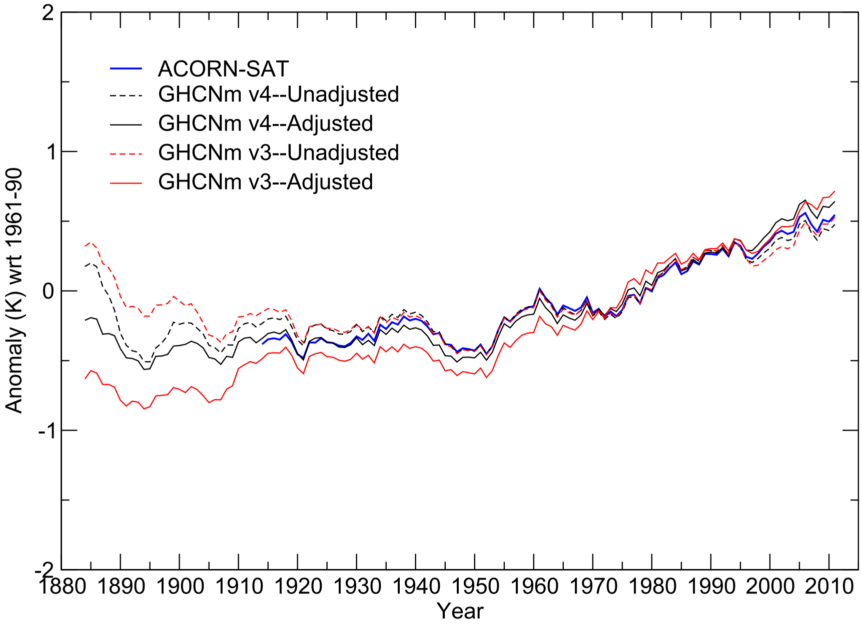

FIG. 17. Mean annual temperature for Australia. Anomalies are shown using a 9-yr moving average.

datasets homogenized at the national level wherever possible in building their global land surface air temperature datasets. Their reasoning is that local researchers are in a better position to know the details of station changes and may have access to additional station data not available in global archives, which may improve homogenization efforts. On the other hand, Berkeley Earth and GHCNm use uniform and comprehensive approaches to homogenization and provide both unadjusted and adjusted versions of the data, which makes the impact of adjustments easier to trace, and the independent approaches provide some measure of the structural uncertainty in the data. Our independent homogenization approach also facilitates keeping the station series updated consistently since nationally homogenized datasets may be one-off efforts or updated on an irregular basis. We argue therefore that there are advantages to both rationales, and the approaches used to produce CRUTEM and CMA-LSAT are perhaps best viewed as complementary to the methods used in Berkeley Earth and GHCNm. Comparisons like those shown in Figs. 16 and 17 provide some insight into whether national groups are making similar adjustments to the global-scale applications as GHCNm. These two national cases, as well as comparisons at the global mean level, suggest that these independent adjustments are leading to greater consistency among the different datasets.

8. Conclusions

In summary, GHCNm v4 is made up of more than 25 000 land-based station records, an increase of 18 000 stations from versions 2 and 3. The larger number of stations relative to earlier versions as well as the expanded use of stations in the calculation of anomalies improves spatial coverage of the dataset throughout the period of record. The majority of the monthly average land surface air temperatures—approximately 75%—that make up GHCNm, version 4, station records are calculated directly from GHCNd using World Meteorological Organization standards. This improves the traceability of how monthly means are calculated. The monthly averages are routinely updated and quality controlled for random errors. Systematic errors are addressed via homogenization. In addition, uncertainties for each station series are provided by running the homogenization algorithm as an ensemble and then adding additional uncertainties associated with incomplete homogenization and conversion of the time series to anomalies. Further uncertainties are addressed at the regional level; the most important of which is spatial coverage uncertainty. Uncertainties provided at the station level can be summed consistently up to the regional and global averages. This ensemble approach can be combined with ensembles of the Extended Reconstructed Sea Surface Temperature (ERSST) dataset (Liu et al. 2015; Huang et al. 2016) for a more comprehensive assessment of global surface temperature uncertainties.

Compared to GHCNm v3, homogenization has a smaller impact on v4 data, but the adjustments lead to greater consistency between GHCNm v3 and v4 than when the unadjusted data are compared, and the v3 series falls within the calculated range of uncertainty for the v4 dataset. Adjustments for shifts also lead to greater consistency between datasets homogenized independently at the national level. Likewise, the uncertainty ranges of the other major independent land surface air temperature datasets overlap with the GHCNm v4 uncertainty ranges throughout the period of record, and the impact of accounting for shifts is broadly consistent between GHCNm and the Berkeley Earth dataset. Finally, it should be noted that this initial release of GHCNm v4 contains only mean monthly temperature. Mean monthly maximum and mean monthly minimum temperatures will be provided in a future release.

Acknowledgments. Thanks to Deke Arndt and three anonymous reviewers for their comments on earlier versions of this manuscript. Thanks also to Colin Gallagher for work with the lead author on case studies evaluating the nature of the urban heat island signal in monthly temperature series at the SAMSI/IMAGe workshop in Boulder, Colorado, in July 2014. Data for GHCNm v4 are available online (at ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/v4). Further technical documentation and code will be available and a DOI for the dataset will be minted upon acceptance of this manuscript for publication.

REFERENCES

Alexandersson, H., 1986: A homogeneity test applied to precipitation data. *J. Climatol.*, **6**, 661–675, https://doi.org/10.1002/joc.3370060607.

Begert M. and C. Frei, 2017: Area-mean temperatures of Switzerland. MeteoSwiss, accessed 10 December 2017, https://doi.org/10.18751/Climate/Timeseries/CHTM/1.0.

Bradley, R. S., P. M. Kelly, P. D. Jones, H. F. Diaz, and C. M. Goodess, 1985: A climatic data bank for Northern Hemisphere land areas, 1851–1980. U.S. Department of Energy Tech. Rep. TR017, 335 pp., http://www.cru.uea.ac.uk/publications/crurp.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.*, **111**, D12106, https://doi.org/10.1029/2005JD006548.

Chenoweth, M., 1993: Nonstandard thermometer exposures at U.S. cooperative weather stations during the late nineteenth century. *J. Climate*, **6**, 1787–1797, https://doi.org/10.1175/1520-0442(1993)006<1787:NTEAUC>2.0.CO;2.

Clayton, H. H. 1927: *World Weather Records*. Smithsonian Miscellaneous Collections, Vol. 79, Smithsonian Press, 1196 pp.

Cowtan, K., and R. G. Way, 2014: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quart. J. Roy. Meteor. Soc.*, **140**, 1935–1944, https://doi.org/10.1002/qj.2297.

Diamond, H. J., and Coauthors, 2013: U.S. Climate Reference Network after one decade of operations: Status and assessment. *Bull. Amer. Meteor. Soc.*, **94**, 485–498, https://doi.org/10.1175/BAMS-D-12-00170.1.

Durre, I., M. J. Menne, and R. S. Vose, 2008: Strategies for evaluating quality control procedures. *J. Appl. Meteor. Climatol.*, **47**, 1785–1791, https://doi.org/10.1175/2007JAMC1706.1.

Folland, C. K., and Coauthors, 2001: Global temperature change and its uncertainties since 1861. *Geophys. Res. Lett.*, **28**, 2621–2624, https://doi.org/10.1029/2001GL012877.

Hansen, J., R. Ruedy, J. Glascoe, and M. Sato, 1999: GISS analysis of surface temperature change. *J. Geophys. Res.*, **104**, 30 997–31 022, https://doi.org/10.1029/1999JD900835.

——, ——, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, https://doi.org/10.1029/2010RG000345.

Hausfather, Z., M. J. Menne, C. N. Williams, T. Masters, R. Broberg, and D. Jones, 2013: Quantifying the impact of urbanization on U.S. Historical Climatology Network temperature records. *J. Geophys. Res. Atmos.*, **118**, 481–494, https://doi.org/10.1029/2012JD018509.

——, K. Cowtan, M. J. Menne, and C. N. Williams Jr., 2016: Evaluating the impact of U.S. Historical Climatology Network homogenization using the U.S. Climate Reference Network. *Geophys. Res. Lett.*, **43**, 1695–1701, https://doi.org/10.1002/2015GL067640.

Hawkins, D. M., 1976: Point estimation of the parameters of a piecewise regression model. *Appl. Stat.*, **25**, 51–57, https://doi.org/10.2307/2346519.

Huang, B., and Coauthors, 2016: Further exploring and quantifying uncertainties for Extended Reconstructed Sea Surface Temperature (ERSST) version 4 (v4). *J. Climate*, **29**, 3119–3142, https://doi.org/10.1175/JCLI-D-15-0430.1.

Jones, P. D., 1994: Hemispheric surface air temperature variations: A reanalysis and an update to 1993. *J. Climate*, **7**, 1794–1802, https://doi.org/10.1175/1520-0442(1994)007<1794:HSATVA>2.0.CO;2.

——, and A. Moberg, 2003: Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001. *J. Climate*, **16**, 206–223, https://doi.org/10.1175/1520-0442(2003)016<0206:HALSSA>2.0.CO;2.

——, S. C. B. Raper, R. S. Bradley, H. F. Diaz, P. M. Kelly, and T. M. L. Wigley, 1986a: Northern Hemisphere surface air temperature variations: 1851–1984. *J. Climate Appl. Meteor.*, **25**, 161–179, https://doi.org/10.1175/1520-0450(1986)025<0161:NHSATV>2.0.CO;2.

——, ——, and T. M. L. Wigley, 1986b: Southern Hemisphere surface air temperature variations: 1851–1984. *J. Climate Appl. Meteor.*, **25**, 1213–1230, https://doi.org/10.1175/1520-0450(1986)025<1213:SHSATV>2.0.CO;2.

——, T. J. Osborn, and K. R. Briffa, 1997: Estimating sampling errors in large-scale temperature averages. *J. Climate*, **10**, 2548–2568, https://doi.org/10.1175/1520-0442(1997)010<2548:ESEILS>2.0.CO;2.

——, 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, https://doi.org/10.1029/2011JD017139.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kintisch, E., 2014: Climate outsider finds missing global warming. *Science*, **344**, 348–348, https://doi.org/10.1126/science.344.6182.348.

Klein Tank, A. M. G., and Coauthors, 2002: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *Int. J. Climatol.*, **22**, 1441–1453, https://doi.org/10.1002/joc.773.

Lanzante, J. R., 1996: Resistant, robust and non-parametric techniques for analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.*, **16**, 1197–1226, https://doi.org/10.1002/(SICI)1097-0088(199611)16:11<1197::AID-JOC89>3.0.CO;2-L.

Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie, 2011: An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *J. Geophys. Res.*, **116**, D19121, https://doi.org/10.1029/2011JD016187.

Liu, W., and Coauthors, 2015: Extended Reconstructed Sea Surface Temperature version 4 (ERSST.v4): Part II. Parametric and structural uncertainty estimations. *J. Climate*, **28**, 931–951, https://doi.org/10.1175/JCLI-D-14-00007.1.

Menne, M. J., and C. N. Williams Jr., 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate*, **22**, 1700–1717, https://doi.org/10.1175/2008JCLI2263.1.

——, C. N. Williams Jr., and R. S. Vose, 2009: The United States Historical Climatology Network Monthly Temperature Data, version 2. *Bull. Amer. Meteor. Soc.*, **90**, 993–1007, https://doi.org/10.1175/2008BAMS2613.1.

——, I. Durre, R. S. Vose, B. G. Gleason, and T. Houston, 2012: An overview of the Global Historical Climatology Network Daily dataset. *J. Atmos. Oceanic Technol.*, **29**, 897–910, https://doi.org/10.1175/JTECH-D-11-00103.1.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, https://doi.org/10.1029/2011JD017187.

NOAA, 2007: Datzilla error and tracking system quick reference fact sheet. National Weather Service Rep., 2 pp., http://training.weather.gov/nwstc/DATAACQ/datacq4mgrs/lesson2/Datzillaref.pdf.

Parker, D. E., 1994: Effects of changing exposure of thermometers at land stations. *Int. J. Climatol.*, **14**, 1–31, https://doi.org/10.1002/joc.3370140102.

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

Quayle, R. G., T. C. Peterson, A. N. Basist, and C. S. Godfrey, 1999: An operational near-real-time global temperature index. *Geophys. Res. Lett.*, **26**, 333–335, https://doi.org/10.1029/1998GL900297.

Rennie, J. J., 2015: International Surface Temperature Initiative global land surface databank, version 1.1.0. NOAA Tech. Rep., 10 pp., ftp://ftp.ncdc.noaa.gov/pub/data/globaldatabank/monthly/stage3/ISTI_Databank_Technical_Report_v1.1.0.pdf.

——, and Coauthors, 2014: The International Surface Temperature Initiative global land surface databank: Monthly temperature data release description and methods. *Geosci. Data J.*, **1**, 75–102, https://doi.org/10.1002/gdj3.8.

Rohde, R., and Coauthors, 2013: A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinf. Geostat. Overview*, **1**, https://doi.org/10.4172/2327-4581.1000103.

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore, 2008: Improvements to NOAA's historical Merged Land–Ocean Surface Temperature analysis (1880–2006). *J. Climate*, **21**, 2283–2296, https://doi.org/10.1175/2007JCLI2100.1.

Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears, 2005: Uncertainties in climate trends: Lessons from upper-air temperature records. *Bull. Amer. Meteor. Soc.*, **86**, 1437–1442, https://doi.org/10.1175/BAMS-86-10-1437.

——, and Coauthors, 2011: Guiding the creation of a comprehensive surface temperature resource for twenty-first-century climate science. *Bull. Amer. Meteor. Soc.*, **92**, ES40–ES47, https://doi.org/10.1175/2011BAMS3124.1.

——, and Coauthors, 2016: Reassessing changes in diurnal temperature range: A new data set and characterization of data biases. *J. Geophys. Res. Atmos.*, **121**, 5115–5137, https://doi.org/10.1002/2015JD024583.

Trewin, B., 2010: Exposure, instrumentation, and observing practice effects on land temperature measurements. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 490–506, https://doi.org/10.1002/wcc.46.

——, 2013: A daily homogenized temperature data set for Australia. *Int. J. Climatol.*, **33**, 1510–1529, https://doi.org/10.1002/joc.3530.

Venema, V. K. C., and Coauthors, 2012: Benchmarking homogenization algorithms for monthly data. *Climate Past*, **8**, 89–115, https://doi.org/10.5194/cp-8-89-2012.

Vose, R. S., R. L. Schmoyer, P. M. Steurer, T. C. Peterson, R. Heim, T. R. Karl, and J. Eischeid, 1992: The Global Historical Climatology Network: Long-term monthly temperature, precipitation, sea level pressure, and station pressure data. Carbon Dioxide Information Analysis Center Tech. Rep. ORNL/CDIAC-53, 325 pp.

——, S. Applequist, M. J. Menne, C. N. Williams Jr., and P. W. Thorne, 2012: An intercomparison of temperature trends in the U.S. Historical Climatology Network and recent atmospheric reanalyses. *Geophys. Res. Lett.*, **39**, L10703, https://doi.org/10.1029/2012GL051387.

——, and Coauthors, 2014: Improved historical temperature and precipitation time series for U.S. climate divisions. *J. Appl. Meteor. Climatol.*, **53**, 1232–1251, https://doi.org/10.1175/JAMC-D-13-0248.1.

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

Willmott, C. J., S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell, and C. M. Rowe, 1985: Statistics for the evaluation and comparisons of models. *J. Geophys. Res.*, **90**, 8995–9005, https://doi.org/10.1029/JC090iC05p08995.

WMO, 2017: WMO guidelines on the calculation of climate normal. WMO/TD-1203, 29 pp.

Xu, W., and Coauthors, 2018: A new integrated and homogenized global monthly land surface air temperature dataset for the period since 1900. *Climate Dyn.*, **50**, 2513–2536, https://doi.org/10.1007/s00382-017-3755-1.

Zhang, J., W. Zheng, and M. J. Menne, 2012: A Bayes factor model for detecting artificial discontinuities via pairwise comparisons. *J. Climate*, **25**, 8462–8474, https://doi.org/10.1175/JCLI-D-12-00052.1.