Uncertainties, Trends, and Hottest and Coldest Years of U.S. Surface Air Temperature since 1895: An Update Based on the USHCN V2 TOB Data

SAMUEL S. P. SHEN AND CHRISTINE K. LEE

Department of Mathematics and Statistics, San Diego State University, San Diego, California

JAY LAWRIMORE

NOAA/National Climatic Data Center, Asheville, North Carolina

(Manuscript received 23 February 2011, in final form 17 December 2011)

ABSTRACT

This paper estimates the sampling error variances of gridded monthly U.S. Historical Climatology Network, version 2 (USHCN V2), time-of-observation-biases (TOB)-adjusted data. The analysis of mean surface air temperature (SAT) assesses uncertainties, trends, and the rankings of the hottest and coldest years for the contiguous United States in the period of 1895–2008. Data from the USHCN stations are aggregated onto a 2.5° × 3.5° latitude–longitude grid by an arithmetic mean of the stations inside a grid box. The sampling error variances of the gridded monthly data are estimated for every month and every grid box with data. The gridded data and their sampling error variances are used to calculate the contiguous U.S. averages and their trends and associated uncertainties. The sampling error variances are smaller (mostly less than 0.2°C^2) over the eastern United States, where the station density is greater and larger (with values of 1.3°C^2 for some grid boxes in the earlier period) over mountain and coastal areas. In the period of 1895–2008, every month from January to December has a positive linear trend. February has the largest trend of 0.162°C (10 yr)$^{-1}$, and September has the smallest trend at 0.020°C (10 yr)$^{-1}$. The three hottest (coldest) years measured by the mean SAT over the United States were ranked as 1998, 2006, and 1934 (1917, 1895, and 1912).

1. Introduction

Many applications require the knowledge of data errors to quantitatively understand the relevant uncertainties (Brohan et al. 2006; Folland et al. 2001) and to make meaningful statistical inferences for scientific conclusions. For example, an uncertainty assessment of the optimal global average annual mean surface air temperature (SAT) requires information on data errors (Jones et al. 1997; Shen et al. 1998, 2007; Folland et al. 2001). The uncertainties of SAT trend and the statistical inference of extreme SAT also need error data.

The uncertainties of a climate dataset have many aspects. Three fundamental types are 1) observational errors due to station data quality; 2) sampling errors due to data gridding, reconstruction, or spatial or temporal averaging; and 3) temporal interpolation errors when filling in the missing values of station data. This paper will focus on investigating the sampling errors of data gridding and spatial averaging as well as associated uncertainties when including observational errors for the U.S. Historical Climatology Network, version 2 (USHCN V2), dataset. This dataset was recently developed by the U.S. National Oceanic and Atmospheric Administration's (NOAA) National Climatic Data Center (NCDC) (Menne et al. 2009) and is an improved dataset from USHCN V1 (Easterling et al. 1996). The USHCN V2 contains the monthly means of daily maximum (Tmax), daily minimum (Tmin), and daily mean (Tmean) SAT as well as precipitation data from 1218 stations over the contiguous United States. These primarily compose a subset of stations from NOAA's Cooperative Observer Program (COOP) selected using various criteria, including completeness and length of record. USHCN V2's development went through three steps: 1) data quality evaluation and database construction, 2) time-of-observation-biases (TOB) adjustment, and 3) pairwise homogeneity adjustment (to correct for artificial discontinuities) and missing-value estimation from surrounding

Corresponding author address: Samuel S. P. Shen, Dept. of Mathematics and Statistics, San Diego State University, 5500 Campanile Dr., San Diego, CA 92182.
E-mail: shen@math.sdsu.edu

DOI: 10.1175/JCLI-D-11-00102.1

stations. These three steps respectively generated three USHCN V2 datasets named RAW, TOB, and F52 (Menne and Williams 2009). The TOB data are used in this study for our analysis of sampling error, trends, and hottest and coldest years. We average the SAT anomaly data of the USHCN stations within 2.5° × 3.5° latitude–longitude grid boxes and use the resulting values within each grid box to calculate the contiguous U.S. (CONUS) temperature for each month and each year from January 1895 to December 2008. Further, we quantify the uncertainties of the gridded USHCN V2 TOB SAT data by estimating the sampling errors on the basis of the theory of averaging correlated time series (Wigley et al. 1984; Shen et al. 2007). The presence of sampling errors within each grid box influences the calculation of the national average and creates uncertainty in any analysis of CONUS SAT trends.

The detailed objectives of our current work include 1) an estimate of the error variance of the gridded USHCN V2 SAT data, 2) a calculation of the monthly and annual means of the CONUS SAT and their uncertainties by using the gridded data and their errors, 3) an analysis of the U.S. SAT trends and their errors from 1895 to 2008, and 4) identification of the 10 hottest and coldest years between 1895 and 2008.

The rest of the paper is arranged as follows. Section 2 describes the USHCN V2 data and the method for calculating the sampling error variances. Section 3 presents the results of this paper. Section 4 contains the conclusions and discussion.

2. Data and method

a. Data

The USHCN V2 dataset (Menne et al. 2009) consists of long-term stations selected from the COOP network on the basis of many factors, including length of record, spatial coverage, and stability. A criterion for a station to be included in the USHCN is a minimum continuous record of 80 yr. However, in the process of developing the USHCN V2 dataset, the station-sparse areas of the United States required that 208 series be formed as composites. Each of these series is a joined record of two or more neighboring stations of consecutive records of less than 80 yr. The total of 1218 stations in USHCN V2 includes these 208 composite series.

Among the three USHCN V2 datasets (RAW, TOB, and F52), TOB data have corrected the bias in the RAW data that is due to different local observational times of a day. The F52 data, also called the fully adjusted data, contain corrections for TOB, a homogenization process of pairwise comparison, and fill-in- the-network (FILNET) estimates. The FILNET estimates fill in all of the missing data of a station, and hence F52 data of each station are complete from 1895. This data-filling process may have introduced some temporal and spatial smoothness into the USHCN dataset. Detailed estimation of the FILNET estimation error and a quality assessment of the FILNET procedures are still to be made. We chose to use the TOB-adjustment data for assessing the sampling error, trend, and extremes for the U.S. gridded data and U.S. spatially averaged data.

The history of the number of these USHCN V2 TOB stations from January 1895 to December 2008 is shown in Fig. 1. The number of stations increased from a minimum of 481 in January 1895 to a maximum of 1209 in March 1962. Never have all 1218 USHCN stations been operating at the same time. The recent drop in the number of stations as shown in Fig. 1 is primarily due to station closures. The spatial distribution of the 1218 stations is shown in Fig. 2. This figure indicates that the CONUS is well covered, with a higher station density in the more heavily populated eastern United States.

We aggregate the TOB station data onto 2.5° × 3.5° latitude–longitude grid boxes. Figure 3 shows the 2.5° × 3.5°grid with the northwestmost grid box's northwest vertex as 50°N, 126°W. Altogether, the 1218 USHCN V2 stations fall within 120 grid boxes. The number of stations in a grid box has a maximum of 32 (the centroid of this grid box is 41.25°N, 75.25°W) and a minimum of 1 for several grid boxes on the border of the CONUS.

We have excluded these seven 1-station grid boxes on the southern and northern borders in our error calculation. Four of these boxes are along the Atlantic Ocean and Gulf of Mexico and have their stations on islands. Each of these boxes has most of its area over water. Another one-station box is farther west and near the U.S.–Mexico border and has about one-third of its area over water. Inclusion of these one-station grid boxes yields more noise. If these one-station grid boxes are used in calculating a national average, then each grid box should be weighted by its land area. Thus, the exclusion of these boundary grid boxes does not significantly affect the applications of the gridded USHCN data and conclusions regarding the CONUS temperature change. Two other excluded one-station grid boxes are near the U.S.–Canada border. A strong spatial discontinuity of the SAT anomalies at the one-station grid boxes was identified over Minnesota when compared with the adjacent grid boxes. This discontinuity led us to exclude these two boxes near the Canadian border. Thus, only 113 grid boxes are left for our analysis.

To estimate the sampling error variance for each grid box that had monthly-mean SAT data, we used the correlation-factor method of Shen et al. (2007). The

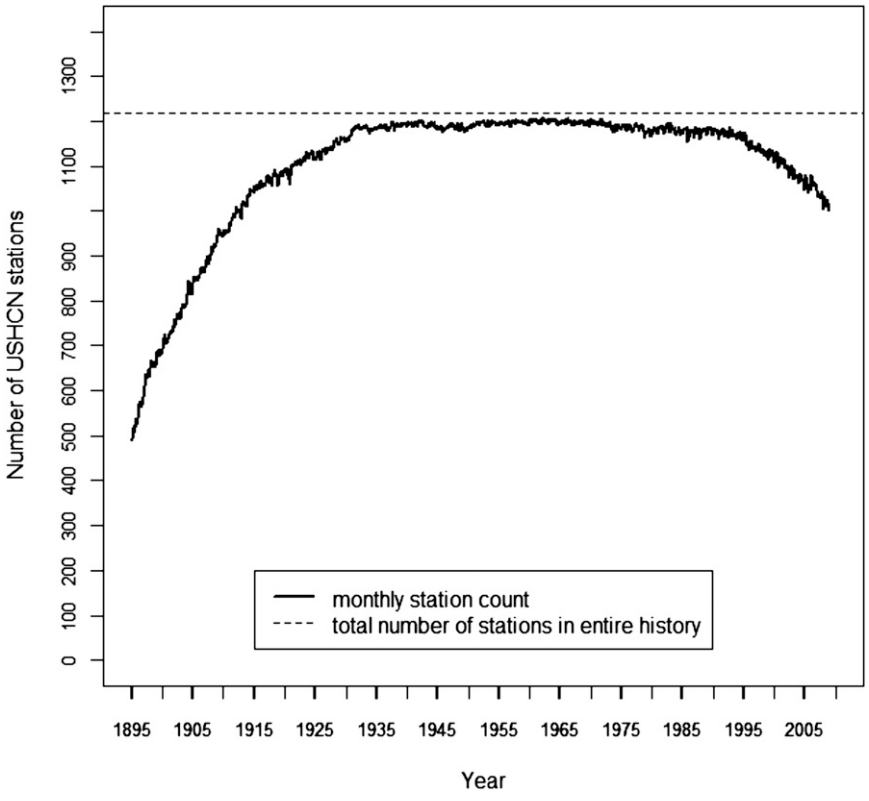


FIG. 1. History of the number of USHCN V2 stations from January 1895 to December 2008.

aggregated SAT data and their estimated errors are used to calculate the CONUS SAT time series and their uncertainties from 1895 to 2008 at the monthly, seasonal, and annual time scales. The trends of the mean SAT for 1895–2008 are assessed.

Our calculations are made with temperature data expressed as anomalies, that is, departures from the climatology as defined by the mean in the base period of 1961–90. Almost all of the USHCN stations have at least 21 yr of data during this base period. However, for each

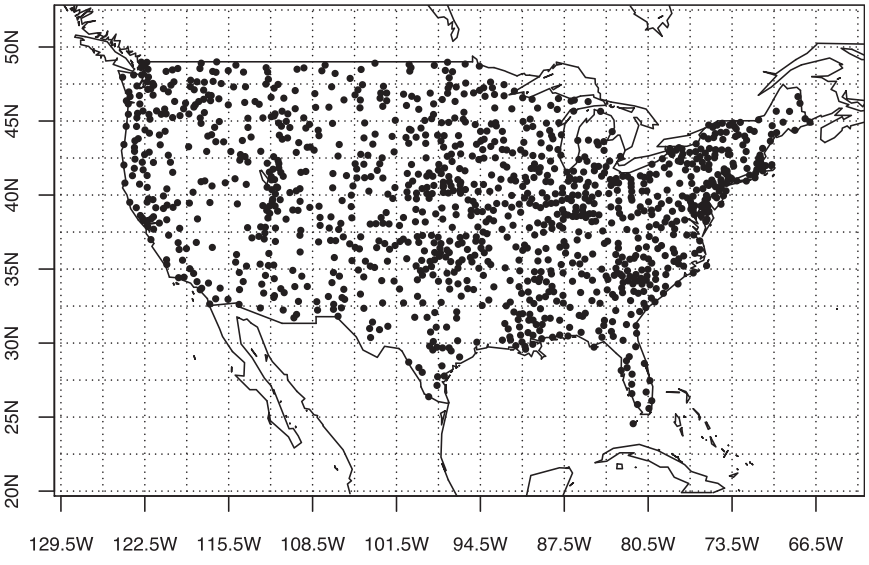


FIG. 2. Spatial distribution of the 1218 USHCN V2 stations.

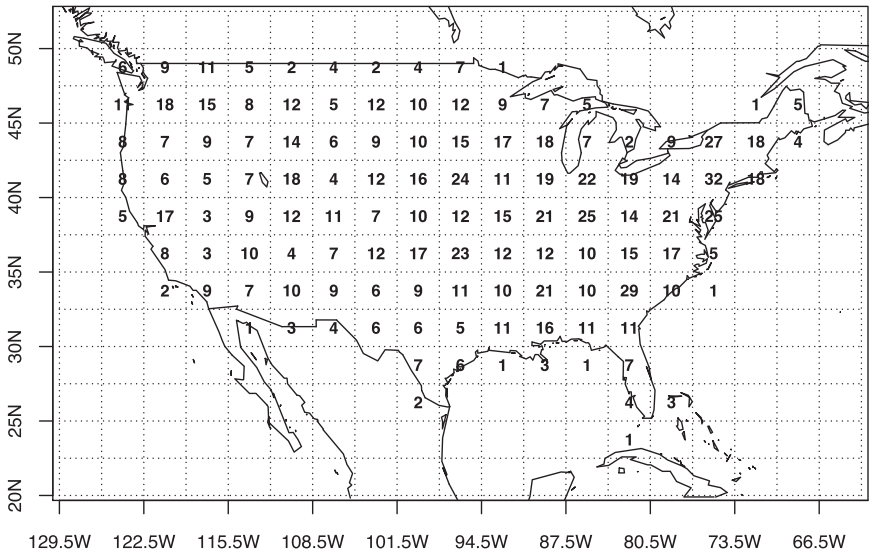


FIG. 3. Total number of stations N in a box in the USHCN history for the 2.5° × 3.5° gridbox resolution over the CONUS. The sum of the numbers in all of the grid boxes is 1218.

month there are still a few stations that did not have 21 yr of data in 1961–90. For example, for November, eight stations had fewer than 21 yr of data. Their station identifiers are 47195, 49490, 105559, 300023, 300321, 308906, 352135, and 355362. We cannot calculate the climatology and anomalies for these stations. Thus, the data from these stations are not used to derive the results of this paper.

b. Method

Consider an SAT anomaly field $T(\mathbf{r}, t)$, over a grid box Ω, where $\mathbf{r}$ is the position vector and t is time. Let $\overline{T}$ be the true average of the SAT field over the grid box:

$$\overline{T}(t) = \frac{1}{\|\Omega\|}\int_{\Omega} T(\mathbf{r}, t)\, d\Omega, \tag{1}$$

where $\|\Omega\|$ is the grid box's area. An estimator of this spatial average from station data is

$$\widehat{\overline{T}}(t) = \frac{1}{N}\sum_{i=1}^{N} T_i(t). \tag{2}$$

In the above equation, $T_i(t) = T(\mathbf{r}_i, t)$ is a sampling-anomaly datum of the station at $\mathbf{r}_i$ and time t and N is the number of stations in the grid box.

The SAT field over a grid box is not white noise and is inhomogeneous. The data of different stations in a grid box are correlated. Hence, the aggregation of the data for the stations inside a grid box is basically a problem of finding the average value of correlated time series (Wigley et al. 1984; Shen et al. 2007). The conventional convenient sampling error formula s^2/N does not apply. Here, we adopt the method of Shen et al. (2007) that uses spatial variances and a correlation factor to estimate the standard error of the gridbox data, that is, the mean-square error (MSE) of $\widehat{\overline{T}}$. The MSE is also referred to as error variance, and its estimation formula is

$$E^2 = \langle(\widehat{\overline{T}} - \overline{T})^2\rangle = \alpha_s\frac{\sigma_s^2}{N}, \tag{3}$$

where

$$\sigma_s^2 = \left\langle \frac{1}{N}\sum_{j=1}^{N}[T_j(t) - \overline{T}(t)]^2 \right\rangle \tag{4}$$

is the spatial variance,

$$\alpha_s = 1 + \frac{1}{N}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\left\langle \frac{(T_i - \overline{T})}{\sigma_s}\frac{(T_j - \overline{T})}{\sigma_s}\right\rangle \tag{5}$$

is the correlation factor, and the angle brackets stand for the operation of ensemble mean, or expected value.

The spatial variance σ_s^2 is estimated by

$$\hat{\sigma}_s^2(t) = \frac{1}{\|\mathrm{MTW}(t)\|}\sum_{\tau\in\mathrm{MTW}(t)}\frac{1}{N}\sum_{j=1}^{N}[T_j(\tau) - \widehat{\overline{T}}(\tau)]^2, \tag{6}$$

where $\mathrm{MTW}(t) = \{t - \tau_0, \ldots, t-2, t-1, t\}$ denotes the set of a moving time window (MTW) and $\|\mathrm{MTW}(t)\|$ is the cardinality of the set. We follow the idea of piecewise

stationarity (Folland et al. 2001) and use a 5-yr backward MTW. If a grid box has data for every year in the MTW, then $\tau_0 = 4$ and MTW $= \{t-4, t-3, t-2, t-1, t\}$. If an MTW does not have data for every year, then we require a minimum of 3 yr of data in an MTW to make a calculation that is based on Eq. (6). For 1897, the MTW has only 3 yr of data: MTW $= \{1895, 1896, 1897\}$ and hence $||\text{MTW}(1897)|| = 3$. For spatial variance, $N = 4$ is chosen as the minimum number of stations within a box, because the following regression estimate of α_s needs at least four stations.

The correlation factor α_s is estimated by using a regression rather than being computed directly from Eq. (5) (Shen et al. 2007). Suppose that a box has N (larger than or equal to 4) station anomalies. We treat the data of these N stations as a statistical population. The population mean of the station temperature anomalies in the box is

$$\widehat{\overline{T}}_N(t) = \frac{1}{N}\sum_{i=1}^{N} T_i(t). \tag{7}$$

If a simple random sampling of n stations is taken from the population (Cochran 1977), then the sample mean of the n stations is

$$\widehat{\overline{T}}_n(t) = \frac{1}{n}\sum_{i=1}^{n} T_{n,i}(t), \tag{8}$$

where $T_{n,i}$ is the ith station's anomaly temperature in the subsample network of size n. Following the method of Shen et al. (2007), the mean-square difference $\hat{E}_n^2$ between the above two quantities over 1000 samples is used as an initial estimate of the sampling error. We then apply a regression procedure using the following data:

$$\left(\frac{\hat{E}_n^2}{\hat{\sigma}_s^2}, \frac{1}{n}\right)(n = 1, 2, 3, \ldots, N-1). \tag{9}$$

The regression coefficient is the $\hat{\alpha}_S$ value, which is the estimate of α_s in Eqs. (3) and (5).

We then populate the $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values onto the grid boxes with fewer than four stations by using a simple interpolation method. Starting from the farthest southeastern grid box, if a grid box G has three or fewer stations, it will be assigned $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values according to the following search procedure. On the same latitude band, a search is done for the first box to the west. If this box has four or more stations, the $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values of this box are assigned to grid box G. Otherwise, we search to the first box on the same latitude band to the east. If the box does not meet the criterion, then the search goes to the second box to the west. This west and east alternation can acquire the $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values for grid box G if the latitude band has at least one grid box that has four or more stations. If it does not, we search for a grid box on the first latitude band to the north. The first search is to the box directly north of grid box G, then to the first box to the west, and then to the right. The alternation goes on until $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values are found or the entire latitude band has been exhausted. If the entire band is exhausted without $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values, we do the same search for the first latitude band south. This north and south alternation will eventually find at least one pair of $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values, which are assigned to grid box G. This search procedure has a preference to a western box and a northern box. The temperature's spatial variance over the CONUS is more relevant to the west than to the east because of the prevailing westerly atmospheric circulation patterns (Washington and Parkinson 1986, chapter 2). However, it is unknown whether our search procedure is optimum to this atmospheric circulation. Alternative search procedures can also be used to fill in the values for grid box G. For example, Shen et al. (2007) considered the errors of the gridded data for the entire globe, searched from a northwestmost grid box, and had a preference to the south latitude band.

Last, the sampling error variance of the USHCN gridbox data for a given box and a given month is computed by

$$E^2 = \hat{\alpha}_S \frac{\hat{\sigma}_s^2}{N}. \tag{10}$$

3. Results

a. Error variances of the gridded USHCN

Error variances $\hat{E}_n^2$ are calculated for all 113 grid boxes (Fig. 3) for each month from January 1897 to December 2008 for which a box had data. Figure 4 shows the error variances of 4 months: January of 1897 and 1990 and July of 1897 and 1990. The year 1897 is the earliest year for which the error variance can be computed because the MTW requires at least 3 yr of data within the 5-yr MTW. The sampling error variances are inversely proportional to the number of stations. Thus, the error variances for the eastern United States are smaller than those of the western United States and those of the boundary grid boxes. The large sample error variances occurred over the grid boxes with few stations and large spatial variances. These grid boxes are distributed along the mountain regions, the northern and southern national borders, and some of the coastal grid boxes. Although these error variances are relatively larger than those over the eastern United States, they

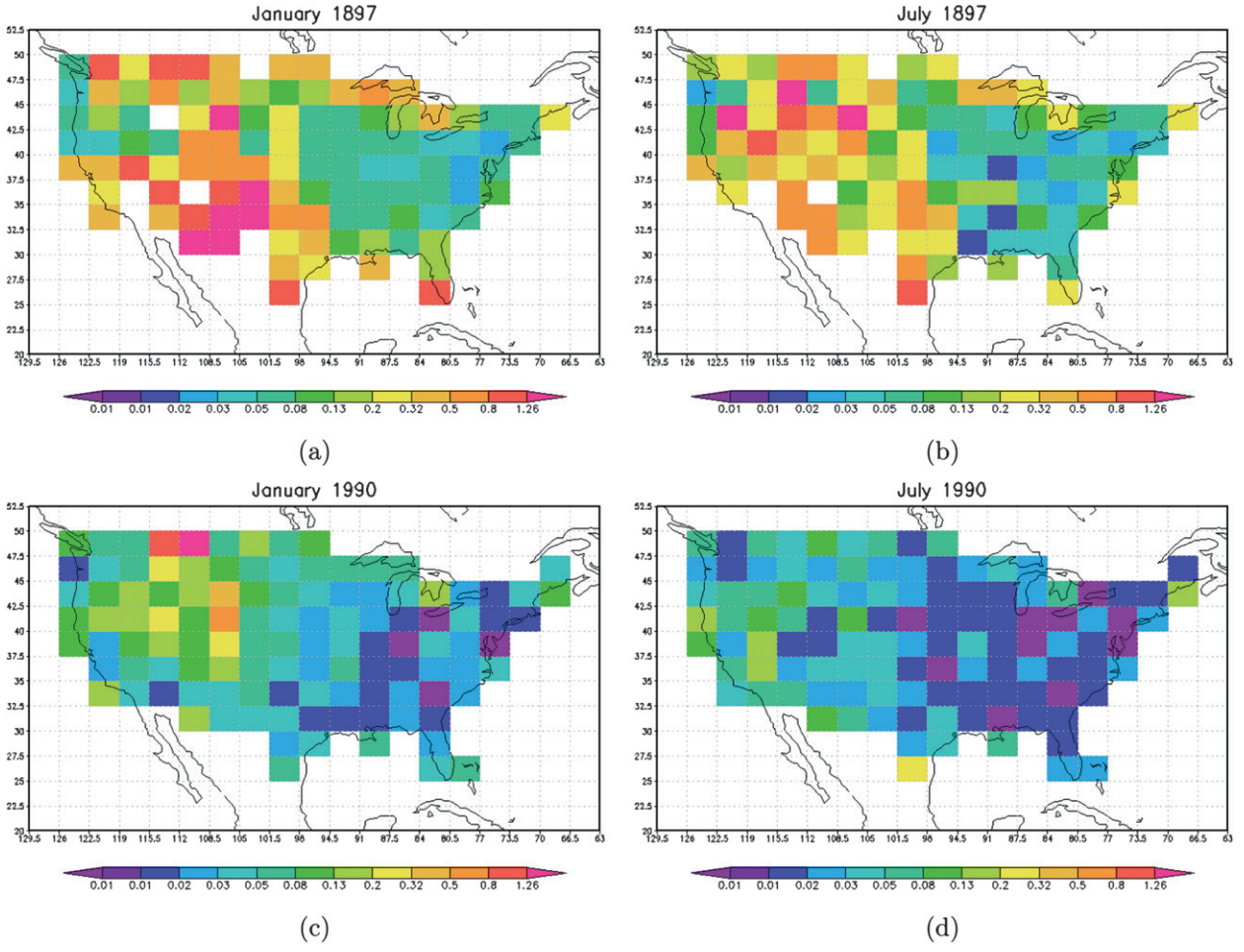


FIG. 4. Spatial distribution of the sample error variances (°C^2) of the USHCN TOB-adjusted Tmean anomaly data (relative to the 1961–90 climatology) over the grid boxes: (a) January 1897, (b) July 1897, (c) January 1990, and (d) July 1990.

are in general still less than 1.3°C^2. In 1897, the eastern U.S. error variances were still smaller than 0.3°C^2 when only about 500 stations existed for the CONUS as a whole. However, the sampling errors were large in 1897 in the western mountain regions and the southwestern United States, and many grid boxes had sampling error variances greater than 1.0°C^2. Several grid boxes had no station data in 1897, and they are left blank in Figs. 4a and 4b. Because of the nontrivial sizes of the sampling errors in the early part of the temperature record, when identifying the SAT trend over certain regions, the associated sampling uncertainties need to be considered.

The station data have errors also. The errors include the time-of-observation bias caused by the change of observational time during a day, bias caused by the changes that result from station moves and changes in observing practice, heat island effects, instrumental random measurement errors, and others (Brohan et al. 2006; Folland et al. 2001; Karl et al. 1986; Menne et al. 2009; Vose et al. 2003). Some known biases, such as the TOB, have been corrected in USHCN V2 (Menne et al. 2009). The random errors, some of which were introduced during the adjustment process, still remain. These random errors have expected values that are close to 0 and are uncorrelated among the stations (Brohan et al. 2006). The quantitative assessment of the remaining observational errors, including both random and bias errors, is a challenging task and remains to be addressed, as pointed out by Brohan et al. (2006) and Menne et al. (2009). Menne et al. (2009) used the pairwise comparison and further homogenized the USHCN data. However, the procedure resulted in a larger warming trend of the U.S. SAT. We have chosen to use the TOB-only data and used grid boxes larger than those of Menne et al. (2009). Our grid-box size ensures that at least one station exists in every grid box across the country during the climatological period 1961–90. Our warming trend is 0.057°C (10 yr)$^{-1}$, whereas

Menne et al. (2009)'s trend from F52 data is 0.064°C $(10\ \text{yr})^{-1}$. When we apply our method to F52, the trend is much larger: 0.075°C $(10\ \text{yr})^{-1}$. In future research we will address detailed assessments of the results from RAW, TOB and F52 data. Because the observational error and the sampling error constitute the overall uncertainties of the observed climate changes, the uncertainty of the gridded data should incorporate both errors. Consequently, the overall uncertainty is larger than the sampling error shown in Fig. 4.

The observational random error variance of the gridded data is denoted by E_o^2. The gridded SAT field $\overline{T}$ may be statistically modeled by

$$\overline{T} = \widehat{\overline{T}} + \epsilon_s + \epsilon_o, \tag{11}$$

where $\overline{T}$ is the true gridded SAT field, $\widehat{\overline{T}}$ is the gridded datum calculated by Eq. (7), ϵ_s is the sampling error, and ϵ_o is the observational error. For our data, we assume that both errors are normally distributed and have a mean of 0 (Brohan et al. 2006):

$$\epsilon_s \sim N(0, E^2) \quad \text{and} \quad \epsilon_o \sim N(0, E_o^2). \tag{12}$$

This assumption for the sampling error has little question, but that for observational error may be questionable since bias may still exist after TOB and pairwise corrections are applied (Williams et al. 2012). Thus, one may regard this assumption for the observational error as a mathematical approximation, which can still be subject to change.

By definition, the sampling error and the observational errors may be assumed to be uncorrelated. Thus, when both sampling and observational errors are taken into account, the total error variance for a gridbox datum is

$$\epsilon^2 = E^2 + E_o^2. \tag{13}$$

Figure 4 shows that the sampling error variance of the gridded data is bounded by $1.3°C^2$. The standard sampling errors for most grid boxes are thus less than 1.1°C. For many boxes, the errors are close to zero. Figure 1 indicates that the number of stations reached a maximum from the 1930s through the 1980s. Thus, the sampling error variances were large in 1895 and gradually diminished to almost zero in the 1930s for most grid boxes. Figure 8 of Menne et al. (2009) shows the error bars for the annual minimum SAT (Tmin) at Reno, Nevada. This station may be the worst-case scenario for the error-bar size. Those Monte Carlo–simulated error bars $\pm 2E_o$ may be regarded as an upper bound of the observation uncertainty, although the data have gone through the full pairwise adjustment. In the earlier years, $2E_o$ was about 1.0°C. The error diminished to zero in the 1990s. Brohan et al. (2006) and Folland et al. (2001) postulated their observational error sizes. Here, we do the same by comparing Fig. 4 of this paper with Fig. 8 of Menne et al. (2009) and tentatively postulate that the observational errors are about one-half of the sampling errors for each grid box in the contiguous United States. This assumption and Eq. (13) lead to $\varepsilon^2 = (5/4)E^2$.

b. U.S. average SAT and its uncertainty

The CONUS spatial average of the mean SAT is calculated at monthly, seasonal, and annual time scales by an area-weighted averaging method applied to the 113 grid boxes. For the U.S. spatial average to truly reflect the entire CONUS area, all 113 grid boxes are used to calculate the U.S. average. The data-void grid boxes acquire their data from interpolation in the same way as the correlation factor.

To find the standard error of the U.S. average SAT, we use Eqs. (11) and (12) of Jones et al. (1997) on the basis of the definition of degrees of freedom using the *S* method (Wang and Shen 1999):

$$\overline{E}^2 = \frac{\sum_{i=1}^{113} E_i^2 \cos\varphi_i \Big/ \sum_{i=1}^{113} \cos\varphi_i}{N_{\text{eff}}}, \tag{14}$$

where E_i^2 is the sampling error variance determined by Eq. (10) for the ith grid box, φ_i is the latitude of the grid box's centroid, and N_{eff} is the effective degrees of freedom of the SAT anomaly field over the United States. In effect, the area-weighted average is calculated for the sampling error variances over grid boxes. The average is then divided by the effective degrees of freedom [see Eq. (11) of Jones et al. (1997) and Eq. (7) of Smith et al. (1994)], and the square root of this result is the standard error of the U.S. average. Here, we have assumed that $\widehat{\overline{T}}$ is an unbiased estimate of $\overline{T}$ as indicated by Eqs. (11) and (12). Thus, the expected value of $\overline{T}$ is equal to $\widehat{\overline{T}}$, and hence the *S*-method definition of the degrees of freedom applies (Wang and Shen 1999). The degrees of freedom of the monthly SAT over the Northern Hemisphere were estimated in Wang and Shen (1999) by using four different methods and varied between 40 and 80. From Eqs. (13) and (14) of Jones et al. (1997), the effective degrees of freedom is the ratio of the investigation area to the spatial characteristic area. The area of the Northern Hemisphere divided by 40 is the upper bound of the characteristic area, which is $5.6 \times 10^6\ \text{km}^2$. The CONUS area divided by this characteristic area is equal to 1.4. This is the lower bound of N_{eff}. The upper bound

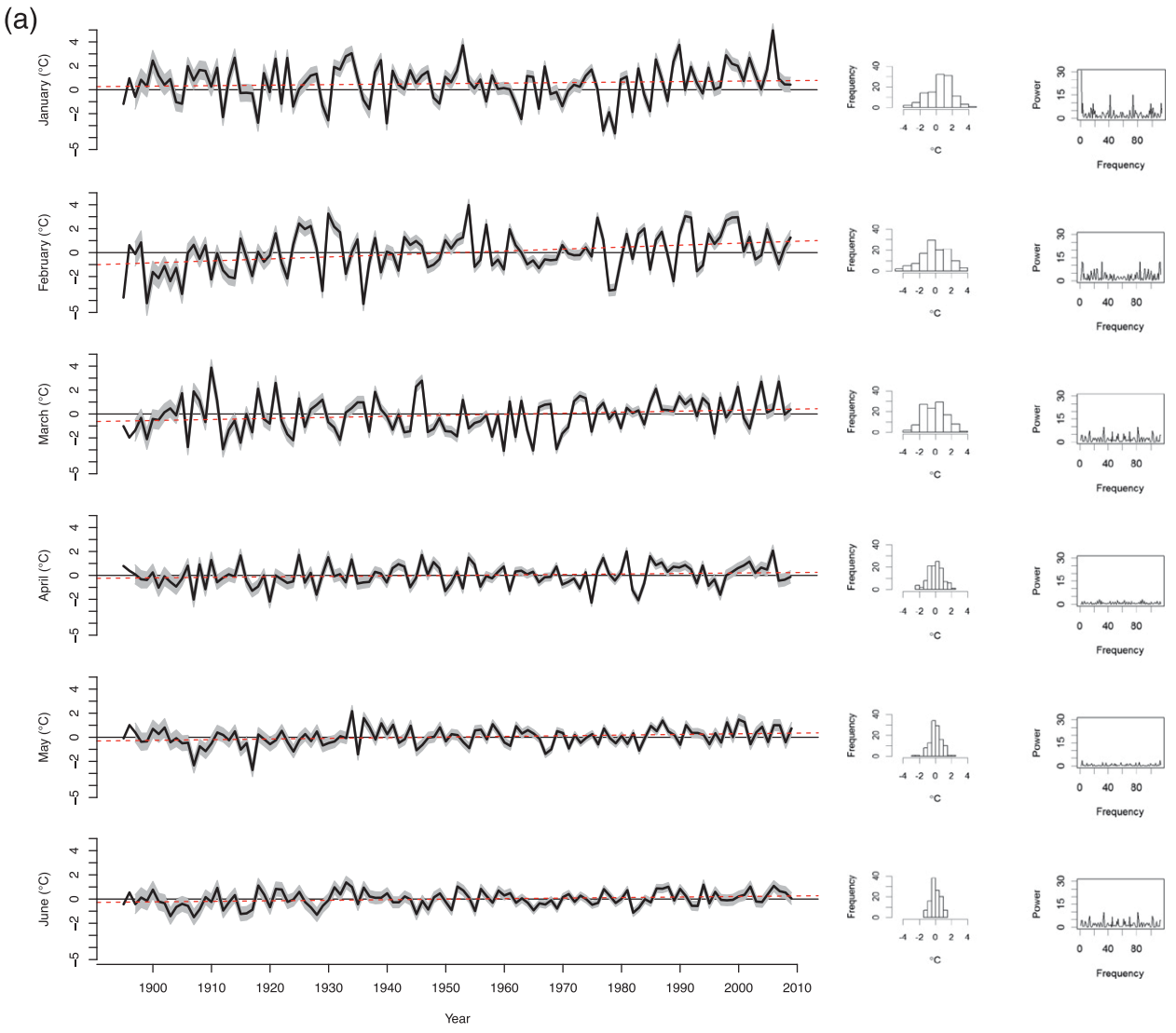


FIG. 5. (left) Time series with 2-sigma error margin, (center) histograms, and (right) spectral power of the U.S. average monthly Tmean anomalies (relative to the 1961–90 climatology) for (a) January–June and (b) July–December. The red dashed line is the linear trend. The shaded area is the 2-sigma error margin.

is 2 times this value. We thus estimate the effective degrees of freedom N_{eff} for the CONUS to be 2 for the monthly SAT. The solid lines in the first column of Fig. 5 show the area-weighted average of the contiguous U.S. SAT anomalies on the basis of the gridded USHCN data for each month from January to June (Fig. 5a) and from July to December (Fig. 5b). The shading indicates the 2-sigma error bound (i.e., the 95% confidence interval). The red dashed straight lines are the linear trends whose slopes and their error ranges are given in Table 1.

Figure 5 considers both sampling error $\overline{E}^2$ and observational error. When taking the random observational error into account, the actual confidence interval at the 95% confidence level is slightly larger than $\pm 2\overline{E}$. The observational errors are approximately one-half of the sampling error for the same grid box as we have postulated; that is, $E^2 = (1/4)E_o^2$ and then $\epsilon = (E^2 + E_o^2)^{1/2} = 1.12E$. From this, the actual confidence interval is $\pm 2.24E$, shown by the shaded region for the time series in Fig. 5. This may be the worst-error scenario. Despite these two errors and other uncertainties, the total errors are not yet large enough to alter the conclusions about the upward or downward trends of SAT in a given period of time on the basis of the USHCN V2 TOB-adjusted data or fully adjusted F52 data (Fig. 10 of Menne et al. 2009).

Figure 5 clearly demonstrates larger temperature trend, variance, and error in the winter than in the summer. The

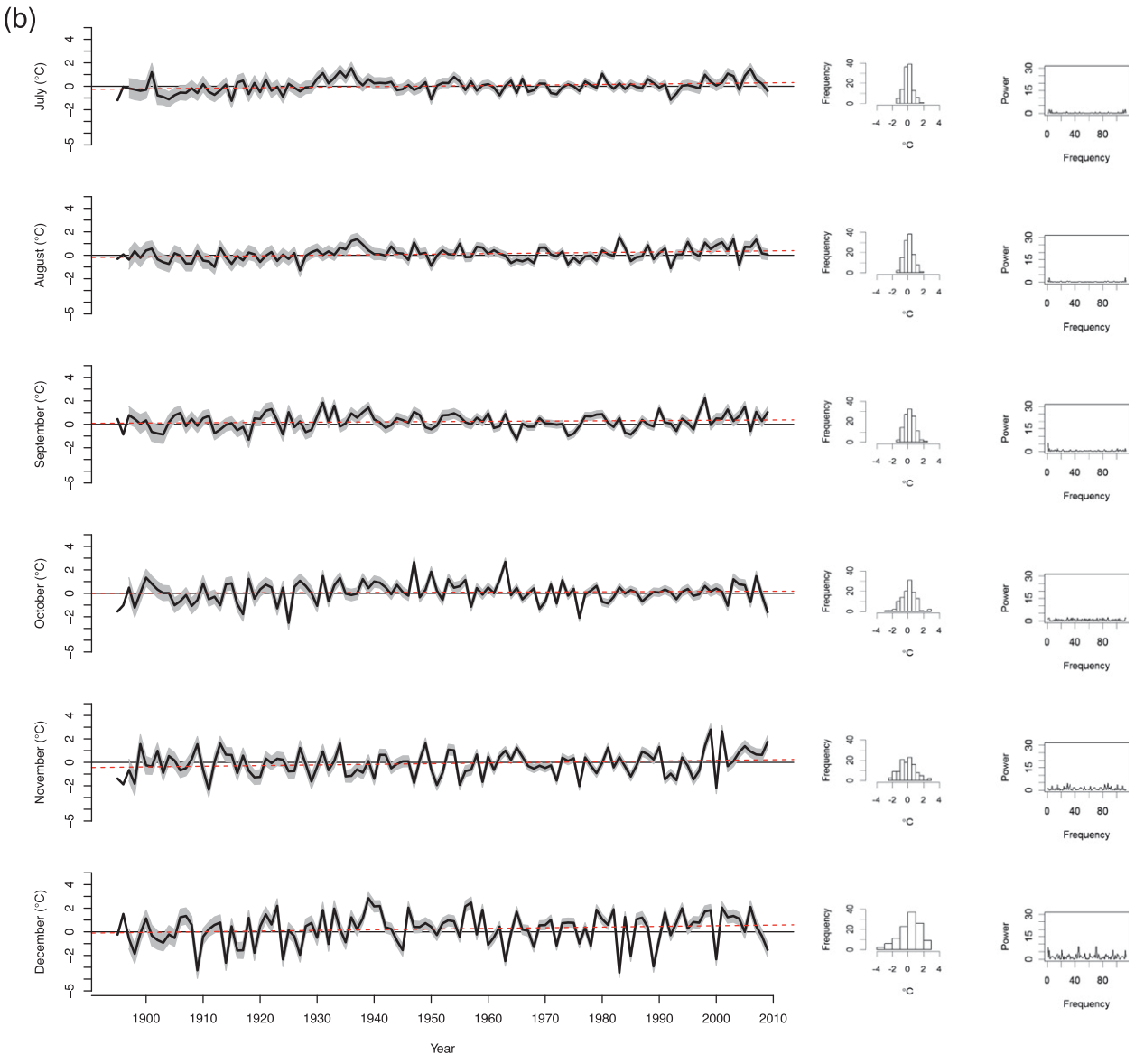


FIG. 5. (*Continued*)

histograms show that the February anomalies have the largest variance. February SAT has a clear warming trend from 1900 to the 1930s (the ''Dust Bowl'' drought period). Both January and February SAT show warming trends in the 1980s and 1990s. The warming trend is also shown in December's SAT in the 1990s. The power spectra of January SAT anomalies indicate the existence of a strong low-frequency oscillation, and a similar but weaker oscillation appears in February's and December's power spectra. The first peak at zero frequency is due to the large positive mean, which is reflected in both the time series and the histogram. This peak is statistically significant at the 5% significance level (Wei 2006, section 13.1.3). The other two peaks at periods of approximately 6 yr (114/20) and 3 yr (114/40) may be the U.S. SAT response to El Niño. The other two identical high peaks at the higher frequencies are the mirror images reflected at the frequency of 56/114 and seem to have no physical meaning. The peaks of February, March, and December also reflect the U.S. SAT variations at the periods in the range of 2–7 yr. However, none of these peaks is statistically significant even at the 10% significance level according to an F test (Wei 2006, section 13.1.3). Therefore, it is likely that these low-frequency oscillations shown in the power spectra of December, January, February, and March are noisy but statistically

TABLE 1. Linear trends of the monthly, seasonal, and annual mean CONUS Tmean anomalies (relative to the 1961–90 climatology) from 1895 to 2008 [°C (10 yr)$^{-1}$] using USHCN V2 TOB-adjusted data. The boldface type indicates that a trend is statistically significant at the 5% significance level. The plus/minus sign indicates a 1-sigma confidence interval at the 68% confidence level.

Month	Trend
Jan	0.044 ± 0.047
Feb	0.162 ± 0.048
Mar	0.087 ± 0.042
Apr	0.042 ± 0.028
May	0.053 ± 0.024
Jun	0.045 ± 0.022
Jul	0.050 ± 0.020
Aug	0.048 ± 0.020
Sep	0.020 ± 0.022
Oct	0.023 ± 0.027
Nov	0.049 ± 0.032
Dec	0.066 ± 0.040
Annual	0.057 ± 0.017
Winter (Dec–Feb)	0.086 ± 0.030
Spring (Mar–May)	0.061 ± 0.022
Summer (Jun–Aug)	0.048 ± 0.017
Autumn (Sep–Nov)	0.031 ± 0.020

insignificant SAT responses to El Niño. In other months, from April to November, the SATs have little warming trends, small variances, and very weak power spectra.

The histograms of Fig. 5 clearly indicate that the summer SAT anomalies have much smaller variances than do those of winter. For the purpose of convenient comparison, we have fixed the scales of both the horizontal and vertical axes of the histogram. Thus, for the summer months of small variance, the histogram bars are thinner than those for the winter months of large variance. The histograms also seem to indicate that the summer months' SATs are more symmetrically distributed than the winter ones, perhaps because of smaller summer climate variations. Although it may not be conclusive whether the U.S. mean SAT anomalies are skewed left, the histograms seem to suggest a longer tail toward the left. This is most likely due to the fact that during the 1895–2008 period most years before 1960 were colder than the 1961–90 mean. Other inferences-based nonparametric methods, such as the Kolmogorov–Smirnov test, can be made to confirm whether the probability density functions (pdf) have shifted from one period to another (Regele 2010). One can also assess temporal variations of the variance, skewness, and other higher statistical moments (Shen et al. 2011). A comprehensive study of the inference that is based on the three USHCN datasets will be included in future work.

Figure 6 shows the annual average of the monthly SAT for the CONUS TOB-adjusted time series in Fig. 5. The red bars indicate the positive anomalies of the U.S. annual mean SAT with respect to the 1961–90 average, and the blue bars indicate negative anomalies. The thin and black error bars are the 2-σ confidence interval at the 95% confidence level. Here, we have accounted for the errors from both sampling and measurement and have used an upper-bound estimate that the measurement error is one-half of the sampling error. We also assume that the error variances from different months are independent. The standard error $\overline{E}_{\rm Ann}$ of the annual mean is thus estimated by the 12-month mean of the monthly error variance, divided by 12. Namely,

$$\overline{\varepsilon}_{\rm Ann} = \left(\frac{1}{12}\sum_{m=1}^{12}\overline{\varepsilon}_m^2/12\right)^{1/2} = \left[\frac{1}{12}\sum_{m=1}^{12}(\overline{E}_m^2 + \overline{E}_{o,m}^2)/12\right]^{1/2} = \left(\frac{5}{48}\overline{E}_{\rm Ann}^2\right)^{1/2}, \quad (15)$$

where $\overline{E}_m^2$ and $\overline{E}_{o,m}^2$ are respectively the U.S. average SAT's sampling error and the random observational error for the month m and

$$\overline{E}_{\rm Ann}^2 = \sum_{m=1}^{12}\overline{E}_m^2/12$$

is the annual mean of the monthly U.S. spatially averaged sampling errors. Thus, the 95% confidence error bars (i.e., 2σ) of our annual time series are calculated by $\pm 2\overline{\varepsilon}_{\rm Ann}$. The thick black solid line is the 10-yr moving average of the annual SAT. The moving average shown in Fig. 6 is from 1899 to 2003. Moving average is only one of many types of low-pass filters that may be used to smooth the annual time series. The results from different filters may have little difference in the middle section of a time series but can show different tendencies at the end of a time series (Mann 2004). A more systematic way of finding a nonlinear and nonstationary trend may be the empirical model decomposition method (Huang and Shen 2005). Still, the tendencies near the end points should not be used as time series extrapolations.

Figure 6 shows the warm anomalies in the 1930s and the last two decades and the cold anomalies in the first two decades and the 1960s and 1970s. Although the U.S. SAT's long-term warming trend is synchronized with that of the global SAT, the United States's persistent warm anomalies in the 1930s and the short-lived warm anomalies in the early 1950s are different from the global SAT (Fig. 3.6 of Solomon et al. 2007; Karl et al. 2009). For the globe, the 1930s SAT was cooler than the 1961–90 climatology. The recent distinct warmth over the United States in 1998 and 2006 and other warmth observed in the

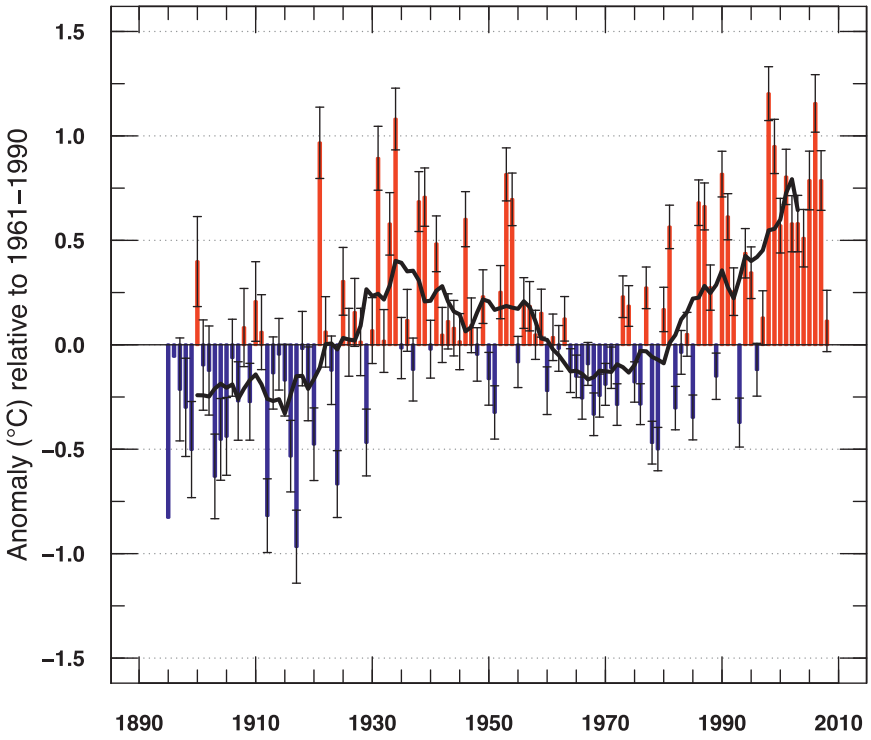


FIG. 6. The annual mean SAT Tmean for the contiguous United States from the TOB-adjusted data. The error bar represents the 95% confidence interval. The blue curve is the 10-yr moving average.

last two decades were well synchronized with the global SAT.

c. U.S. SAT trend and its uncertainty

The linear trend from 1895 to 2008 for each month is shown in Table 1. The trend is positive for every month. The largest trend is 0.162°C (10 yr)$^{-1}$ for February, and the smallest trend is 0.020°C (10 yr)$^{-1}$ for September. The error of the linear trend assessment is calculated by using the following linear statistical model:

$$T_d = \beta_0 + \beta_1 t + \varepsilon + \varepsilon_E. \tag{16}$$

Here, T_d is the monthly U.S. SAT data, $\beta_0 + \beta_1 t + \varepsilon$ is the linear statistical model to represent the true monthly U.S. SAT, ε is the model error, and ε_E is the data error we have just estimated, called the measurement error in statistical literature (Carroll et al. 2006), relative to the model and is assumed to be independent of ε. The variance of ε_E is the sum of the sampling error variance $\overline{E}_m^2$ and the observational error variance $\overline{E}_{o,m}^2$ according to Eq. (15). Equation (16) implies that

$$\mathrm{Var}(\hat{T}_d) = \mathrm{Var}(\varepsilon) + \mathrm{Var}(\varepsilon_E), \tag{17}$$

where $\hat{T}_d = \hat{\beta}_0 + \hat{\beta}_1 t$ is the estimated trend and $\hat{\beta}_1$ for each month is shown in Table 1. For example, $\hat{\beta}_1 =$ 0.044°C (10 yr)$^{-1}$ for January and $\hat{\beta}_1 = 0.162$°C (10 yr)$^{-1}$ for February. The uncertainty for this slope is measured by the standard deviation of $\hat{\beta}_1$, which is

$$\mathrm{SD}(\hat{\beta}_1) = [\mathrm{Var}(\hat{T}_d)/S_{xx}]^{1/2}, \tag{18}$$

where S_{xx} is the variance of the explanatory variable t and Var(ε) is estimated by sum of squared errors (SSE): $S_{xx} = \mathrm{SSE}/(n-2)$, with $n = 113$ being the total number of data points (Carroll et al. 2006; Wackerly et al. 2002). Standard statistical software packages can calculate SD($\hat{\beta}_1$) but do not include the data error. For the mth month, $\mathrm{Var}(\varepsilon_E) = \overline{E}_m^2 + \overline{E}_{o,m}^2 = (5/4)\overline{E}_m^2$ is the sum of the sampling error variance and the random observational error variance, which is assumed to be one-quarter of the former; Var(ε_E) varies from year to year. When this quantity is added to Var(ε) [see Eq. (17)], a more realistic uncertainty for the slope measured by $\pm$SD($\hat{\beta}_1$) [see Eq. (18)] is calculated (shown in Table 1). This regression error in winter months is much larger than that in summer months, because the winter SAT has much larger variances than does summer SAT.

Despite some unknown data errors, according to the error formulations in Eqs. (17) and (18), the positive trend signals are significant at a 5% significance level for February, March, May, June, July, August, winter, spring, summer, and the full year (see Table 1 for the boldface values). Although the trend of December is 0.066°C (10 yr)$^{-1}$ and is the second largest among the 12 months, it is obscured by the large noise of the month and is not statistically significant. The same can be said for January, April, and November.

There are still some unknown errors in the climate data, however. The remaining errors after the TOB or full adjustments of Menne et al. (2009) may still include nonrandom errors associated with observational practices that are not fully accounted for, such as changes in station location, changes in the station environment, and changes in instrumentation through time. Thus, the remaining error in the USHCN V2 data may not be spatial white noise. Before the errors of the sampling and observations are fully determined and before the spatial numerical integration errors are accurately estimated, the significance of the weaker trends cannot be assessed with very high certainty. In the future, more effort needs to be given to resolving these issues.

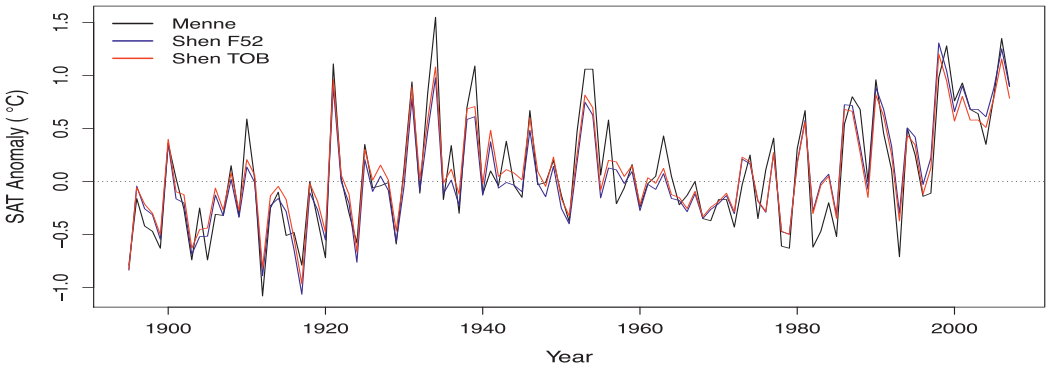
d. Comparison of three time series of the annual mean U.S. average SAT

For the annual mean, Menne et al. (2009) calculated the U.S. average SAT anomaly by using the area-weighted average of the gridded data at a 0.25° × 0.25° resolution. They interpolated the TOB- and pairwise-adjusted station data—that is, the F52 data—onto the 0.25° × 0.25° grid first by using an improved inverse-distance weighting method (Willmott et al. 1985). We consider the difference between the Menne et al. (2009) Tmean results (i.e., their Fig. 12) and our TOB Tmean results (Figs. 6, 7a) that are based on the area-weighted average of the 2.5° × 3.5° gridbox data. To examine the pairwise adjustment and FILNET effect, we have also included our results that are based on the F52 data over the 2.5° × 3.5° grid in the comparison in Fig. 7. The three time series in Fig. 7(a) follow a similar upward trend, with trends of 0.059, 0.075, and 0.064°C (10 yr)$^{-1}$ for Shen TOB, Shen F52, and Menne. The increase of the trend from TOB to F52 is expected because the full adjustment procedure included corrections for siting and instrument changes, such as the transition to the maximum-minimum temperature system that took place in the 1980s (Menne et al. 2009). Menne et al. (2009) used a finer-resolution grid that has smoothed both the spatial and temporal variances; hence, Menne's trend is less than that of Shen F52 as discussed below in the description of Fig. 7d. The correlation coefficients between each pair of the time series in Fig. 7a are 0.94 between Shen TOB and Menne, 0.93 between Shen F52 and Menne, and 0.99 between Shen TOB and Shen F52. Thus, Shen TOB and Shen F52 are synchronized almost perfectly in phase, achieving extremes at the same time but with different magnitudes, as demonstrated by the red and blue lines of Fig. 7a. The phase synchronization between Menne and Shen TOB or between Menne and Shen F52 is not as good, particularly in the early 1940s and late 1950s. This implies that interpolation method and grid resolution may play a very important role in determining the warming trend and identifying temperature extremes.

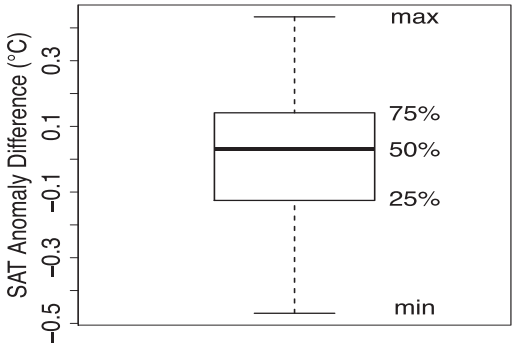
Figures 7b–e show the statistics of differences between Shen TOB and Menne, as well as Shen F52 and Menne, with box plots and power spectra plots. Figure 7b shows the statistical properties of the difference resulting from the Shen TOB result minus Menne et al. (2009)'s time series. The mean and standard deviation of the differences of the Shen TOB time series minus Menne's are 0.01° and 0.19°C, respectively. The positive mean 0.01°C reflects Shen TOB's slightly higher estimate of the U.S. SAT. The nontrivial standard deviation of the differences 0.19°C might be caused by using grids of different resolutions, as well as the pairwise adjustment. The maximum absolute difference is 0.47°C, which occurred for 1934 for which our current estimate of the U.S. SAT anomaly is 1.08°C while Menne's is 1.55°C. The second-largest absolute difference is 0.43°C, which occurred for 1983 with ours equal to −0.04°C and Menne's equal to −0.47°C. All together there are 16 yr with absolute differences that are larger than 0.3°C.

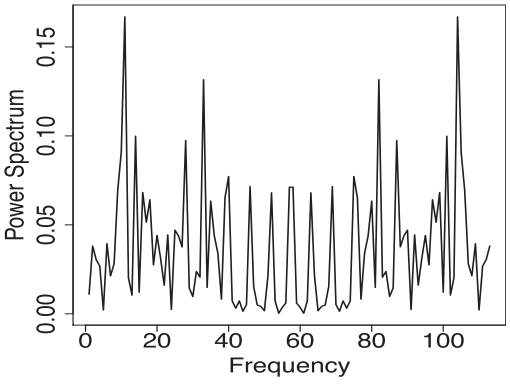
Figure 7c shows the power spectra of the differences between Shen TOB and Menne, which are similar to those for random noise and show no particular dominant cycles, an indication that no systematic cyclic bias has been introduced to the system of Menne's or our current work.

Figure 7d shows the box plot of the differences of Shen F52 and Menne, both of which are based on the fully adjusted F52 data. The mean is −0.02°C, and the standard deviation is 0.20°C. Thus, Shen F52 yields a slightly lower U.S. SAT because of the coarser grid and the different spatial interpolation and averaging methods relative to Menne. The largest positive difference is 0.46°C in 1983, and the largest negative difference is −0.57°C, which occurred in 1934. These two extreme values contribute to the larger trend of Shen F52 relative to Menne. The 0.20°C standard deviation is similar to the standard deviation 0.19°C of the difference between Shen TOB and Menne, which implies a similar magnitude of variations of the two differences. In contrast, the standard deviation of the difference between Shen TOB and Shen F52 is only 0.07°C.

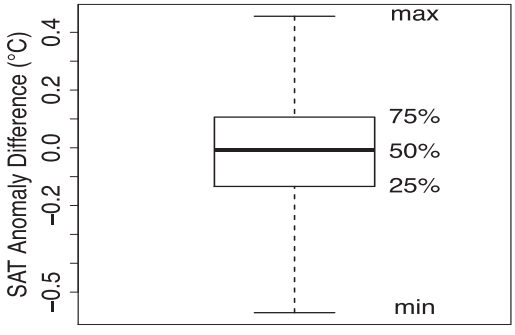


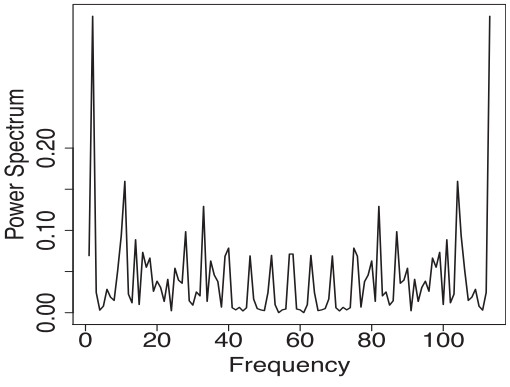
(a)





(b) (c)





(d) (e)

FIG. 7. Comparison between the CONUS average annual Tmean anomaly (relative to the 1961–90 climatology) of this paper and that of Menne et al. (2009): (a) time series of Menne (black) and this paper: Shen F52 (blue) and Shen TOB (red); (b) box plot of the differences of Shen TOB and Menne: the mean is 0.01°C, the standard deviation is 0.19°C, the maximum is 0.43°C, and the minimum is −0.47°C; (c) power spectra of the difference in (b); (d) box plot of the differences of Shen F52 and Menne: the mean is −0.02°C, the standard deviation is 0.20°C, the maximum is 0.46°C, and the minimum is −0.57°C; and (e) power spectra of the difference in (d).

TABLE 2. The 10 hottest and coolest summers between 1895 and 2008 from Tmax, and the 10 warmest and coldest winters from Tmin from the TOB-adjusted data. The numbers in the parentheses are the corresponding anomalies relative to the 1961–90 climatology (°C).

10 hottest summers (JJA avg Tmax)	10 coolest summers (JJA avg Tmax)	10 warmest winters (DJF avg Tmin)	10 coldest winters (DJF avg Tmin)
1934 (2.12)	1939 (−2.14)	1990/91 (1.66)	1939/40 (−1.94)
1955 (2.11)	1919 (−1.81)	1987/88 (1.62)	1906/07 (−1.47)
1910 (1.71)	1970 (−1.71)	2002/03 (1.55)	1951/52 (−1.40)
2000 (1.32)	1937 (−1.45)	1897/98 (1.54)	1931/32 (−1.32)
1985 (1.06)	1991 (−1.45)	1954/55 (1.46)	1915/16 (−1.02)
1979 (0.97)	2003 (−1.35)	2004/05 (1.15)	1984/85 (−1.01)
1945 (0.95)	1961 (−1.35)	1936/37 (1.11)	1924/25 (−0.94)
1957 (0.90)	1916 (−1.34)	1930/31 (1.09)	1933/34 (−0.67)
2005 (0.89)	1949 (−1.30)	1999/2000 (0.89)	1948/49 (−0.60)
1905 (0.86)	1963 (−1.12)	1972/73 (0.72)	1997/98 (−0.59)

Figure 7e shows the power spectra of the difference of Shen F52 and Menne. The distinct peak of low frequency implies a nonrandom bias between Shen F52 and Menne, which is again caused by different grid resolution and data gridding methods. Therefore, Figs. 7d and 7e imply that gridding and averaging methods may cause nontrivial biases, the sizes of which can be comparable to or larger than the adjusted bias for station data. It is hence important to quantify the uncertainties of each gridding or averaging method (Shen et al. 1998, 2007).

e. Ranking of the hottest and coldest years according to the U.S. average SAT

Table 2 displays the results of the top 10 extreme winters and summers. The first two columns of Table 2 display the 10 hottest and coolest summers [June–August (JJA) average], respectively, between 1895 and 2008 according to Tmax. The third and fourth columns display the 10 warmest and coldest winters [December–February (DJF) average], respectively, between 1895 and 2008 according to Tmin. The values in parentheses are the U.S. average seasonal mean Tmax or Tmin anomalies with respect to the 1961–90 normal. The years are sorted in descending order of the absolute values of the anomalies. In the 114 yr from 1895 to 2008, 5 of the 10 warmest winters according to the DJF Tmin occurred after 1987. Among the other five, two occurred during the Dust Bowl period of the 1930s, one occurred in 1897/98, one occurred in 1954/55 (in the 1950s warming period), and one occurred in 1972/73. The top three warmest winters all occurred after 1987, and the warmest winter was 1990/91. Eight of the 10 coldest winters, also according to the DJF Tmin, occurred before 1952, with the coldest winter being 1939/40. The U.S. annual warming trend is mainly attributed to this winter warming: frequent cold winters in the earlier years and frequent warm winters in the recent period. This is consistent with Gleason et al. (2008)'s finding that the amount of U.S. area that experiences hot extremes has increased in recent years. Other noticeable warm events occurred during the long Dust Bowl drought of the 1930s and a short warm period during the 1950s. The hottest summer on record occurred in 1934, with the JJA Tmax anomaly being 2.12°C. The second hottest summer was 1955, with the JJA Tmax anomaly being 2.11°C. These two anomalies are well within the range of uncertainty, making it impossible to definitely say which year was warmest. The 1954/55 winter was warm too: the fifth-warmest winter in 1895–2008, with a DJF Tmin anomaly of 1.46°C.

Tmin also provides a good measure of the severity of summer heat since, for example, high minimum temperatures during a summer heat wave can result in significant heat stress on people, animals, and plants (D'Ippoliti 2010; Gaffen and Ross 1998). Similarly, Tmax also provides a measure of the severity of winter coldness since maximum temperature during a winter cold spell can indicate the severity of the cold.

Table 3 shows the top 10 hottest and coolest summers according to Tmin and the top 10 warmest and coldest winters according to Tmax between 1895 and 2008. According to Tmin, the top two hottest summers were 1955 (during the short period of the 1950s warming) and 1934 (during the U.S. Dust Bowl period). The top two coolest summers were 1963 and 1919, both of which were among the two CONUS cool periods: 1895–1930 and 1961–85 (Weithmann 2011). On the basis of Tmax, the warmest winter was 1987/88, which occurred during a strong El-Niño episode. The coldest winter was 1906/07, which occurred during a La Niña episode. Another noticeable feature is that the top 10 coldest winters occurred before 1961, much earlier than when the U.S. SAT entered its ascending mode in the 1980s.

Table 4 shows the SAT rankings of annual means. Six of the 10 hottest years occurred after 1990 according to the annual average Tmean, and two occurred during the Dust Bowl drought period in the 1930s. Also, according

TABLE 3. The 10 hottest and coolest summers between 1895 and 2008 from Tmin, and the 10 warmest and coldest winters from Tmax from the TOB-adjusted data. The numbers in the parentheses are the corresponding anomalies relative to the 1961–90 climatology (°C).

10 hottest summers (JJA avg Tmin)	10 coolest summers (JJA avg Tmin)	10 warmest winters (DJF avg Tmax)	10 coldest winters (DJF avg Tmax)
1955 (1.81)	1963 (−1.53)	1987/88 (1.76)	1906/07 (−1.92)
1934 (1.68)	1919 (−1.49)	1897/98 (1.52)	1931/32 (−1.77)
1910 (1.62)	1906 (−1.39)	1943/44 (1.37)	1915/16 (−1.42)
1979 (1.25)	2003 (−1.31)	1990/91 (1.31)	1901/02 (−1.39)
2005 (1.11)	1937 (−1.28)	1954/55 (1.31)	1939/40 (−1.37)
1997 (1.11)	1949 (−1.22)	1956/57 (1.24)	1933/34 (−1.21)
1900 (0.92)	1954 (−1.17)	1913/14 (1.19)	1951/52 (−0.95)
1985 (0.85)	1964 (−1.07)	1898/99 (1.18)	1934/35 (−0.90)
1931 (0.85)	1961 (−1.06)	2002/03 (1.13)	1960/61 (−0.89)
1996 (0.75)	1939 (−1.01)	1936/37 (1.09)	1918/19 (−0.80)

to the annual average Tmean, 8 of the 10 coldest years occurred before 1924. This upward trend is more dramatic for the annual average Tmin. Among the top 10 hottest years, 9 occurred after 1986. Among the 10 coldest years, 9 occurred before 1929. This Tmin warming is not only the primary contributor to the U.S. warming climate but also the main contributor to the decrease of the diurnal range of temperature (Karl et al. 2009). The Dust Bowl of the 1930s and the short-lived early-1950s drought are well reflected in the annual mean Tmax. According to Tmax, 5 of the 10 hottest years were in these two periods and 1934 was the hottest year. These high Tmax values make primary contributions to the 1930s warmth, whereas the high Tmin values explain the recent warming.

The recent warming in CONUS temperature appears to be synchronized with the global Tmean, but the U.S. 1930s Dust Bowl warming was not. The hottest three of the U.S. average annual Tmean from 1895 to 2008 were 1998 (1.20°C), 2006 (1.16°C), and 1934 (1.08°C). According to Hansen et al. (2010), on the basis of a 1951–80 mean, the hottest three of the global annual Tmean from 1895 to 2008 were 2005 (0.63°C), 2007 (0.58°C), and 1998 (0.56°C). According to results of Jones et al. (2011) and Brohan et al. (2006) that are based on a 1961–90 mean, the hottest three of the global annual Tmean from 1895 to 2008 were 1998 (0.529°C), 2005 (0.474°C), and 2003 (0.467°C). The warmest three years globally as calculated at NCDC from departures from the twentieth-century mean are 2005 (0.63°C), 1998 (0.62°C), and 2003 (0.60°C). In these datasets, the strong El Niño influence for 1998 stands out, but the hot year 1934 is only reflected in the U.S. Tmean.

The absolute values of the Tmax anomalies (columns 2 and 5) and Tmin anomalies (columns 3 and 6) in Table 4 are in general much larger than those for the Tmean anomalies (columns 1 and 4), since Tmean is a smoother variable than Tmax and Tmin. Mathematically, this is related to the following formula:

$$\mathrm{Var}\left(\frac{T_{\max} + T_{\min}}{2}\right) \leq \frac{\mathrm{Var}(T_{\max}) + \mathrm{Var}(T_{\min})}{2}. \qquad (19)$$

Table 4 does not imply a temporal increase of SAT variations. As a matter of fact, the SAT variances demonstrate a decreasing trend in the twentieth century (Shen et al. 2011), which is due to the fact that the pdf has

TABLE 4. The 10 hottest and coldest years between 1895 and 2008 from annual Tmean, Tmax, and Tmin (TOB-adjusted data). The numbers in the parentheses are the corresponding anomalies relative to the 1961–90 climatology (°C).

10 hottest years			10 coldest years		
Tmean	Tmax	Tmin	Tmean	Tmax	Tmin
1998 (1.20)	1934 (1.39)	1998 (1.58)	1917 (−0.97)	1912 (−0.87)	1917 (−1.25)
2006 (1.16)	2006 (1.09)	2006 (1.20)	1895 (−0.83)	1993 (−0.71)	1895 (−0.96)
1934 (1.08)	1921 (1.09)	2005 (1.01)	1912 (−0.82)	1895 (−0.68)	1924 (−0.87)
1921 (0.97)	1953 (1.05)	2001 (0.92)	1924 (−0.67)	1917 (−0.68)	1912 (−0.76)
1999 (0.95)	1939 (1.04)	2007 (0.89)	1903 (−0.63)	1920 (−0.60)	1916 (−0.74)
1931 (0.89)	1999 (1.01)	2004 (0.89)	1916 (−0.53)	1903 (−0.59)	1903 (−0.67)
1990 (0.82)	1931 (0.98)	1986 (0.88)	1899 (−0.50)	1905 (−0.55)	1904 (−0.67)
1953 (0.82)	1954 (0.95)	1999 (0.87)	1979 (−0.50)	1982 (−0.54)	1976 (−0.67)
2001 (0.80)	1990 (0.86)	1921 (0.83)	1920 (−0.48)	1978 (−0.54)	1929 (−0.52)
2005 (0.79)	1998 (0.81)	1991 (0.82)	1978 (−0.47)	1979 (−0.52)	1899 (−0.50)

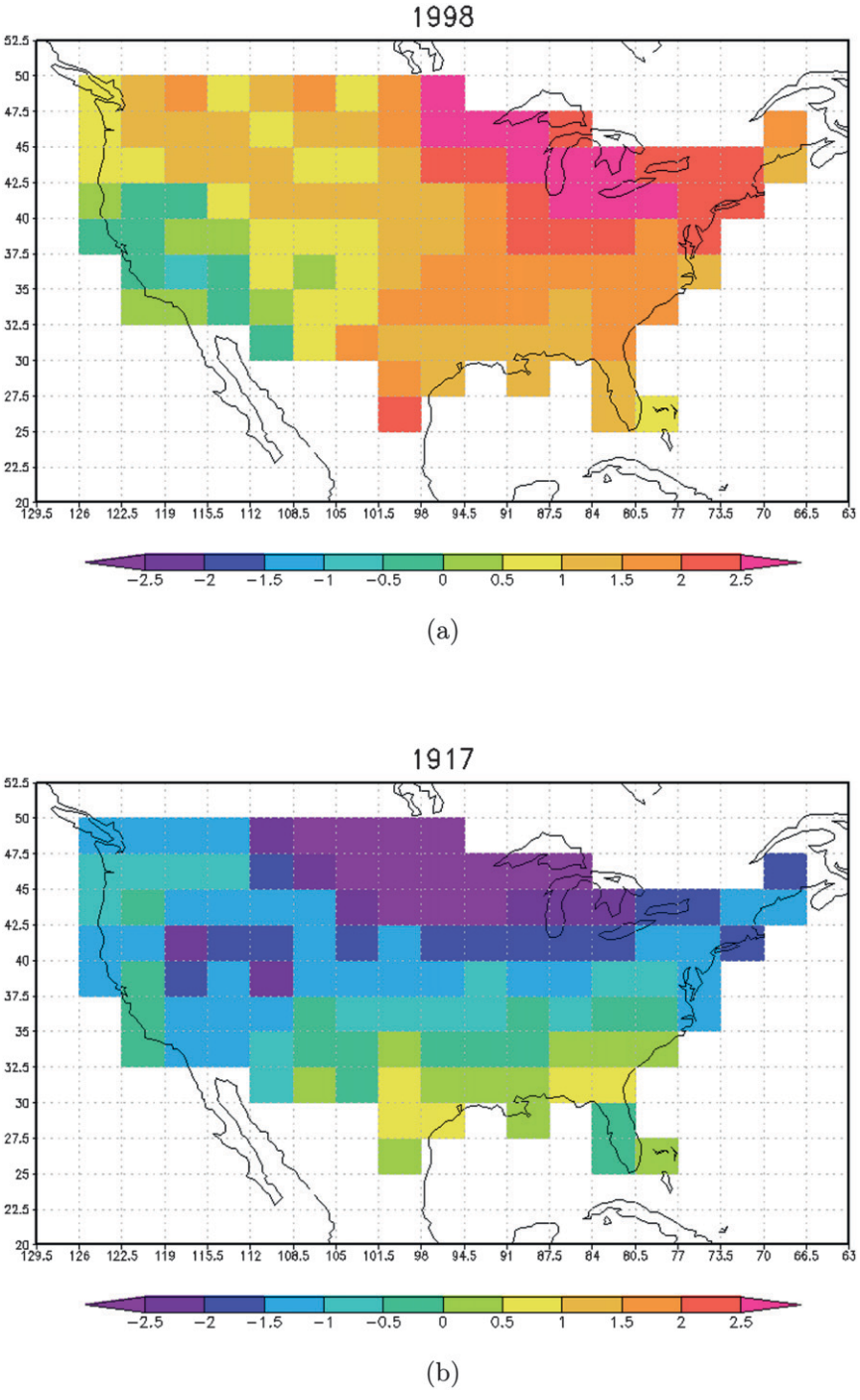


FIG. 8. The spatial distributions of the temperature anomalies (relative to the 1961–90 climatology; °C) of the hottest and coldest years in the United States: (a) 1998, the hottest year, and (b) 1917, the coldest year.

become taller and slimmer in the late warming period (1976–2000) than the earlier cooling time (1946–75).

Figure 8 shows the spatial distribution of the SAT anomalies of the U.S. hottest year (1998) and the coldest year (1917). For the hottest year 1998, according to the annual average temperature Tmean, the entire CONUS except the southwestern region had positive anomalies. The largest anomalies are distributed over the Great Lakes region, where the anomalies exceeded 3°C in many places. These strong warm anomalies may be a result of

the superposition of the warming SAT trend, the strong 1997–98 El Niño, and a weak Arctic Oscillation (AO) (Karl et al. 2009). The temporal mean of the U.S. SAT during the El Niño winters (DJF) demonstrates a clear warm anomaly over the entire northern CONUS, with the largest anomalies over the Great Lakes. The weak AO in the 1997/98 winter made the northern United States dry, and the associated pressure and geopotential height patterns (Thompson and Wallace 1998) extended the warm anomalies farther south. Cold anomalies still occurred over California and its vicinity. Thus, the 1998 U.S. SAT demonstrated strong spatial inhomogeneities.

Figure 8b displays the SAT distribution of the coldest SAT: 1917. The strongest cold anomalies were over the northern Great Plains, Midwest, and Great Lakes, with some regions showing a negative anomaly exceeding −3°C. The southeastern United States, including Georgia, however, still showed weak positive anomalies. Between 1895 and 1998, 1917 had the strongest La Niña episode with a duration of 21 months from 1916 to 1918 (Giese and Ray 2011). The persistent cold tropical Pacific SST induced midlatitude atmospheric circulation changes and might have caused the cooling of the 1917 SAT over the entire western United States and other CONUS regions that are normally affected by the easterly moving weather patterns (Fig. 8b). The AO of 1917 was not strong (Thompson and Wallace 1998). Hence, the 1917 SAT over Minnesota and other northern states east of the Great Plains and west of the Great Lakes was not heavily influenced by the dominant La Niña and was warmer than average.

4. Discussion and possible future work

Most often a linear trend assessment in climate research performed using standard statistical software does not consider the data error. However, a full assessment of trends requires that data error be considered, particularly if the trend is weak and data errors are not small (section 3c). Our method of regression including explanatory data errors can be useful for future studies when considering other errors and biases, such as urban warming. We concluded in section 3d that the full-adjustment F52 leads to a warmer trend than does TOB. Another contribution to warming is the heat island effect, quantified by Hansen et al. (2010) for the Global Historical Climatology Network–Monthly data and the USHCN data. They concluded that the urban warming effect for the CONUS is less than 0.1°C for the entire period of 1900–2009 according to the annual Tmean linear trend. This is consistent with their urban warming conclusion for the entire globe, which is less than 0.1°C $(100\ \text{yr})^{-1}$. This is also consistent with other studies that show the urbanization influence on global average temperature trend is insignificant (Jones et al. 1990; Parker 2006; Peterson et al. 1999). Although these errors or biases do not alter the overall warming conclusion qualitatively, their influence on the uncertainty of the warming trend should be comprehensively quantified in the future. Future studies will also need to attribute the uncertainties of other types of errors and biases for Tmax and Tmin (Parker and Horton 2005; Weithmann 2011).

While our data aggregation from stations to a grid box is through simple averaging, to further reduce the errors and uncertainties in the product of both gridded USHCN and the spatial average SAT for the United States, future work could include the use of optimal averaging (OA) theory (Shen et al. 1998). When applying OA to develop the 2.5° × 3.5° gridded data, fine-resolution climate model data are needed, for example, a 0.5° × 0.5° reanalysis dataset for 50 yr or more. The high resolution is necessary to resolve the station locations in the model data and to demonstrate the spatial inhomogeneity within a grid box. Before this dataset becomes available, one can still explore the OA for calculating the large-scale contiguous U.S. average for the monthly or annual Tmean by using gridded USHCN V2 data and following the method of Folland et al. (2001) and Shen et al. (1998). The method can incorporate the observational errors and the sampling errors. Bias correction (Menne et al. 2009) and optimal averaging (Shen et al. 1998) are essential procedures to reduce uncertainties in a climate change assessment. The optimal average has an advantage of using fewer stations. One can choose non-urban long-term stations to reduce the uncertainty of the heat island effect. Various kinds of accurate calculation of spatial averages of observed data and rigorous statistical inference are helpful in climate model validations for not only SAT but also other parameters, including precipitation and radiation. An example of this model validation is a critical assessment of why there is a discrepancy between the radiosonde data and general circulation model simulations of the tropical tropospheric temperature trend since 1979 (Titchner et al. 2009).

Tables 2 and 3 show very small differences in SAT anomalies for some years. Some differences are discerned at the third decimal place of a Celsius degree. As shown in Fig. 7 and section 3d, these differences can be much smaller than differences that result from the selection of data aggregation methods and spatial averaging methods. Therefore, it is desirable to accurately calculate the U.S. average and to establish a rigorous statistical inference and understanding of uncertainty when determining rankings of the hottest and coldest years. Although these two processes will help to quantify the uncertainties in ranking the CONUS hottest and coldest years and in

ranking the hot and cold temporal regimes, dynamical explanations will still need to be developed using reanalysis and other model data in conjunction with the improved temperature statistics (Compo et al. 2011).

5. Conclusions

We have aggregated the USHCN V2 monthly daily-mean SAT data onto 113 2.5° × 3.5° grid boxes over the contiguous United States from January 1895 to December 2008 and estimated the sampling error variances for each grid box and each month when station data are available in the box. The data were used to assess the trends as well as the hottest and coldest years for the CONUS since 1895. The sampling error variances are smaller over the eastern United States than those over the western mountain regions and the southern coastal areas, mainly because of the station density differences. The spatial correlation length-scale difference between the eastern and western United States that results from different land cover may play a role also. The SAT increase has mainly been attributed to winter warming, particularly in February, which has a warming trend of 0.162°C (10 yr)$^{-1}$. Two major warm periods in 1895–2008 were identified: 1) the 1930s Dust Bowl drought until 1955 and 2) the recent persistent and strong warmth since the 1980s. Eight of the 10 warmest winters according to the seasonal Tmin and 9 of the 10 hottest years from annual Tmean were in these two periods. The sampling error analysis, the previous studies on observational errors, and the comparison between our current work and Menne et al. (2009) reveal the impact of station errors, sampling errors, and consequences resulting from different grid sizes and data aggregation methods. Although these errors may be of nontrivial magnitude and may influence the rank of the hottest or coldest years, they are not large enough to alter the trend of the CONUS SAT.

Acknowledgments. This study was supported in part by the U.S. National Oceanic and Atmospheric Administration (Award EL133E09SE4048), the U.S. National Science Foundation (Awards AGS-1015926 and AGS-1015957), and the U.S. Department of Energy (Award DE-SC002763). Alex Weithmann, Tobias Regele, Max Velado, Julien Pierret, David New, and Olaf Wied helped to produce some figures. Discussion with Barbara Bailey was very helpful. We thank the anonymous reviewers for their valuable suggestions that helped to significantly improve the quality of this paper.

REFERENCES

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, doi:10.1029/2005JD006548.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, 2006: *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Chapman and Hall, 455 pp.

Cochran, W. G., 1977: *Sampling Techniques*. 3rd ed. Wiley, 428 pp.

Compo, G. P., and Coauthors, 2011: The Twentieth Century Reanalysis Project. *Quart. J. Roy. Meteor. Soc.,* **137,** 1–28.

D'Ippoliti, D., and Coauthors, 2010: The impact of heat waves on mortality in 9 European cities: Results from the EuroHEAT project. *Environ. Health,* **9,** doi:10.1186/1476-069X-9-37.

Easterling, D. R., T. R. Karl, E. H. Mason, P. Y. Hughes, and D. P. Bowman, 1996: United States Historical Climatology Network (U.S. HCN) monthly temperature and precipitation data. Oak Ridge National Laboratory Carbon Dioxide Information Analysis Center Rep. ORNL/CDIAC-87, NDP-019/R3, 280 pp.

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

Gaffen, D. J., and R. J. Ross, 1998: Increased summertime heat stress in the US. *Nature,* **396,** 529–530.

Giese, B. S., and S. Ray, 2011: El Niño variability in simple ocean data assimilation SODA). *J. Geophys. Res.,* **116,** C02024, doi:10.1029/2010JC006695.

Gleason, K. L., J. H. Lawrimore, D. H. Levinson, T. R. Karl, and D. J. Karoly, 2008: A revised U.S. climate extremes index. *J. Climate,* **21,** 2124–2137.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.,* **48,** RG4004, doi:10.1029/2010RG000345.

Huang, N. E., and S. S. P. Shen, Eds., 2005: *Hilbert–Huang Transform and Its Applications*. Interdisciplinary Mathematical Sciences, Vol. 5, World Scientific, 360 pp.

Jones, P. D., P. Ya. Groisman, M. Coughlin, N. Plummer, W.-C. Wang, and T. R. Karl, 1990: Assessment of urbanization effects in time series of surface air temperatures over land. *Nature,* **347,** 169–172.

——, T. J. Osborn, and K. R. Briffa, 1997: Estimating sampling errors in large-scale temperature averages. *J. Climate,* **10,** 2548–2568.

——, D. E. Parker, T. J. Osborn, and K. R. Briffa, cited 2011: Global and hemispheric temperature anomalies - Land and marine instrumental records. [Available online at http://cdiac.ornl.gov/trends/temp/jonescru/jones.html.]

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

——, J. M. Melillo, and T. C. Peterson, Eds., 2009: *Global Climate Change Impacts in the United States*. Cambridge University Press, 188 pp.

Mann, M. E., 2004: On smoothing potentially non-stationary climate time series. *Geophys. Res. Lett.,* **31,** L07214, doi:10.1029/2004GL019569.

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

——, ——, and R. S. Vose, 2009: The U.S. Historical Climatology Network monthly temperature data, version 2. *Bull. Amer. Meteor. Soc.,* **90,** 993–1007.

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

——, and B. Horton, 2005: Uncertainties in central England temperature 1878–2003 and some improvements to the maximum and minimum series. *Int. J. Climatol.,* **25,** 1173–1188.

Peterson, T. C., K. P. Gallo, J. Lawrimore, T. W. Owen, A. Huang, and D. A. McKittrick, 1999: Global rural temperature trends. *Geophys. Res. Lett.,* **26,** 329–332.

Regele, T., 2010: USHCN data gridding, error estimation, and the changes of the extreme weather over the US since 1895. M.S. thesis, Dept. of Mathematics and Statistics, San Diego State University, 103 pp.

Shen, S. S. P., T. M. Smith, C. F. Ropelewski, and R. E. Livezey, 1998: An optimal regional averaging method with error estimates and a test using tropical Pacific SST data. *J. Climate,* **11,** 2340–2350.

——, H. Yin, and T. M. Smith, 2007: An estimate of the error variance of the gridded GHCN monthly surface air temperature data. *J. Climate,* **20,** 2321–2331.

——, A. B. Gurung, H.-S. Oh, T. Shu, and D. R. Easterling, 2011: The twentieth century contiguous US temperature changes indicated by daily data and higher statistical moments. *Climatic Change,* **109,** 287–317.

Smith, T. M., R. W. Reynolds, and C. F. Ropelewski, 1994: Optimal averaging of seasonal sea surface temperatures and associated confidence interval (1860–1989). *J. Climate,* **7,** 949–964.

Solomon, S., D. Qin, M. Manning, M. Marquis, K. Averyt, M. M. B. Tignor, H. L. Miller Jr., and Z. Chen, Eds., 2007: *Climate Change 2007: The Physical Science Basis*. Cambridge University Press, 996 pp.

Thompson, D. W. J., and J. M. Wallace, 1998: The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett.,* **25,** 1297–1300.

Titchner, H. A., P. W. Thorne, M. P. McCarthy, S. F. B. Tett, L. Haimberger, and D. E. Parker, 2009: Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments. *J. Climate,* **22,** 465–485.

Vose, R. S., C. N. Williams Jr., T. C. Peterson, T. R. Karl, and D. R. Easterling, 2003: An evaluation of the time of observation bias adjustment in the U.S. Historical Climatology Network. *Geophys. Res. Lett.,* **30,** 2046, doi:10.1029/2003GL018111.

Wackerly, D. D., W. Mendenhall III, and R. L. Scheaffer, 2002: *Mathematical Statistics with Applications*. 6th ed. Duxbury, 853 pp.

Wang, X., and S. S. P. Shen, 1999: Estimation of spatial degrees of freedom of a climate field. *J. Climate,* **12,** 1280–1291.

Washington, W. M., and C. L. Parkinson, 1986: *An Introduction to Three-Dimensional Climate Modeling*. University Science Books, 422 pp.

Wei, W. W. S., 2006: *Time Series Analysis: Univariate and Multivariate Methods*. 2nd ed. Pearson Addison Wesley, 614 pp.

Weithmann, A., 2011: Optimal averages of U.S. temperature, error estimates and inferences. M.S. thesis, Dept. of Mathematics and Statistics, San Diego State University, 96 pp.

Wigley, T. M. L., K. R. Briffa, and P. D. Jones, 1984: On the average value of correlated time series, with applications in dendroclimatology and hydrometeorology. *J. Climate Appl. Meteor.,* **23,** 201–213.

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, doi:10.1029/2011JD016761.

Willmott, C. J., C. M. Rowe, and W. D. Philpot, 1985: Small-scale climate maps: A sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring. *Amer. Cartogr.,* **12,** 5–16.