Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1029/2018JD028355

Key Points:

- Four land surface air temperature data sets are considerably different at regional and local scales
- Largest data set differences are over the tropics, high latitudes and Africa where the number of stations is limited
- Data set differences could lead to inconsistent or even contrasting regional/local surface warming trend estimation

Supporting Information:

- Supporting Information S1

Correspondence to:
S. Liang,
sliang@umd.edu

Citation:
Rao, Y., Liang, S., & Yu, Y. (2018). Land surface air temperature data are considerably different among BEST-LAND, CRU-TEM4v, NASA-GISS, and NOAA-NCEI. *Journal of Geophysical Research: Atmospheres*, *123*, 5881–5900. https://doi.org/10.1029/2018JD028355

Received 19 JAN 2018
Accepted 19 MAY 2018
Accepted article online 28 MAY 2018
Published online 7 JUN 2018

©2018. American Geophysical Union.
All Rights Reserved.

Land Surface Air Temperature Data Are Considerably Different Among BEST-LAND, CRU-TEM4v, NASA-GISS, and NOAA-NCEI

Yuhan Rao[1], Shunlin Liang[1,2], and Yunyue Yu[3]

[1]Department of Geographical Sciences/Cooperative Institute for Climate and Satellites-Maryland, University of Maryland, College Park, MD, USA, [2]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei, China, [3]Center for Satellite Applications and Research, NESDIS, NOAA, College Parkm, MD, USA

Abstract Several groups routinely produce gridded land surface air temperature (LSAT) data sets using station measurements to assess the status and impact of climate change. The Intergovernmental Panel on Climate Change Fifth Assessment Report suggests that estimated global and hemispheric mean LSAT trends of different data sets are consistent. However, less attention has been paid to the intercomparison at local/regional scales, which is important for local/regional studies. In this study we comprehensively compare four data sets at different spatial and temporal scales, including Berkley Earth Surface Temperature land surface air temperature data set (BEST-LAND), Climate Research Unit Temperature Data Set version 4 (CRU-TEM4v), National Aeronautics and Space Administration Goddard Institute for Space Studies data (NASA-GISS), and data provided by National Oceanic and Atmospheric Administration National Center for Environmental Information (NOAA-NCEI). The mean LSAT anomalies are remarkably different because of the data coverage differences, with the magnitude nearly 0.4°C for the global and Northern Hemisphere and 0.6°C for the Southern Hemisphere. This study additionally finds that on the regional scale, northern high latitudes, southern middle-to-high latitudes, and the equator show the largest differences nearly 0.8°C. These differences cause notable differences for the trend calculation at regional scales. At the local scale, four data sets show significant variations over South America, Africa, Maritime Continent, central Australia, and Antarctica, which leads to remarkable differences in the local trend analysis. For some areas, different data sets produce conflicting results of whether warming exists. Our analysis shows that the differences across scales are associated with the availability of stations and the use of infilling techniques. Our results suggest that conventional LSAT data sets using only station observations have large uncertainties across scales, especially over station-sparse areas. In developing future LSAT data sets, the data uncertainty caused by limited and unevenly distributed station observations must be reduced.

1. Introduction

The Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report concludes that it is confident that the global land surface air temperature (LSAT) has warmed since the 1900s, and the increase after the 1970s has been much faster than previous years (Hartmann et al., 2013). This high confidence is based on consistent results using four LSAT data sets produced independently by Berkley Earth, National Aeronautics and Space Administration (NASA) Goddard Institute of Space Studies (GISS), National Oceanic and Atmospheric Administration (NOAA) National Center for Environmental Information (NCEI), and Climate Research Unite (CRU) at University of East Anglia (Hartmann et al., 2013; Jones, 2016). This consistency is only achieved after improvements of these data sets in the recent decade. Jones (2016) asserts that the consistency of the large-scale temperature estimates mainly resulted from (1) similar raw input station data, (2) similar methods for correcting biases and adjusting inhomogeneity of the raw data, and (3) spatial autocorrelation of the temperature data.

Despite the global LSAT being one of the most direct indicators of climate change, it has very little direct impact on ecosystems and human societies, which are mainly influenced by local and regional temperature variations. An analysis based on the global LSAT reflects the general status of the surface temperature over the global land but misses the crucial spatial pattern of the surface temperature changes that directly influence ecosystems and millions of people (Editorial, 2017). This spatial pattern of the LSAT change will directly affect the essential functions of human and natural systems, such as vegetation productivity, hydrological events (e.g., snow melting and surface runoff), and human health. To produce global temperature records, these institutions usually generate gridded data sets first with various methods using preprocessed station-based observations at coarse grid boxes (e.g., 1–5°; Hansen et al., 2010; Jones et al., 2012; Muller et al., 2013; Vose et al., 2012). These gridded data sets have been used in various studies to quantify LSAT changes and assess LSAT's impact on human and natural systems at different spatial scales.

Unfortunately, the confidence regarding spatial details of the LSAT change is still low, especially for regions with sparse stations (Hartmann et al., 2013). Most regional- and local-scale studies mainly focus on regions with abundant ground-based observations, such as Continental United States, China, Australia, and Europe. Regional studies like these could draw relatively confident conclusions of the regional mean LSAT change. However, the confidence is usually low when it comes to the spatial details of the LSAT change, which could partially be attributed to the station data quality and different preprocessing and gridding methods. Fall et al. (2011) question the potential large biases in observations collected by the U.S. Historical Climatology Network, of which many stations cannot meet the official World Meteorology Organization siting guidance. However, the overall biases of the network in recent decades can be better explained by instrumental changes rather than siting biases (Hartmann et al., 2013; Menne et al., 2010). The confidence of regional analysis is worse for observation-sparse regions, such as the Antarctic, high mountains, and other sparsely populated areas. Research can usually agree on the sign of the LSAT change (i.e., warming or cooling) for the regional mean LSAT by interpolating available sparse ground-based observations. However, a significant inconsistency or even disagreement in the magnitude and spatial pattern of the LSAT change has been observed (Thorne et al., 2016).

Surprisingly, little attention has been paid to quantify differences among different gridded temperature data sets and to assess their impact on the LSAT trend calculation at regional and local scales with the importance and wide application of these data sets (Thorne et al., 2016). Since IPCC's Fourth Assessment Report, many efforts have been made to improve the data quality of individual gridded LSAT data set by improving spatial coverage, preprocessing, and gridding methods (Hansen et al., 2010; Jones et al., 2012; Muller et al., 2013; Vose et al., 2012). The ultimate goal of these improvements is to reconcile the differences of the global mean LSAT calculated from different data sets, which has been proven successful (Thorne et al., 2016). However, no comprehensive quantification of the data set differences and the impact on the LSAT trend analysis has been made at the regional and local scales (Thorne et al., 2016). Vose et al. (2005) conduct an intercomparison between the gridded LSAT data sets produced by CRU, GISS, and NCEI at the global, hemispheric, and grid box levels. Their study focused on the impact of different gridding techniques and averaging methods on the global and hemispheric mean LSAT (Vose et al., 2005). They only provide a comparison of the estimated linear LSAT trends at the grid box level for the CRU and NCEI data and conclude that a general agreement is met at the grid box level with large regional variations (Vose et al., 2005). Furthermore, with the recent improvements of these data sets and the new development of the Berkley Earth LSAT data set, this topic must be comprehensively revisited (Muller et al., 2013).

In addition, gridded LSAT data sets are all constructed using available observations collected by national and regional station networks, which are constantly changing through time (Menne & Williams, 2009). In general, the availability of observations has significantly increased, especially during 1950–1980 (Hansen et al., 2010). However, the change of the data availability is not steady through time, and not even across continents (Hegerl et al., 2014). Different data sets have extended their spatial coverage over different regions by including networks from various agencies and research groups over high latitudes, such as the Antarctic and Greenland (Hansen et al., 2010; Jones et al., 2012). Despite these *data-hunting* efforts, the number of stations used in most, if not all, data sets even decreased in the recent decades (Hansen et al., 2010; Hegerl et al., 2014). This reduction is mainly caused by the elimination of stations in South America and Africa. The reduction could be alarming for both developers and users of these data sets mainly because this will increase the dependence of the LSAT data sets on specific stations over data-sparse regions. With recent effort, the Global Historical Climatology Network (GHCN)-monthly data (version 4) reverse this decreasing pattern by including more station data since 1980s (Rennie et al., 2014). This new station data will be used as main input for both NOAA and NASA for future products. It is still worthwhile to understand how the variation across data sets evolves through time in response to the changing data availability.

The present study focuses on quantifying the differences among four gridded LSAT data sets and their impact on the LSAT trend estimation. Even though most of these data sets are using similar raw input data

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1
The Summary of BEST-LAND, CRU-TEM4v, NASA-GISS, and NOAA-NCEI

Data	Grid size	Climatology period	No. of stations	Homogenization method	Interpolation method	Notes
BEST-LAND	1° × 1°	1951–1980	36,866	*scalpel*: Split time series using detected break points and automatically adjust weight for each time series	Gaussian process regression/Kriging	Muller et al. (2013) and Rohde et al. (2013)
CRU-TEM4v	5° × 5°	1961–1990	5,583	Comparing with neighbor stations	No interpolation implemented	Jones et al. (2012)
NASA-GISS	2° × 2°	1951–1980	7,290	Comparing with neighbor stations; urbanization adjustment	Distance-dependent weighted average of station observations within 1,200-km radius	Hansen et al. (2010)
NOAA-NCEI	5° × 5°	1961–1990	7,280	Comparing with neighbor stations	Two-step (low and high frequency) reconstruction using Empirical Orthogonal Teleconnection	Smith et al. (2008) and Vose et al. (2012)

Note. BEST-LAND = Berkley Earth Surface Temperature land surface air temperature data set; CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

(except BEST which uses much more stations than others), different quality control procedures, homogenization methods, and gridding techniques could lead to different spatial coverages and values at the grid box levels (Jones, 2016). This analysis intends to perform a comprehensive assessment to (i) evaluate the data coverage biases of LSAT at the global and regional scales and (ii) quantify the data set differences and their impact on the LSAT trend estimation at the regional and local scales. In this study, we use local scale and grid box scale interchangeably since the grid box level is the finest scale of gridded data sets. Section 2 summarizes the information of individual data sets necessary for the intercomparison and preprocessing procedures to make the intercomparison meaningful. Section 3 presents the difference of global and hemispheric mean LSAT caused by data coverage differences. Sections 4 and 5 present the intercomparison results. The discussion and the conclusion are provided at last to summarize the intercomparison results and the implications for the future LSAT change analysis and the new LSAT data development.

2. Data and Methods

Table 1 summarizes the basic information of four major gridded LSAT data sets used in this comparison. These four data sets have been widely used for assessing the status of climate change and its impact on ecosystem and society (Hartmann et al., 2013; Jones, 2016) because of their rigorous quality control, routine (monthly) updates, good documentation, and completeness of their archive. Each data set is briefly described in the following section to ensure appropriate interpretation of current intercomparison. More detailed information of individual data sets should be directed to the references listed in Table 1.

2.1. Berkley Earth Surface Temperature Data Set

The Berkley Earth Surface Temperature land surface air temperature data set (BEST-LAND) combines station temperature measurements from 14 different sources with a total archive of 44,455 sites (Muller et al., 2013; Rohde et al., 2013). A total of 36,866 sites has been kept for the final BEST-LAND process after removing duplicate stations in different data sources and stations with measurements less than 1 year or missing location metadata (Rohde et al., 2013). The largest data source used by BEST-LAND is the GHCN-Daily (more than 25,000 stations) and GHCN-monthly (GHCNM-v3 with 7280 stations, which will be replaced by GHCNM-v4 later this year) data archive managed by NOAA NCEI. These data contain temperature measurements from 180 countries (Menne et al., 2012). After preprocessing and monthly averaging, the station measurements are interpolated into 15,984 equal-area grid cells (with nearly 1.25° resolution at the equator) for the Earth surface using Gaussian process regression (i.e., Kriging interpolation) and then regridded into 1° × 1° grid boxes (Rohde et al., 2013; Thorne et al., 2016).

The station data used in BEST-LAND are raw data from each data source with no homogenization and limited data quality control. Before the interpolation process, it uses a pairwise method to identify statistical breakpoints within the original data for each station compared with its neighboring stations (Rohde et al., 2013). Unlike other groups, BEST-LAND does not correct the detected discontinuities potentially caused by site relocation, instrument changes, and urbanization effects. Instead, it separates original data into different fragments at detected breakpoints and treats these fragments as data from different stations. This process, called *scalpel*, is intentionally designed to reduce the human bias caused by adjustment (Muller et al., 2013; Rohde et al., 2013). An original archive of 36,866 stations produces 179,928 data fragments with the scalpel process (Rohde et al., 2013). These fragments are then used to produce the temperature data for each grid using Kriging interpolation with an integrated iterative bias adjustment and outlier deweighting process. This process is designed to be tolerant of data records with a limited length, thereby allowing majority of the reliable station observations to be used in the analysis. BEST-LAND provides temperature anomalies from 1850 to present time comparing to the climatology period of 1951–1980.

2.2. Climate Research Unit Temperature Data Set Version 4

CRU Temperature Data Set Version 4 (CRU-TEM4v) is a gridded LSAT data set provided by the Climate Research Unit at the University of East Anglia with the variance adjusted for changing

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

station numbers within each grid (i.e., 5° × 5°). The station temperature measurements used in CRU-TEM4v are combined from multiple data sources with a total of 5,583 stations (Jones et al., 2012). The main data source is the National Meteorology Services (NMSs) of countries across the world comprising 2,444 stations in the final archive. Another important data source for CRU-TEM4v is the decadal World Weather Report starting from 1950s onward, which provides data records for underrepresented nations in NMS data (mainly South America, Africa, Asia, and many island groups; Jones et al., 2012). With its most recent data archive update, CRU-TEM4v has improved its data coverage over the Arctic compared to its predecessor (i.e., CRU-TEM3v) by including more data for the Arctic region provided by Bekryaev et al. (2010) and the Danish Meteorological Institute reports.

Instead of performing homogenization for all stations, CRU mostly relies on homogenized temperature records provided by NMSs (Jones et al., 2012). Additionally, CRU identifies 219 data records for additional adjustment of homogeneity caused by various factors, such as change of instruments, site locations, and local environments. The adjustment is performed by comparing with multiple neighboring stations (Jones et al., 2012). The time series of absolute temperature are then converted into anomalies by simply subtracting the long-term average for each month derived from a base period (i.e., 1961–1990 for CRU-TEM4v) for each station, which is referred to as the climate anomaly method. Finally, a temperature anomaly of each grid box is generated by simple averaging of all the available station anomalies within each grid box (Jones et al., 2012). Station availability changes through time for some grid boxes; hence, CRU-TEM4v has also adjusted the variance of each grid box to account for this factor using the method outlined in Brohan et al. (2006).

2.3. NASA GISS Surface Temperature Data Set

The NASA Goddard Institute for Space Studies (NASA-GISS) produces a global land surface air temperature data set using the reference station method (Hansen et al., 2010). The majority of station temperature time series used by the NASA-GISS is obtained from GHCN-monthly version 3 with a total of 7,280 stations. By selecting stations with at least 20 years of records, nearly 6,300 stations are kept for further analysis. Hansen et al. (2010) also use monthly data collected by the Scientific Committee on Antarctic Research since 1957 to fulfill data gaps in the GHCN station archives. Similar with CRU-TEM4v, the original time series are inspected for homogeneity and adjusted for inhomogeneity if necessary. Moreover, Hansen et al. (2010) utilize satellite night-light radiance data to identify the stations affected by urban effects. They correct these station measurements by comparing them with their neighboring rural stations.

After the adjustments, the station measurements are converted to anomalies compared to their long-term average for each month derived from the base period of 1951–1980. The temperature anomalies for 2° × 2° grid cells are generated by weighted averaging of the anomalies for all stations within 1,200 km of that grid. The weight for each station is a linear function decreasing with its distance from the grid center (Hansen et al., 2010).

2.4. NOAA National Climate Data Center Surface Temperature Data Set

The NOAA-NCEI surface temperature data set at a 5° × 5° grid box is created by separately constructing low-frequency and high-frequency variations. The station observations used by NOAA-NCEI are from the GHCN-monthly data set (GHCNM-v3). The inhomogeneity adjustments for each time series are implemented for each station using a pairwise method compared with its neighboring stations (Menne & Williams, 2009). All homogenized data records are then converted to anomalies with the base period for 1961–1990. For each grid box, the anomalies for the stations within the grid box are averaged into a *super observation*, which is used for the final reconstruction process (Vose et al., 2012).

The reconstruction for the NOAA-NCEI assumes that temperature anomaly time series can be divided into two different components: low-frequency variations that reflect long-term changes and high-frequency variations that represent temperature variability over short-time periods (Smith & Reynolds, 2005; Smith et al., 2008; Vose et al., 2012). Therefore, the low-frequency component is first derived by a simple spatiotemporal smoothing method (Smith et al., 2008). The residual anomaly time series for each grid box is then fitted to a group of large-scale spatial-covariance modes derived from the modern era data (1982–1991, with a maximum spatial coverage) using the empirical orthogonal teleconnections (Smith & Reynolds, 2005; van den Dool et al., 2000). The final temperature anomaly time series for each grid box is obtained by simply adding the smoothed low-frequency time series and the fitted high-frequency time series. The residuals of the

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2
Geographical Boundaries of the Regions Used for Calculating the Spatial Average

Continent	Min. latitude (degrees north)	Max. latitude (degrees north)	Min. longitude (degrees east)	Max. longitude (degrees east)
North America 1 (NA1)	15	50	−165	−50
North America 2 (NA2)	50	85	−165	−50
South America 1 (SA1)	−23.5	15	−90	−30
South America 2 (SA2)	−60	−23.5	−80	−40
Europe (EUR)	35	80	−15	60
Africa (AFR)	−35	30	−20	50
Asia 1 (AS1)	5	50	45	150
Asia 2 (AS2)	50	80	60	180
Maritime Continent (MCT)	−10	5	90	165
Australia (AUS)	−50	−10	110	155
Antarctica (ANT)	−90	−60	−180	180
Greenland (GRL)	60	90	−70	−10

reconstructed data compared with the original data are treated as background noise potentially arising from uneven sampling, observation errors, etc. (Vose et al., 2012). The reconstruction is designed to capture the key trends and patterns while neglecting the local, short-term irregularities, and provide anomalies in unsampled areas by identifying spatial-covariance modes (Smith et al., 2008). For consistency with its ocean counterpart, the anomalies over the land grid box (with base period of 1961–1990) are converted to anomalies comparing to base period of 1971–2000 (Vose et al., 2012).

2.5. Data Processing

Each data set has its own spatial resolution and climatology period; hence, all data sets must be adjusted to the same spatial resolution and climatology period to ensure a meaningful intercomparison. First, NOAA-NCEI and CRU-TEM4v are adjusted to the LSAT anomaly values against the monthly climatology of 1951–1980 by subtracting the 30-year mean value (1951–1980) for each month from the original anomaly. After the climatology adjustment, fine-resolution data sets (i.e., BEST-LAND and NASA-GISS) are aggregated to the spatial resolution of 5° × 5° by weighted average considering the grid area change caused by the latitude.

The comparison is performed at different spatial scales, including global, hemispheric, latitudinal, regional, and local scales. The spatial average is calculated using weighted average considering the grid area change caused by the latitude. The latitudinal average is calculated for each 20° latitudinal band, while the regional average is calculated based on the regions described in Table 2. In addition, the spatial coverage is different across data sets because of the different gridding methods and source data used by each group. Hence, we define two valid spatial coverages for calculating the spatial average: native and common data coverage. The native coverage for a data set includes all grid boxes, of which the data set has a nonmissing anomaly value, whereas the common coverage only includes grid boxes, of which all four data sets have nonmissing anomaly values. All spatial averages are calculated based on both native and common coverages.

The data sets are also compared at different temporal scales, including annual and seasonal mean. Therefore, the monthly anomaly is averaged through a given year or season. In this study, we define four seasons as December–January–February (DJF), March–April–May (MAM), June–July–August (JJA), and September–October–November (SON), where the December value is from the previous year.

For trend calculation, we use ordinary least squares (OLS) method to estimate the linear trend for the given time periods (i.e., 1901–2017, 1951–2017, 1981–2017, and 1998–2017). However, temperature data are usually strongly autocorrelated, which will lead to underestimation of standard error for OLS-estimated trends (Hausfather et al., 2017; Lee & Lund, 2004; Santer et al., 2000). To address this issue, we consider an autoregressive-moving-average model with the order of 1 for each component (i.e., ARMA [1,1]) to adjust standard errors of estimated trends for global, hemispheric, regional, and local scales. More details of the adjustment method can be found in Hausfather et al. (2017) and Lee and Lund (2004).

To test whether LSAT differences across data sets cause significant impact on linear trend calculation, we adopt the method from Hausfather et al. (2017), which calculates the statistical significance of linear trends of data set difference time series. In theory, when data coverage is the same, difference between trends

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

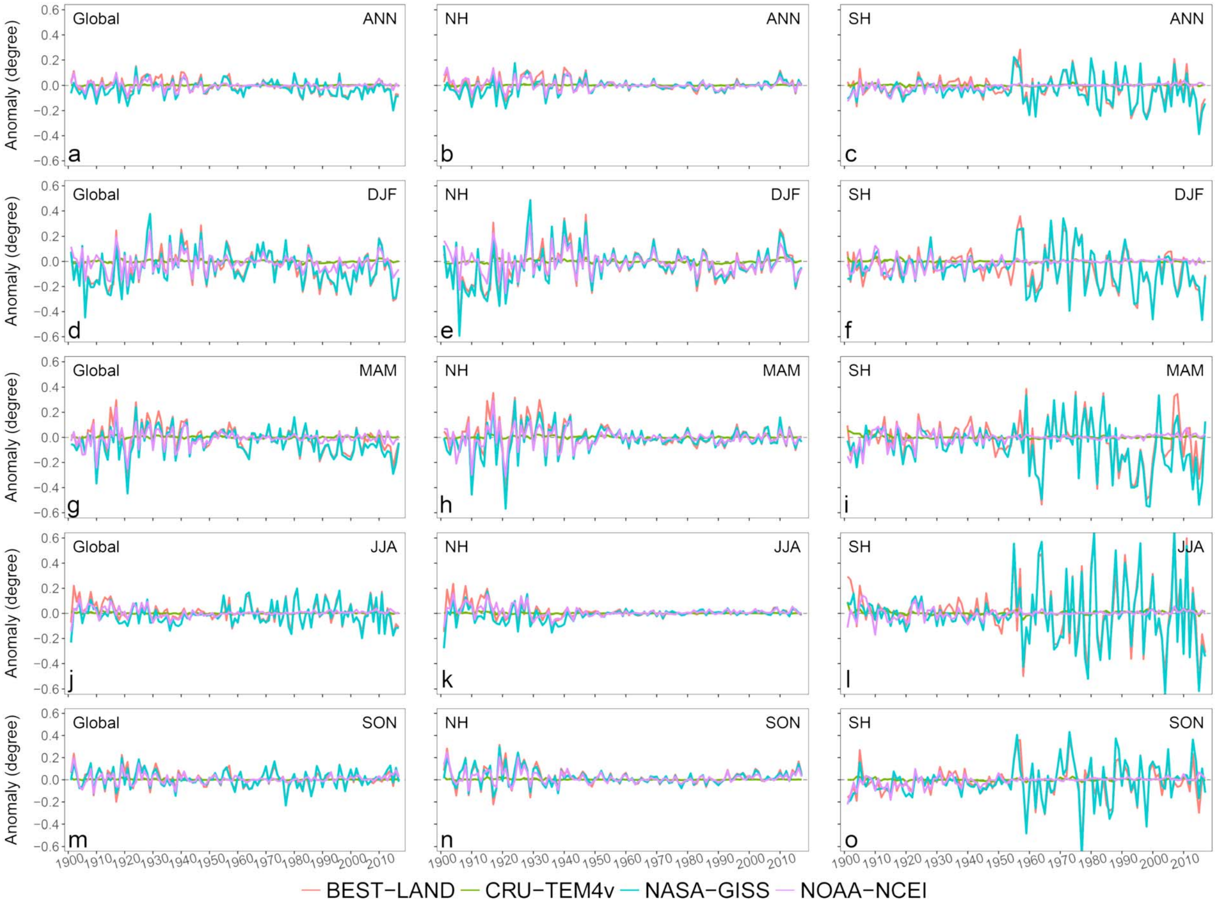

Figure 1. Differences between the temperature anomalies calculated using the native coverage and the common coverage for four data sets (ANN = Annual; DJF = December–January–February; MAM = March–April–May; JJA = June–July–August; SON = September–October–November). BEST-LAND = Berkley Earth Surface Temperature land surface air temperature data set; CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

estimated from two time series is the same as trends estimated from the difference time series. This method can avoid the dependency of different data sets caused by similar source data (Hausfather et al., 2017). Because BEST-LAND uses much more stations to generate the gridded product, we use it as our reference for the difference time series calculation. The difference time series of CRU-TEM4v, NASA-GISS, and NOAA-NCEI comparing to BEST-LAND are then used to estimate the difference trends using OLS. We also use ARMA (1,1) model to address the autocorrelation in difference time series.

3. Differences of the Global and Hemispheric Mean LSATs

We first examine how different spatial coverages affect the calculation of large-scale mean LSAT (i.e., global and hemispheric averages) by comparing the spatial average calculated using different coverages (i.e., native and common coverages) for each data set. Figure 1 shows the differences between the global and hemispheric averages of the native coverage and the ones of the common coverage. The CRU-TEM4v generally has the lowest differences in the mean LSATs calculated using different data coverages because CRU-TEM4v does not use interpolation techniques to fill in the grid boxes with no observations (Cowtan & Way, 2014; Jones et al., 2012). Thus, it has the smallest native coverage similar with the common data coverage for all data sets.

For the global scale, the data coverage-caused differences are relatively stable for the annual mean, ranging between ±0.2°C, but vary remarkably for the seasonal mean, especially for DJF and MAM before 1960, which vary between ±0.4°C. The difference caused by the data coverage is substantial considering that the

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

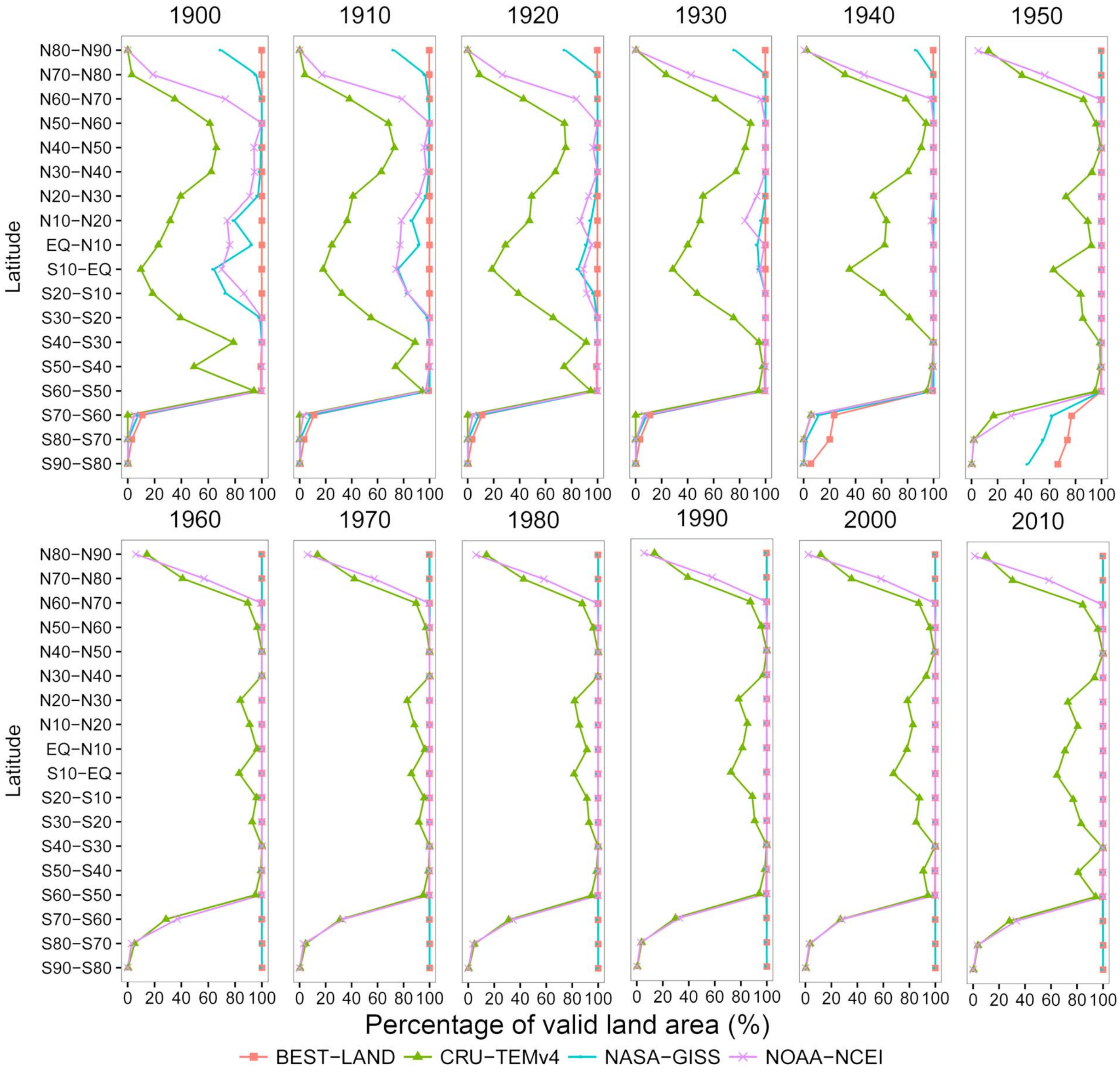

Figure 2. Decadal change of the percentage of land areas with valid land surface air temperature anomaly values for different 10° latitudinal bands for each data set. BEST-LAND = Berkley Earth Surface Temperature land surface air temperature data set; CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

magnitude of global warming is around 1.8°C since 1900 (Jones, 2016). NASA-GISS and BEST-LAND have the largest differences because they both use infilling techniques to estimate temperatures of the grid boxes with no stations. Meanwhile, NOAA-NCEI has smaller differences, followed by CRU-TEM4v. Figure 2 depicts that BEST-LAND and NASA-GISS have the highest percentage of valid land area for the whole study period because of spatial infilling, especially over high latitudes and the tropics.

The difference of the Northern Hemisphere (NH) mean LSAT has the largest variation before 1950s for both annual and seasonal mean LSATs. The variation is relatively smaller for the annual mean and warm seasons (JJA and SON; i.e., ±0.2°C) but much larger for the cold seasons (DJF and MAM) ranging from −0.6 to 0.4°C. The large variation before 1950s is mostly caused by the data coverage differences for 0–70°N (Figure 2). The difference caused by the data coverage of this region is mostly caused by the low data coverage of CRU-TEM4v and NOAA-NCEI, which do not provide LSAT for grid boxes with no station observations (Jones et al., 2012; Smith et al., 2008; Vose et al., 2012).

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In contrast with the NH, the difference for the Southern Hemisphere (SH) mean LSAT shows the largest variations after the 1950s. The difference ranges from −0.4 to 0.3°C for the annual mean LSAT and DJF mean LSAT. It varies within (−0.6, 0.4)°C for MAM/SON and (−0.6, 0.6)°C for JJA. The large variation after 1950s is mostly caused by the rapid increase of the data coverage in BEST-LAND and NASA-GISS (Figure 2). This data coverage improvement is the result of the extensive efforts to add new station observations over the southern high latitudes (Hansen et al., 2010; Rohde et al., 2013). Although the number of stations over the southern high latitudes is still very limited, BEST-LAND and NASA-GISS use spatial interpolation methods to fill in the data gaps, which provides complete data coverage for this region. Notably, the difference in data coverage for the SH results in an obvious trend for certain time periods, such as in MAM 1970–2000 (Figure 1).

Furthermore, we examine the impact of data set differences on the estimated trend for different time periods and seasons (Table S1 in the supporting information). For global mean LSATs, CRU-TEM4v, NASA-GISS, and NOAA-NCEI all show significant trend differences comparing to BEST-LAND for both annual and seasonal during 1951–2017. The difference trends range from 0.01 to 0.02 degree/decade. This positive difference trend pattern also presents in NH for almost all seasons and data sets, while it only appears in NASA-GISS and NOAA-NCEI for SH. During 1981–2017, CRU-TEM4v shows a significant trend difference with BEST-LAND for both global and NH mean, while NASA-GISS and NOAA-NCEI mainly differ with BEST-LAND for SH. Due to relative short time period for 1998–2017, we only find that CRU-TEM4v differs significantly from BEST-LAND for NH trends of mean LSATs.

Although a previous analysis claims that large-scale mean LSAT should be robust against the data set choice (Jones, 2016), our analysis shows that it is sensitive to the data coverage of different data sets. The difference causes significant trend differences estimated from different data. This issue should be examined more carefully in the future IPCC assessment. The users of these data sets should be cautious in terms of the conclusions inferred from the global/hemispheric mean LSAT using only single data set, particularly for seasonal mean temperatures.

4. Latitudinal and Regional LSAT Differences

4.1. Decadal Mean LSAT Comparison

Figures 3 and 4 demonstrate the differences of the regional mean LSAT anomalies for different latitudinal zones and predefined regions in Table 2. We only show the decadal mean values instead of the individual years or months to focus on the systematic differences rather than on the interannual variability. Figure 3 shows that overall LSAT changed remarkably for all regions in all four data sets, especially in recent decades, but with strong latitudinal variations. However, the differences among the four data sets are evident despite the agreement on the general warming patterns.

As shown in Figure 3, the four data sets have the smallest differences at the southern midlatitude and northern middle-to-high latitude (i.e., 30–50°S and 30–70°N), which are regions with rich ground stations (Figure 2). The differences are larger near the equator and the largest in the southern high latitude and polar regions (i.e., 50–70°S, 10°S–10°N, and 70–90°N/S), where only a very limited number of, if any, station observations are available. The differences are as large as 0.8°C for some time periods. Although previous research suggests that the differences among these data sets are reduced at a global scale because of the introduction of more stations (Jones, 2016), we find that the latitudinal differences do not necessarily decrease through time, especially for high latitudes. For example, the difference between CRU-TEM4v and NOAA-NCEI at 70–90°N is approximately 0.6°C for the 21st century, which is much larger than those in the 1980s and 1990s. One possible reason for this pattern is the decline of available stations in certain regions (Figure 2). Additionally, different interpolation methods also contribute to this difference. BEST-LAND and NASA-GISS assign large weights to very high latitude stations to represent unsampled regions in high-latitude regions, while NOAA-NCEI tend to give larger weights to closer regions with more stations available.

For the regional mean LSAT, Figure 4 shows that the four data sets have the highest degree of agreement for North America, Europe, Asia, and Australia. For less-populated regions, such as South America, Greenland, and Antarctica, the differences are much larger than the others. This pattern is highly correlated with the spatial distribution of available ground stations (Figure A1 in Hansen et al., 2010). Similar with the latitudinal comparison, the differences among these data sets show notable increases in recent decades for most continents,

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

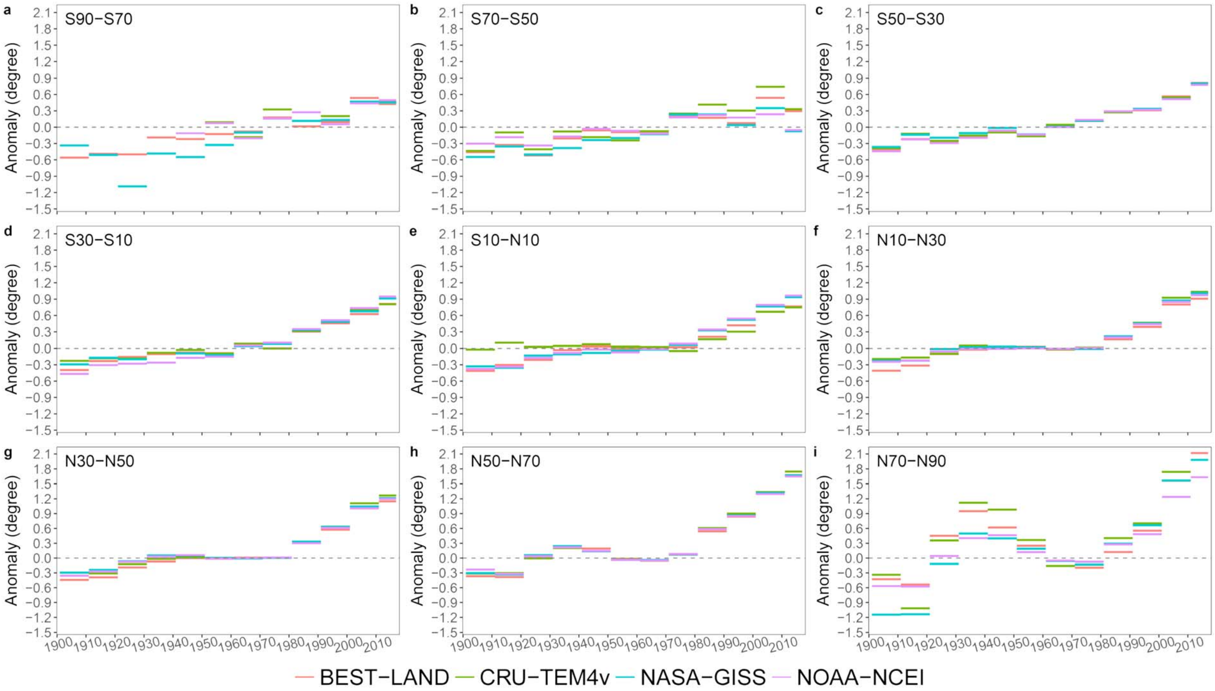

Figure 3. Differences of the decadal annual mean land surface air temperature anomaly between the four data sets for different latitudinal bands. BEST-LAND = Berkley Earth Surface Temperature land surface air temperature data set; CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

including high-latitude North America, South America, Asia, Africa, Maritime Continent, Australia, Antarctica, and Greenland.

4.2. Annual Mean LSAT Comparison

Figure 5 shows the latitudinal and regional comparison using the Taylor diagram (Taylor, 2001). The annual mean LSAT anomaly time series are used in these comparisons. BEST-LAND has the best spatial coverage, and the values calculated from BEST-LAND are used as the reference for the diagram because no *true* values exist for the latitudinal/regional mean LSAT. For the latitudinal comparison (Figure 5a) the Taylor diagram confirms that high-latitude and polar regions have the lowest degree of agreement among these data sets, followed by the tropics. In addition, CRU-TEM4v and NOAA-NCEI tend to have larger deviations from the reference (i.e., BEST-LAND), whereas NASA-GISS has better agreement with BEST-LAND. For the regional comparison (Figure 5b), Antarctica, Greenland, Maritime Continent, and South America show the largest variations, while Europe, Asia, and low- to middle-latitude North America exhibit the best agreement among the four data sets.

4.3. LSAT Warming Trend Comparison for Latitudinal Bands

We compare the latitudinal linear trends estimated using these data sets for different time periods (i.e., 1901–2017, 1951–2017, 1981–2017, and 1998–2017) to examine how the temperature differences among these data sets influence the surface warming trend analysis at different latitudes. Figure 6 illustrates a comparison of the annual and seasonal trends for different latitudes. All data sets generally show consistent latitudinal patterns of the LSAT trends. For the annual trends, the northern middle-to-high latitudes (i.e., 50–90°N) experience the highest warming rates for all time periods. In addition, the warming trends for this region accelerated in the recent decades, especially since 1981. However, the increasing warming trend does not exist in other latitudes. Some latitudes even experience smaller LSAT trends in recent decades. For instance, the LSAT trend of 30–50°N is not significant (around 0.1°C per decade) for 1998–2017, but it is above 0.2°C per decade for 1951–2017 and nearly 0.35°C per decade for 1981–2017. This slowdown

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

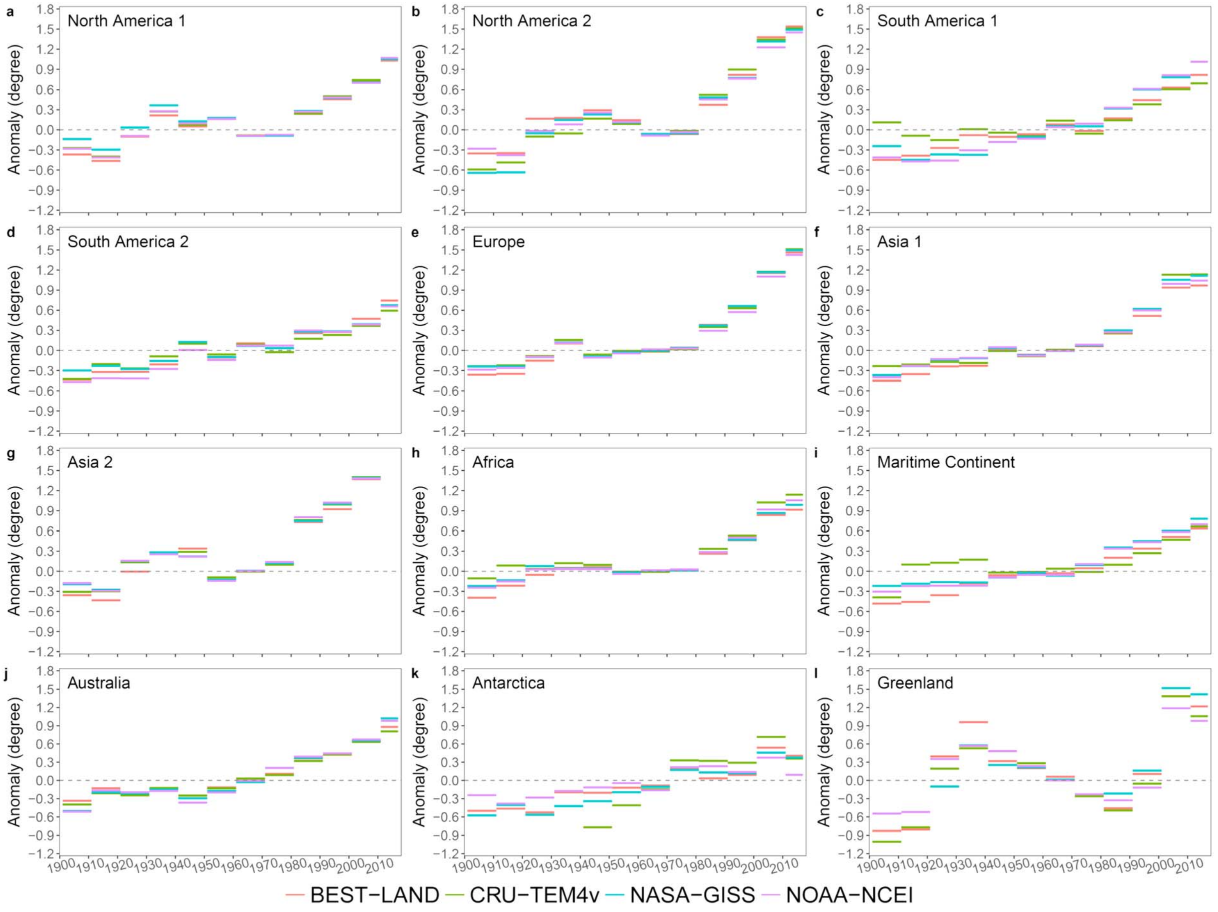

Figure 4. Same as Figure 3 but for different regions described in Table 2. BEST-LAND = Berkley Earth Surface Temperature land surface air temperature data set; CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

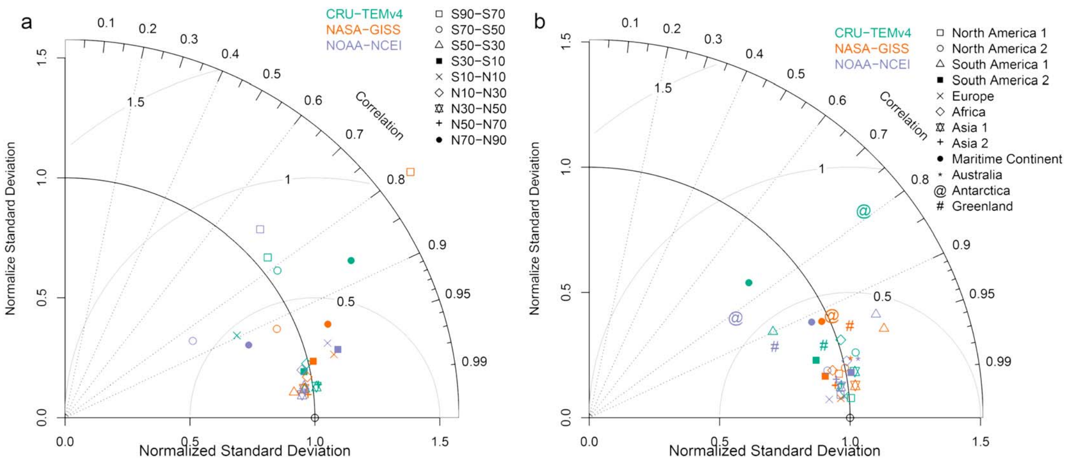

Figure 5. Normalized Taylor diagram for (a) latitudinal zones and (b) selected regions as described in Table 2 (both using Berkley Earth Surface Temperature land surface air temperature data set as the reference). CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

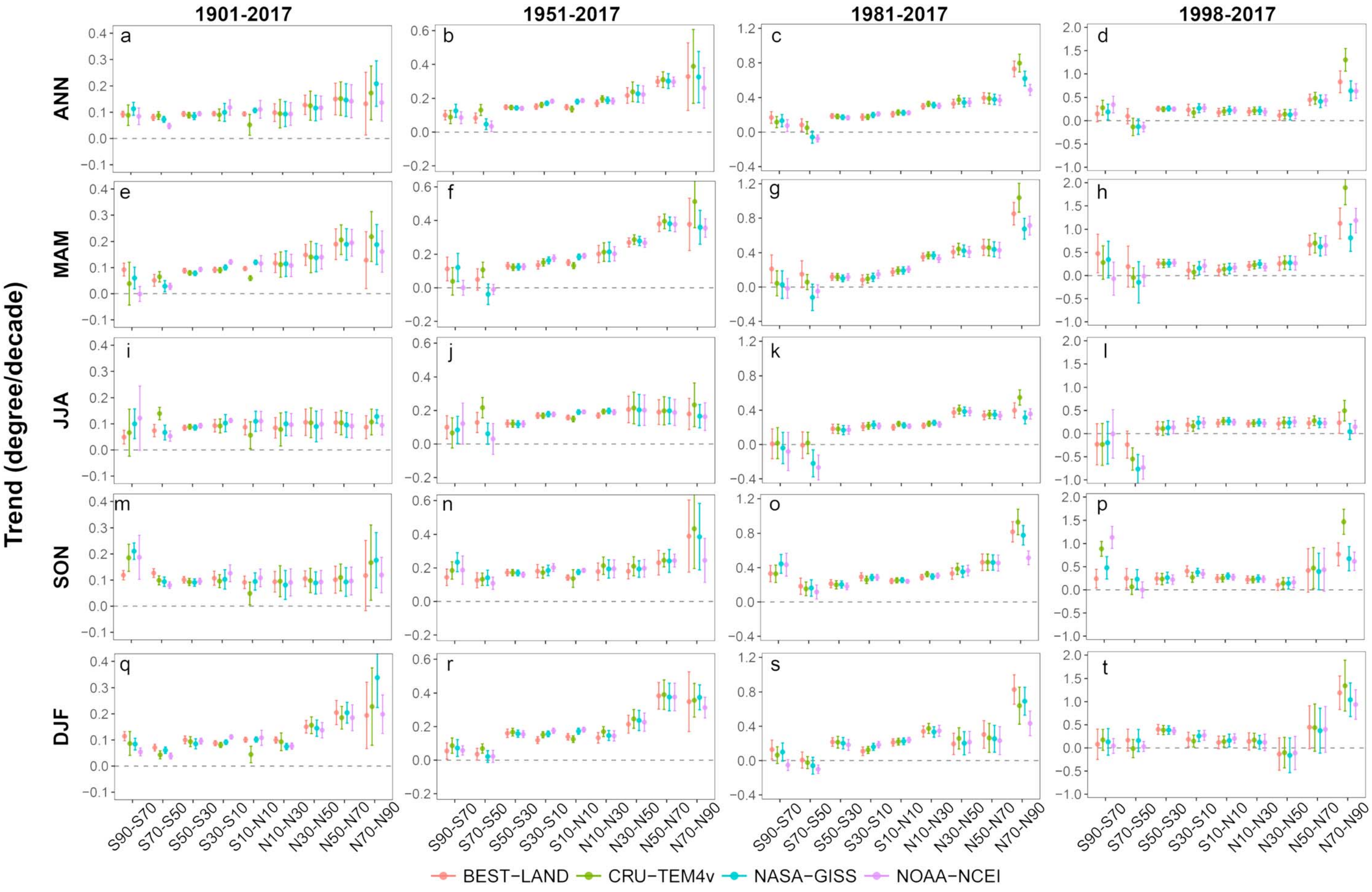

Figure 6. Comparison of the linear trends (unit: Degree per decade) of the annual and seasonal mean land surface air temperature for different latitudinal bands. The error bar around each point indicates the adjusted standard error of the linear trend estimation. The linear trends are calculated for different time periods. BEST-LAND = Berkley Earth Surface Temperature land surface air temperature data set; CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

of warming also happens for 90–50°S. The trend of other latitudes does not notably differ in the recent decades.

The high latitudes (70–90°S/N) have the largest seasonal variations with a larger magnitude of warming during the cold seasons (i.e., DJF and MAM for the NH and JJA and SON for the SH) and a smaller magnitude of warming during the warm seasons. The *warming acceleration* for the northern-high latitudes exists in almost all seasons, except in the summer time (i.e., JJA). The southern-high latitudes (i.e., 90–70°S) only shows a significant warming trend during its spring season (i.e., SON). In addition, this region exhibits the warming acceleration phenomenon with the warming rate reaching almost 1°C per decade in 1998–2017, which is much larger than 0.3°C per decade for the other time periods. The LSAT trends of 30°S–50°N during DJF and MAM for 1998–2017 are insignificant or even negative (i.e., cooling) for these latitudinal bands. However, most of them are significantly positive (i.e., warming) for the other time periods.

Despite the general agreements among the four data sets, notable differences also exist for the LSAT trends across different latitudes and time periods. For the equator (i.e., 10°S–10°N), the LSAT trends estimated from CRU-TEM4v are consistently smaller by more than 50% than those of the other data sets for the whole study period (1901–2017). The LSAT trends for the equator of CRU-TEM4v and BEST-LAND during 1951–2017 are similar but always smaller than the ones of NASA-GISS and NOAA-NCEI for all seasons. For the SH middle-to-high latitudes (i.e., 70–50°S), BEST-LAND has the highest LSAT trends followed by CRU-TEM4v and NOAA-NCEI. Meanwhile, NASA-GISS has the lowest LSAT trend for 1901–2017. The relative difference of

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the LSAT trend for this region between BEST-LAND and NASA-GISS ranges from 62% to 93% for different seasons. For DJF and MAM, the sign of the estimated LSAT trends for this region differs across data sets. This differences also occur during the time period from 1951 to 2017. Moreover, the estimated LSAT trends for the high latitudes in both hemispheres differ across data sets. CRU-TEM4v and BEST-LAND usually have larger estimated trends than NASA-GISS and NOAA-NCEI for the northern high latitude (i.e., 70–90°N) during most seasons and time periods. In the recent decades (i.e., 1981–2017 and 1998–2017), the trend differences across different data sets become more remarkable, especially for the annual trend and the seasonal trends of summer and fall (i.e., JJA and SON). The largest relative differences reach 48%, 61%, and 51% for the annual, summer, and fall LSAT trends during 1981–2017.

Using BEST-LAND as the reference, we test the significance of trend differences for different time periods and different seasons (Table S2). CRU-TEM4v has consistently significant smaller trends for both annual and seasonal mean LSATs during 1901–2017 over equator (i.e., 10°S–10°N). During 1951–2017, CRU-TEM4v shows significantly higher trends over NH (i.e., 10–70°N) for annual mean LSATS, while this pattern expands to high latitudes for fall season (i.e., SON). In the recent decades (1981–2017 and 1998–2017), CRU-TEM4v shows significant positive trend differences for northern latitude bands for most seasons and negative trend differences for southern high latitudes (i.e., 90–50°S). Meanwhile, NASA-GISS shows significant negative trend differences during 1901–2017 for most southern latitudinal bands. It also presents significant trend differences for northern latitudes during summer time (i.e., JJA). However, the significant trend differences between NASA-GISS and BEST-LAND expand to much broader regions from 1951 over the tropics and subtropics (i.e., 30°S–30°N). NOAA-NCEI shows the similar pattern with NASA-GISS but with larger magnitude of differences. In addition, during 1901–2017, NOAA-NCEI also appears to have significant trend differences during MAM season for 30°S–50°N. This trend difference analysis indicates that choice of different data sets leads to significant differences for warming trend calculation.

4.4. LSAT Warming Trend Analysis for the Selected Regions

Similar with the latitudinal band analysis, we also compare the LSAT trends estimated using different data sets of different time periods for the predefined regions (Table 2). The estimated LSAT trends generally have the highest degree of agreement over North America, Europe, and Asia, while they notably differ over South America, Africa, Maritime Continent, Australia, Antarctica, and Greenland (Figure 7). This pattern generally agrees with the spatial distribution of the ground stations commonly used for generating these data sets.

For South America, the LSAT trends estimated using NASA-GISS and NOAA-NCEI are consistently higher than those of BEST-LAND. Meanwhile, CRU-TEM4v has the lowest LSAT trends for the long-term analysis (i.e., 1901–2017 and 1951–2017). The differences are the largest over tropical South America (i.e., SA1), Greenland, and Antarctica. A large part of the tropical South America is covered by dense forests, in which setting up ground stations is difficult and leads to a large uncertainty of the observation-based data sets because of the substantial data gaps (Frenne & Verheyen, 2016).

Trends estimated from four data sets differ notably over Antarctica for both annual and seasonal mean LSATs. The most notable difference for the annual trends is during 1981–2017 when BEST-LAND suggests that Antarctica is warming at a rate of 0.2°C per decade, while other data sets all indicate that it experiences a pause of warming or even slight cooling (insignificant) during the last four decades. For the fall season over Antarctica (i.e., MAM), the LSAT trends estimated from different data sets often disagree with others in terms of whether Antarctica is warming or cooling for different time periods.

For Greenland, these four data sets have a large discrepancy for both annual and seasonal trend calculation. NASA-GISS has the highest LSAT trends for the long term (i.e., 1901–2017 and 1951–2017) for all seasons, while this pattern does not persist in recent decades (i.e., 1981–2017 and 1998–2017). For the spring time (i.e., MAM) during the whole study period (i.e., 1901–2017), NASA-GISS presents an evidence of a warming Greenland with a significant warming rate of 0.15°C per decade, while NOAA-NCEI shows no significant warming during the same periods.

Using BEST-LAND as reference, CRU-TEM4v shows significantly different trends over high-latitude North America, Europe, and South America during 1901–2017. More regions, including lower part of Asia,

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

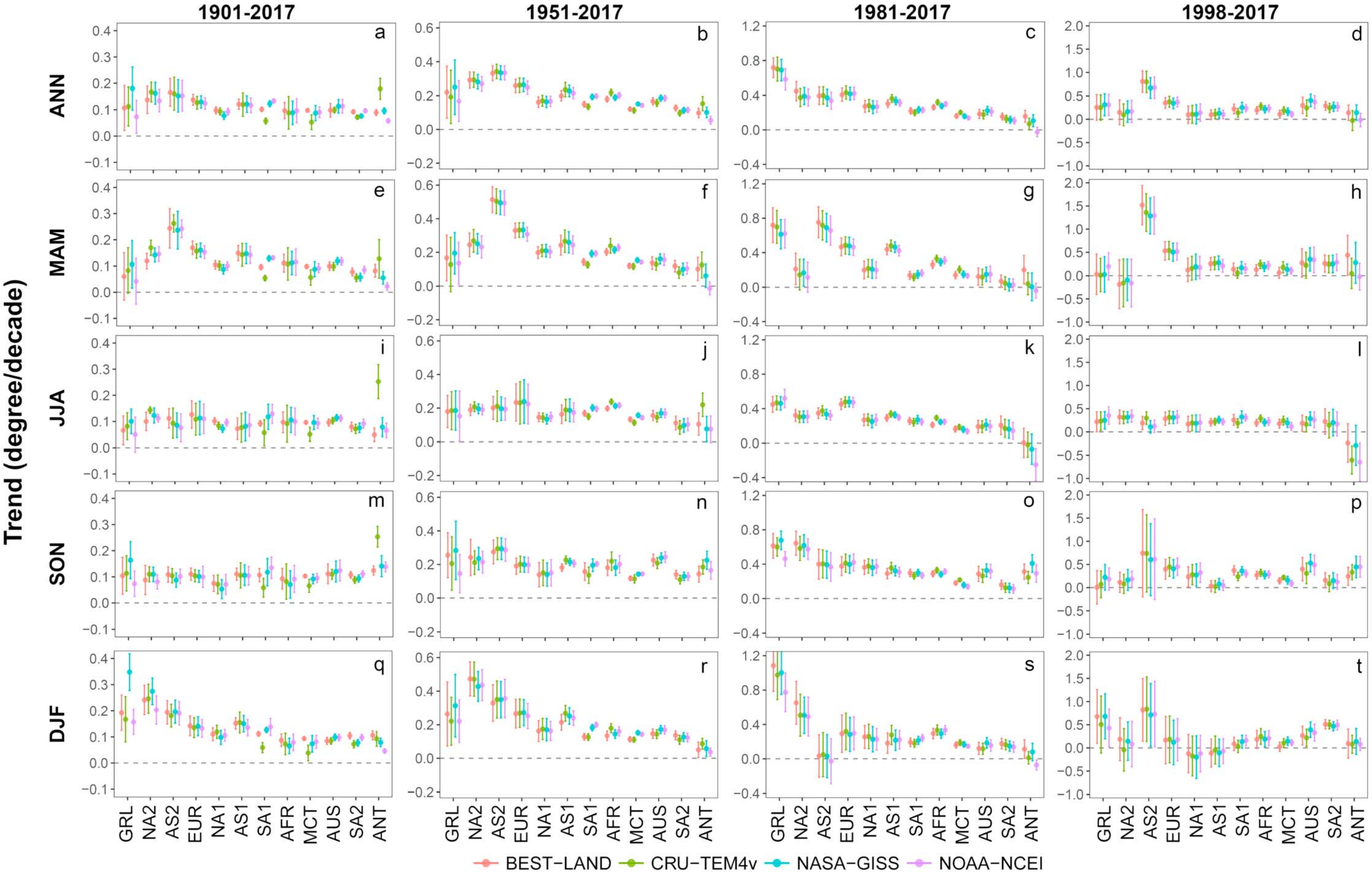

Figure 7. Same as Figure 6 but for different regions described in Table 2. BEST-LAND = Berkley Earth Surface Temperature land surface air temperature data set; CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

Africa, Maritime Continent, and Antarctica, also show significant trend differences after 1951, especially since 1981 (Table S3). Although NASA-GISS has better agreement with BEST-LAND in terms of absolute anomalies, it also appears to have significant trend differences for lower part of Asia, South America, Africa, Maritime Continent, and Antarctica mostly after 1951. Surprisingly, NASA-GISS has significant negative trend differences comparing with BEST-LAND over lower part of North America during 1901–2017, which is a region with abundant station observations. The substantial trend difference is likely the combined impact of different station data (BEST uses a much larger station data archive than CHCNMv3) and different homogenization methods. Further research is necessary to detangle contributions of individual factors. For NOAA-NCEI, it also exhibits significant trend differences with BEST-LAND over South America, Africa, high-latitude regions in Asia and North America, Europe, and Antarctica.

5. Grid Box LSAT and Trend Analysis

Existing studies rarely examine the LSAT difference across data sets at the local scale and its impact on local trend calculation. We present herein the comparison results for both the LSAT and estimated trend at the grid box scale.

5.1. Grid Box LSAT Comparison

Figure 8 demonstrates the coefficients of variations (COV) for the LSATs of the four data sets. We show the temporal evolution of the LSAT variations across data sets by separating the whole study periods into six 20-year periods (i.e., 1901–1920, 1921–1940, 1941–1960, 1961–1980, 1981–2000, and 2001–2017).

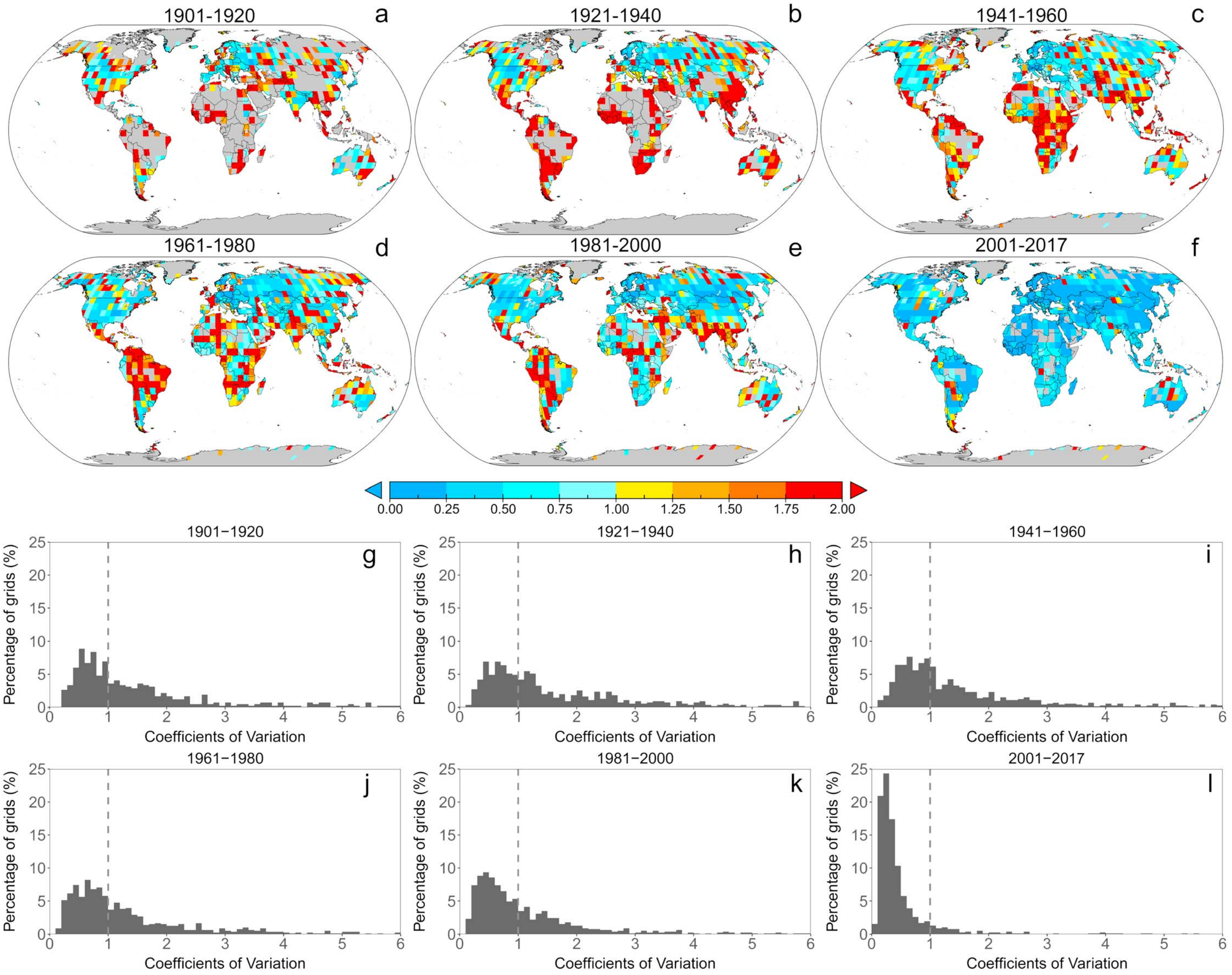

Figure 8. (a–f) Spatial pattern of the coefficients of variation of the annual mean land surface air temperature between four data sets for different time periods (i.e., 1901–1920, 1921–1940, 1941–1960, 1961–1980, 1981–2000, and 2001–2017). Only common data coverage areas are shown in the map. (g–l) Corresponding histograms of the coefficients of variation of the annual mean LSAT between the four data sets for different time periods.

Figures 8a–8f show the spatial patterns of the COV for different time periods, while Figures 8g–8l demonstrate the histograms of the COV for each time period. The COV is only calculated for the land grid boxes, where the four data sets have valid values during the time period.

The continental United States and Europe have the lowest LSAT variations across data sets for all time periods, while other parts of the world experience gradual decreases of the LSAT variations across data sets through time. The decrease of the LSAT variations for most parts of the land, including South America, Africa, and majority of Asia, happens during 1981–2000, which might have been a result of the introduction of new ground stations worldwide. However, some regions still have relatively large variations since 1981 despite more stations being used for these data sets, including the west part of South America, Sahel, Indian subcontinents, Southeast Asia as well as the west and central Australia. The variation drops substantially in the last two decades (i.e., 2001–2017) for most regions. Only a small portion of the grid boxes over central Australia, central Africa, and high latitudes still have large variations. This temporal evolution is clearly captured by the leftward shifting of the histograms across different time periods (Figures 8g–8l).

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

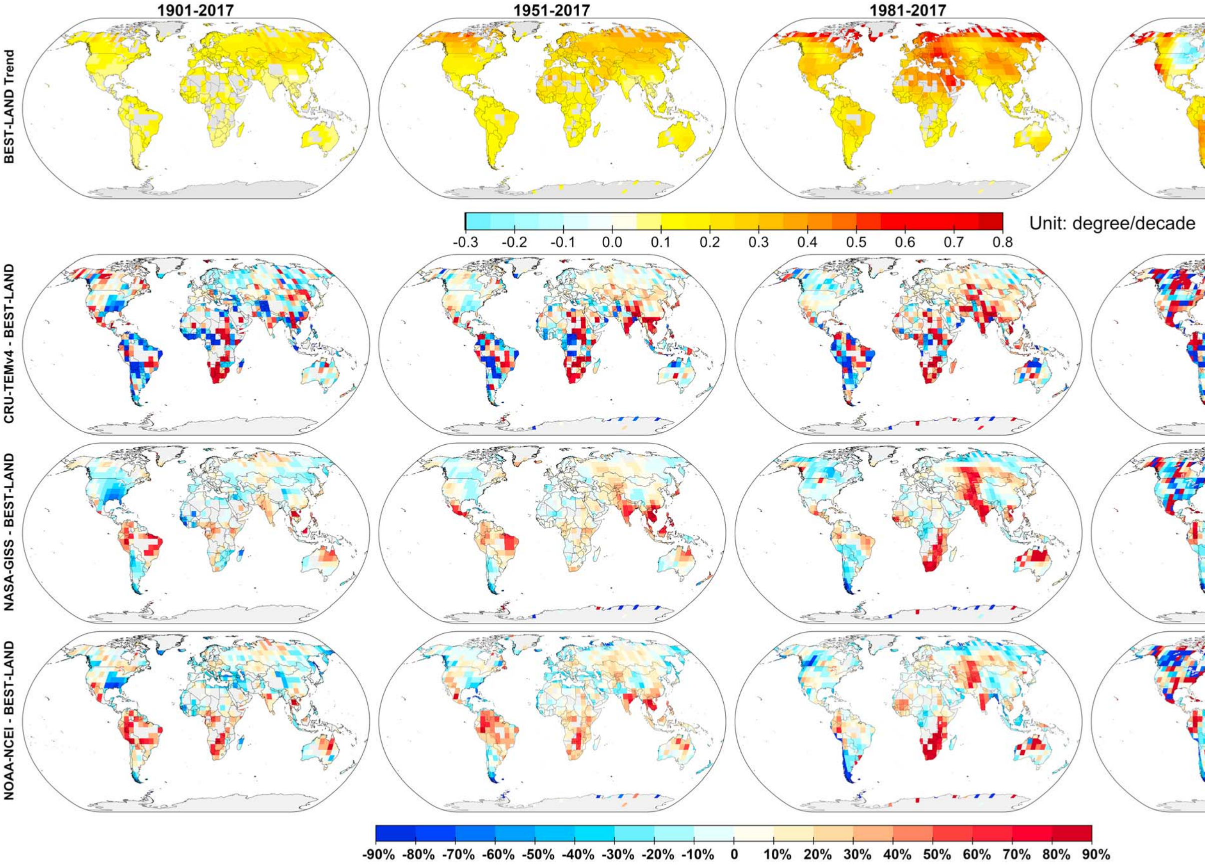

Figure 9. Upper panel: Spatial pattern of the linear trends (unit: Degrees per decade) of the annual mean land surface air temperature of BEST-LAND for 1901–2017, 1951–2017, 1981–2017, and 1998–2017. Bottom panel: Relative trend differences of the annual mean land surface air temperatures for CRU-TEM4v, NASA-GISS, and NOAA-NCEI compared to BEST-LAND over 1901–2017, 1951–2017, 1981–2017, and 1998–2017. Only common data coverage areas are shown in the maps. BEST-LAND = Berkley Earth Surface Temperature land surface air temperature data set; CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

5.2. Grid Box LSAT Trend Comparison

Figure 9 shows the spatial pattern of the annual LSAT trend for the common data coverage areas during different time periods (i.e., 1901–2017, 1951–2017, 1981–2017, and 1998–2017). BEST-LAND has the highest spatial resolution and full data coverage; hence, we use the trends estimated from BEST-LAND as the references because of the lack of *true values* of LSAT trends.

The upper panel in Figure 9 shows a clear spatial pattern of the surface warming for different time periods based on BEST-LAND. Northern middle-to-high latitudes experienced the highest rate of surface warming. The surface warming for majority of the land grid boxes has accelerated since 1951, with the most profound acceleration occurring at Europe, North Africa, northern China, Central Asia, and Siberia. However, certain regions experience smaller warming trends or even cooling trends in the recent two decades (1998–2017). These regions include northeast part of North America, southwest tip of Africa continent, central Asia, and northern China. The spatial patterns of the surface warming based on other data sets (i.e., CRU-TEM4v, NASA-GISS, and NOAA-NCEI) are similar with the one of BEST-LAND (not presented here).

The lower panel of Figure 9 shows the relative difference maps of the annual trends of CRU-TEM4v, NASA-GISS, and NOAA-NCEI compared to BEST-LAND. CRU-TEM4v shows the largest differences compared to BEST-LAND, especially over station-sparse regions, such as Africa. The relative trend difference can even reach 95% (in the central part of Africa). Other regions, such as South America, high latitudes in North America, and Asia also show large relative differences during different time periods. The large discrepancy of the LSAT trends occurs at Africa and South America.

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

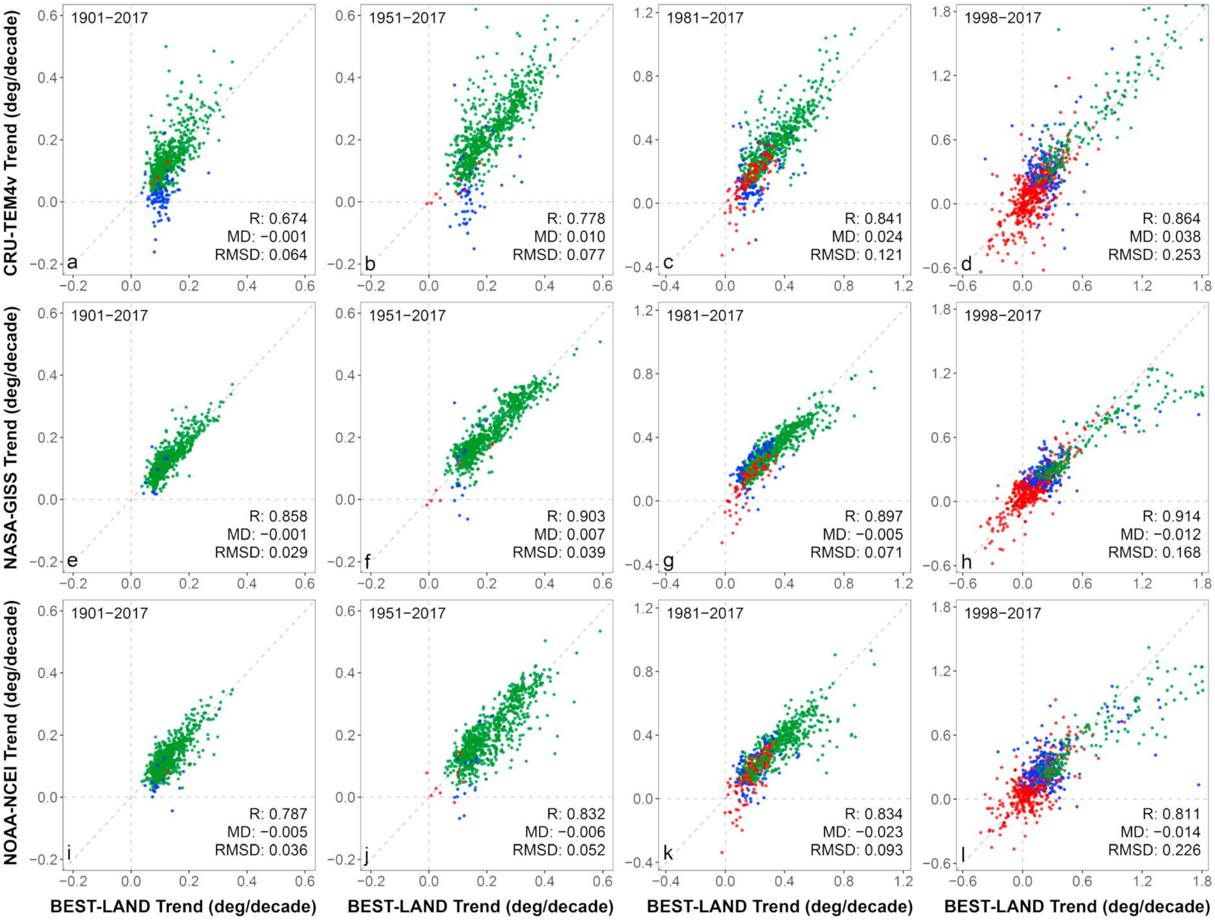

Figure 10. Scatterplots of the linear trend of the annual mean land surface air temperature for different time periods (i.e., 1901–2017, 1951–2017, 1981–2017, and 1998–2017) for CRU-TEM4v, NASA-GISS, and NOAA-NCEI compared to BEST-LAND. Red points indicate that the estimated trends are not significant for either data set; blue points represent that only one estimated trend is significant; green points indicate both estimated trends are significant. BEST-LAND = Berkley Earth Surface Temperature land surface air temperature data set; CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

For NASA-GISS and NOAA-NCEI, they both show relatively small trend differences during long-term analysis (1901–2017 and 1951–2017). But station-sparse regions like tropical South America, Africa, and north Australia demonstrate large trend differences reaching 50%. In recent decades (1981–2017), trend differences increase remarkably over central Asia, India, southern Africa continent, and north Australia (more than 80%). For the past two decades, due to relative short data length, the estimated trends show substantial differences across all data sets.

Figure 10 demonstrates the scatter plots of the annual trend comparison among different data sets for different time periods using BEST-LAND as the reference. CRU-TEM4v appears to be the most inconsistent with BEST-LAND with the lowest correlation coefficients (R), largest mean differences (MD), and largest root-mean-square differences (RMSD) for all time periods. In contrast, NASA-GISS has the highest degree of agreement with BEST-LAND for the annual trend supported by the highest R and smallest RMSD for all time periods. For some grid boxes, different data sets disagree on the sign of estimated LSAT trends. The degree of this disagreement increases in recent decades because of the large standard errors of the estimated trends caused by short data records and large interannual variability.

Table 3 presents the detailed statistics for the annual and seasonal trend comparison across data sets using BEST-LAND's estimations as the reference. CRU-TEM4v generally has the largest differences compared to BEST-LAND, while NASA-GISS has the smallest differences. For the seasonal trend comparison, all data sets seem to have a higher degree of agreement for MAM and DJF for all data sets. Despite the high

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3
Statistics of the Estimated Trend Differences (Unit: Degrees Per Decade) Among CRU-TEM4v, NASA-GISS, and NOAA-NCEI Using BEST-LAND as the Reference for 1901–2017, 1951–2017, 1981–2017, and 1998–2017

Time		CRU-TEM4v			NASA-GISS			NOAA-NCEI		
		R	MD	RMSD	*R*	MD	RMSD	*R*	MD	RMSD
1901–2017	ANN	0.674	−0.001	0.064	0.858	−0.001	0.029	0.787	−0.005	0.036
	MAM	0.781	0.000	0.069	0.923	0.000	0.033	0.872	−0.006	0.044
	JJA	0.602	−0.001	0.065	0.752	0.000	0.035	0.633	−0.001	0.042
	SON	0.689	−0.001	0.066	0.842	−0.002	0.032	0.746	−0.004	0.042
	DJF	0.761	−0.001	0.078	0.906	0.001	0.038	0.857	−0.006	0.051
1951–2017	ANN	0.778	0.010	0.077	0.903	0.007	0.039	0.832	−0.006	0.052
	MAM	0.861	0.004	0.084	0.943	0.005	0.047	0.895	−0.013	0.065
	JJA	0.610	0.008	0.082	0.768	0.004	0.047	0.660	−0.004	0.058
	SON	0.783	0.016	0.086	0.899	0.009	0.042	0.754	−0.001	0.064
	DJF	0.855	0.012	0.096	0.939	0.010	0.056	0.871	−0.005	0.082
1981–2017	ANN	0.841	0.024	0.121	0.897	−0.005	0.071	0.834	−0.023	0.093
	MAM	0.911	0.014	0.127	0.942	−0.008	0.091	0.899	−0.029	0.123
	JJA	0.772	0.027	0.127	0.839	0.003	0.089	0.796	−0.006	0.105
	SON	0.875	0.030	0.143	0.930	−0.001	0.085	0.823	−0.024	0.130
	DJF	0.838	0.020	0.162	0.928	−0.012	0.092	0.833	−0.027	0.137
1998–2017	ANN	0.864	0.038	0.253	0.914	−0.012	0.168	0.811	−0.014	0.226
	MAM	0.922	0.021	0.266	0.957	−0.015	0.216	0.895	−0.033	0.292
	JJA	0.620	0.052	0.253	0.648	−0.002	0.196	0.557	0.014	0.221
	SON	0.846	0.049	0.307	0.897	0.000	0.192	0.730	−0.011	0.285
	DJF	0.893	0.010	0.345	0.936	−0.030	0.231	0.858	−0.022	0.323

Note. *R* = correlation coefficient; MD = mean difference; RMSD = root-mean-square difference; ANN = annual; MAM = March–April–May; JJA = June–July–August; SON = September–October–November; DJF = December–January–February; BEST-LAND = Berkley Earth Surface Temperature land surface air temperature data set; CRU-TEM4v = Climate Research Unit Temperature Data Set version 4; NASA-GISS = National Aeronautics and Space Administration Goddard Institute for Space Studies; NOAA-NCEI = National Oceanic and Atmospheric Administration National Center for Environmental Information.

correlation coefficients in the recent decades, the variations of the LSAT trends increase with time, which are shown by the increasing RMSD values for both annual and seasonal LSAT trends for all data sets. The RMSD for CRU-TEM4v increases from (0.064, 0.078)°C per decade for 1901–2017 to (0.253, 0.345)°C per decade for 1998–2017. The RMSD for NASA-GISS also increases from (0.029, 0.038)°C per decade (1901–2017) to (0.068, 0.216)°C per decade (1981–2017).

6. Discussion and Conclusion

In this study, we thoroughly examine the differences of four LSAT data sets (i.e., CRU-TEM4v, BEST-LAND, NASA-GISS, and NOAA-NCEI) at different spatial and temporal scales and their potential impact on the trend calculation. The data coverage used for calculating the large-scale mean LSAT at global and hemispheric scales has a strong impact on the final time series. For the global annual mean LSAT, different data coverages introduce an LSAT anomaly difference at the magnitude of 0.15°C. This difference is even larger for different seasons (i.e., 0.4°C for DJF and MAM and 0.2°C for JJA and SON). For the hemispheric mean LSAT, the anomaly differences caused by differences in data coverages are nearly 0.6 and 0.3°C for the cold and warm seasons, respectively.

The mean LSAT differences at different latitudes are most prominent at high latitudes (e.g., 70–90°N and 50–90°S) and at the equator (i.e., 10°S–10°N). The decadal mean LSAT differences are as large as 0.6°C and 0.3° for high latitudes and the equator, respectively. These large differences lead to notable differences for the LSAT trend estimation. The relative difference of the LSAT trends for the high latitudes and the equator ranges from 35% to 60% for different seasons and time periods. Meanwhile, the LSAT differences across data sets for the southern high latitudes cause different signs of the estimated LSAT trends for the recent decades. This notable disagreement would lead to different conclusions of whether the warming of the region is accelerating or slowing down, which is essential for understanding the status of the current climate change and its impacts on the ecosystem and the society.

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

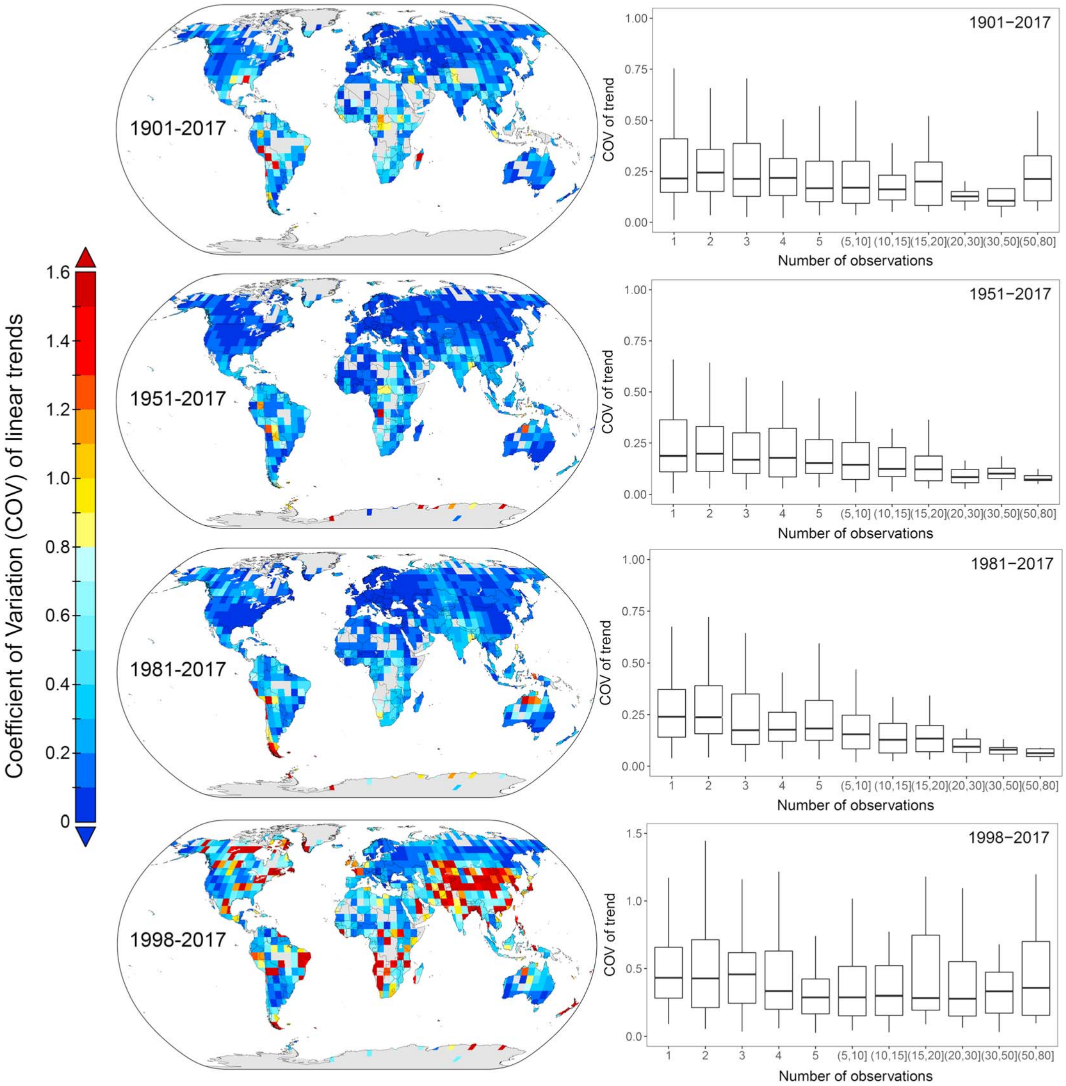

Figure 11. Left panel: Spatial patterns of the linear trend of the annual mean land surface air temperature for different time periods (i.e., 1901–2017, 1951–2017, 1981–2017, and 1998–2017) among the four data sets. Gray areas are grids, where at least two data sets do not have significant trends for that time periods. Right panel: Corresponding box plots of the coefficients of variation of the linear trend for the grid boxes grouped by the number of stations available in the grid boxes used by Climate Research Unit Temperature Data Set version 4 (CRU-TEM4v).

At the regional scale, these data sets have the highest degree of agreement over North America, Europe, and Asia but notably differ over South America, Antarctica, Africa, and Maritime Continent. This difference may be attributed to the skewed distribution of the ground station used to generate these data sets and different interpolation methods. Most stations are clustered in regions that are more developed and populated. The regional mean LSAT differences are nearly 0.4°C even after 2000. As a result, the trends estimated from different data sets differ substantially for those regions. The relative trend difference across data sets ranges from 28% to 93% over different regions and time periods. The LSAT differences across data sets for Antarctica and Greenland even cause different signs of the estimated trends, thereby leading to contrasting conclusions on these most vulnerable regions.

At local scale, our analysis shows that the regions with the largest LSAT variation across data sets are grid boxes over South America, Africa, Indian subcontinent, central and northern Australia, southeast Asia, and Siberia, which is consistent with the regional-scale analysis. The LSAT variation across data sets decreases with time, with the lowest variation among data sets seen after 2000. For the trend analysis, the largest variation of the trends often occurs at the grid boxes with the largest LSAT variation across data sets. The relative

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

difference of trends estimated from different data sets can reach nearly 90% for different regions and time periods. CRU-TEM4v generally appears to have the largest grid box scale differences, while NASA-GISS has the smallest differences compared to BEST-LAND. The uncertainty of the LSAT trend estimation caused by the data set differences (i.e., RMSD) ranges from 0.035 to 0.086°C per decade for the long-term trend (i.e., 1901–2017) to 0.097–0.305°C per decade for recent decades (i.e., 1981–2017).

Based on the previous comparison across different scales, the data set difference strongly depends on the station data availability at any scales. Indeed, the main challenge of generating observation-based LSAT data sets is to obtain homogenized station observations and ingest them into the final product. The stations used for each data set vary significantly because of different quality control procedures and raw data sources (Table 2). Examining the variation of the number of stations used by each data set over different regions and grid boxes would further advance the understanding of the data set variation. CRU provides the number of stations used in each grid box to generate CRU-TEM4v, which is used herein as a proxy since CRU uses least amount of stations in the final data set.

Figure 11 demonstrates the coefficients of variation (COV) of the grid box LSAT trend for different time periods and its relationship with the number of stations available in each grid box. The box plots show that the spread and the mean value of the COV when there are less than five stations in the entire 5° × 5° grid box are significantly larger than the grid boxes with more stations. This large variation can be expected because using a very limited number of stations to capture the full dynamics of LSAT across such a large area (e.g., approximately 500 km × 500 km at low latitudes), especially over regions with complex topography or heterogeneous landscapes, is very challenging (Pepin et al., 2015). Additionally, the within-grid-box distribution can also contribute to this variation since interpolation methods used by BEST-LAND and NASA-GISS tend to give less weights to the clustered stations while they give more weights to the isolated stations in the grid box.

This limitation of the observation-based data sets needs to be addressed to increase the confidence of climate change studies using these data sets. One way to address this issue is to improve the design and implementation of the global station network. Different international initiatives have already started the process of improving the density and the quality of station measurements, such as the International Surface Temperature Initiative (Rennie et al., 2014) and the new network implementation plan described by the Global Climate Observing System (Thorne, Allan, et al., 2017; World Meteorological Organization, 2016).

Although continuing this improvement is critical and necessary, doing so is time and resource consuming. Moreover, the improvement would mostly benefit the data set for the future, which cannot directly reduce the variations across data sets for the past. In contrast, remote sensing data are in a unique niche to provide nearly spatial-complete information over land surface. Several off-the-shelf global surface temperature products at various spatial resolutions for the recent decades (since the 1980s) are currently available, including both land surface temperature and air temperature profiles. Remotely sensed products suffer from their own limitations, such as the observation time change across satellites and biases due to cloud contamination. On the other hand, atmospheric reanalysis data have also been very popular in various applications. Despite its known uncertainties, reanalysis data provide valuable complete global temperature data at various resolutions. With appropriate statistical methods, combining global station network observations, reanalysis temperature data, and remotely sensed temperature products to generate a spatial-complete LSAT data is possible. Efforts in this aspect are already ongoing, which could significantly benefit climate change studies requiring LSAT data sets (Merchant et al., 2013; Thorne, Madonna, et al., 2017).

Acknowledgments

Yuhan Rao is supported by the Chinese Scholarship Council. This study was also partially supported by the NOAA JPSS program and NOAA grant NA14NES4320003 (Cooperative Institute for Climate and Satellites-CICS) at the University of Maryland/ESSIC. We want to thank Kevin E. Trenberth and Yongxin Zhang for providing helpful comments on the intercomparison analysis. The BEST-LAND data are downloaded from Berkeley Earth (http://berkeleyearth.org/data/). The NASA-GISS temperature data are downloaded from Goddard Institute for Space Studies (https://data.giss.nasa.gov/gistemp/). The NOAA-NCEI data are downloaded from National Centers for Environmental Information (https://www.ncdc.noaa.gov/data-access/marineocean-data/noaa-global-surface-temperature-noaaglobaltemp). CRU-TEM4v is accessed through Climate Research Unit (http://www.cru.uea.ac.uk/data). All data were accessed on 15 March 2018. We also want to thank three anonymous reviewers who provided suggestions for the earlier version of our manuscript.

References

Bekryaev, R. V., Polyakov, I. V., & Alexeev, V. A. (2010). Role of polar amplification in long-term surface air temperature variations and modern Arctic warming. *Journal of Climate*, *23*(14), 3888–3906. https://doi.org/10.1175/2010JCLI3297.1

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

Cowtan, K., & Way, R. G. (2014). Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1935–1944. https://doi.org/10.1002/qj.2297

Editorial (2017). Expanding research views. *Nature Climate Change*, *7*(4), 229. https://doi.org/10.1038/nclimate3270

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

Frenne, P. D., & Verheyen, K. (2016). Weather stations lack forest data. *Science*, *351*(6270), 234–234. https://doi.org/10.1126/science.351.6270.234-a

Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, *48*, RG4004. https://doi.org/10.1029/2010RG000345

Hartmann, D. L., Klein Tank, A. M. G., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y., et al. (2013). Observations: Atmosphere and surface. In T. F. Stocker, et al. (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 159–254). Cambridge, United Kingdom and New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.008

Hausfather, Z., Cowtan, K., Clarke, D. C., Jacobs, P., Richardson, M., & Rohde, R. (2017). Assessing recent warming using instrumentally homogeneous sea surface temperature records. *Science Advances*, *3*(1), e1601207. https://doi.org/10.1126/sciadv.1601207

Hegerl, G. C., Black, E., Allan, R. P., Ingram, W. J., Polson, D., Trenberth, K. E., et al. (2014). Challenges in quantifying changes in the global water cycle. *Bulletin of the American Meteorological Society*, *96*(7), 1097–1115. https://doi.org/10.1175/BAMS-D-13-00212.1

Jones, P. (2016). The reliability of global and hemispheric surface temperature records. *Advances in Atmospheric Sciences*, *33*(3), 269–282. https://doi.org/10.1007/s00376-015-5194-4

Jones, P., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., & Morice, C. P. (2012). Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research*, *117*, D05127. https://doi.org/10.1029/2011JD017139

Lee, J., & Lund, R. (2004). Revisiting simple linear regression with autocorrelated errors. *Biometrika*, *91*(1), 240–245. https://doi.org/10.1093/biomet/91.1.240

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, *29*(7), 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

Menne, M. J., & Williams, C. N. (2009). Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, *22*(7), 1700–1717. https://doi.org/10.1175/2008JCLI2263.1

Menne, M. J., Williams, C. N., & Palecki, M. A. (2010). On the reliability of the U.S. surface temperature record. *Journal of Geophysical Research*, *115*, D11108. https://doi.org/10.1029/2009JD013094

Merchant, C. J., Matthiesen, S., Rayner, N. A., Remedios, J. J., Jones, P. D., Olesen, F., et al. (2013). The surface temperatures of Earth: Steps towards integrated understanding of variability and change. *Geoscientific Instrumentation, Methods and Data Systems*, *2*(2), 305–321. https://doi.org/10.5194/gi-2-305-2013

Muller, R. A., Rohde, R., Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., et al. (2013). A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinformatics & Geostatistics: An Overview*, *01*(01). https://doi.org/10.4172/2327-4581.1000101

Pepin, N., Bradley, R. S., Diaz, H. F., Baraër, M., Caceres, E. B., Forsythe, N., et al. (2015). Elevation-dependent warming in mountain regions of the world. *Nature Climate Change*, *5*(5), 424–430. https://doi.org/10.1038/nclimate2563

Rennie, J. J., Lawrimore, J. H., Gleason, B. E., Thorne, P. W., Morice, C. P., Menne, M. J., et al. (2014). The international surface temperature initiative global land surface databank: Monthly temperature data release description and methods. *Geoscience Data Journal*, *1*(2), 75–102. https://doi.org/10.1002/gdj3.8

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., et al. (2013). Berkeley Earth temperature averaging process. *Geoinformatics & Geostatistics: An Overview*, *01*(02). https://doi.org/10.4172/2327-4581.1000103

Santer, B. D., Wigley, T. M. L., Boyle, J. S., Gaffen, D. J., Hnilo, J. J., Nychka, D., et al. (2000). Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. *Journal of Geophysical Research*, *105*(D6), 7337–7356. https://doi.org/10.1029/1999JD901105

Smith, T. M., & Reynolds, R. W. (2005). A global merged land–air–sea surface temperature reconstruction based on historical observations (1880–1997). *Journal of Climate*, *18*(12), 2021–2036. https://doi.org/10.1175/JCLI3362.1

Smith, T. M., Reynolds, R. W., Peterson, T. C., & Lawrimore, J. (2008). Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *Journal of Climate*, *21*(10), 2283–2296. https://doi.org/10.1175/2007JCLI2100.1

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*(D7), 7183–7192.

Thorne, P. W., Allan, R. J., Ashcroft, L., Brohan, P., Dunn, R. J. H., Menne, M. J., et al. (2017). Towards an integrated set of surface meteorological observations for climate science and applications. *Bulletin of the American Meteorological Society*, *98*(12), 2689–2702. https://doi.org/10.1175/BAMS-D-16-0165.1

Thorne, P. W., Donat, M. G., Dunn, R. J. H., Williams, C. N., Alexander, L. V., Caesar, J., et al. (2016). Reassessing changes in diurnal temperature range: Intercomparison and evaluation of existing global data set estimates. *Journal of Geophysical Research: Atmospheres*, *121*, 5138–5158. https://doi.org/10.1002/2015JD024584

Thorne, P. W., Madonna, F., Schulz, J., Oakley, T., Ingleby, B., Rosoldi, M., et al. (2017). Making better sense of the mosaic of environmental measurement networks: A system-of-systems approach and quantitative assessment. *Geoscientific Instrumentation, Methods and Data Systems*, *6*(2), 453–472. https://doi.org/10.5194/gi-6-453-2017

van den Dool, H. M., Saha, S., & Johansson, Å. (2000). Empirical orthogonal teleconnections. *Journal of Climate*, *13*(8), 1421–1435. https://doi.org/10.1175/1520-0442(2000)013<1421:EOT>2.0.CO;2

Vose, R. S., Arndt, D., Banzon, V. F., Easterling, D. R., Gleason, B., Huang, B., et al. (2012). NOAA's merged land–ocean surface temperature analysis. *Bulletin of the American Meteorological Society*, *93*(11), 1677–1685. https://doi.org/10.1175/BAMS-D-11-00241.1

Vose, R. S., Wuertz, D., Peterson, T. C., & Jones, P. D. (2005). An intercomparison of trends in surface air temperature analyses at the global, hemispheric, and grid-box scale. *Geophysical Research Letters*, *32*, L18718. https://doi.org/10.1029/2005GL023502

World Meteorological Organization (2016). The global observing system for climate: Implementation needs. World Meteorological Organization. Retrieved from https://library.wmo.int/opac/doc_num.php?explnum_id=3417

21698996, 2018, 11, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028355 by Noaa Department Of Commerce, Wiley Online Library on [19/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License