

Uncertainty Estimates for Sea Surface Temperature and Land Surface Air Temperature in NOAA GlobalTemp Version 5

BOYIN HUANG,^a MATTHEW J. MENNE,^a TIM BOYER,^a ERIC FREEMAN,^b BYRON E. GLEASON,^a JAY H. LAWRIKORE,^a CHUNYING LIU,^b J. JARED RENNIE,^c CARL J. SCHRECK III,^c FENGYING SUN,^b RUSSELL VOSE,^a CLAUDE N. WILLIAMS,^a XUNGANG YIN,^b AND HUAI-MIN ZHANG^a

^a NOAA/National Centers for Environmental Information (NCEI), Asheville, North Carolina

^b Riverside inc. (government contractor at NOAA/NCEI), Asheville, North Carolina

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

(Manuscript received 29 May 2019, in final form 25 October 2019)

ABSTRACT

This analysis estimates uncertainty in the NOAA global surface temperature (GST) version 5 (NOAAGlobalTemp v5) product, which consists of sea surface temperature (SST) from the Extended Reconstructed SST version 5 (ERSSTv5) and land surface air temperature (LSAT) from the Global Historical Climatology Network monthly version 4 (GHCNm v4). Total uncertainty in SST and LSAT consists of parametric and reconstruction uncertainties. The parametric uncertainty represents the dependence of SST/LSAT reconstructions on selecting 28 (6) internal parameters of SST (LSAT), and is estimated by a 1000-member ensemble from 1854 to 2016. The reconstruction uncertainty represents the residual error of using a limited number of 140 (65) modes for SST (LSAT). Uncertainty is quantified at the global scale as well as the local grid scale. Uncertainties in SST and LSAT at the local grid scale are larger in the earlier period (1880s–1910s) and during the two world wars due to sparse observations, then decrease in the modern period (1950s–2010s) due to increased data coverage. Uncertainties in SST and LSAT at the global scale are much smaller than those at the local grid scale due to error cancellations by averaging. Uncertainties are smaller in SST than in LSAT due to smaller SST variabilities. Comparisons show that GST and its uncertainty in NOAAGlobalTemp v5 are comparable to those in other internationally recognized GST products. The differences between NOAAGlobalTemp v5 and other GST products are within their uncertainties at the 95% confidence level.


1. Introduction

The analysis of global surface temperature (GST) is generally based on in situ measurements of land surface air temperature (LSAT) and sea surface temperature (SST), although satellite-based SST observations may also be included (Rayner et al. 2003). LSAT has been measured by meteorological stations since the late 1600s, and SST has been measured by commercial ships since the early 1700s and by moored and drifting buoy floats since the 1970s (Lawrimore et al. 2011; Kennedy et al. 2011a,b; Freeman et al. 2017). GST is an essential indicator of climate change that has been used

for climate assessment and monitoring (IPCC 2013; USGCRP 2017; Blunden et al. 2018).

There are two notable difficulties in the calculation of GST. First, temperature measurements do not completely cover Earth's surface, particularly near the poles and before the 1950s. Measurement coverage is also better in the Northern Hemisphere (NH) than the Southern Hemisphere (SH) over both the land and the oceans (Menne et al. 2018; Huang et al. 2019). Second, observations may be biased due to changes in thermometers, measuring methods, conventions, and relocations of meteorological stations or changes in the surrounding environment (Menne and Williams 2009; Menne et al. 2012; Kent et al. 2017).

To overcome the first difficulty, different methods have been used to interpolate large-scale variations to the regions without measurements. These methods include empirical orthogonal functions (EOFs), empirical

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

Corresponding author: Boyin Huang, boyin.huang@noaa.gov

orthogonal teleconnection functions (EOTs), kriging or Gaussian regression, and temperature correlations with neighboring measurements (van den Dool et al. 2000; Rayner et al. 2003; Smith and Reynolds 2003, 2004; Hansen et al. 2010; Cowtan and Way 2014).

To overcome the second difficulty, different schemes have been used to correct biases in historical measurements. These schemes include the homogenization of LSAT measurements, the bucket model simulating the heat loss of seawater sampling used for SST measurements, and the use of nighttime marine air temperature (NMAT) as a reference for SST measurements (Folland and Parker 1995; Smith and Reynolds 2003, 2004, 2005; Lawrimore et al. 2011; Kennedy et al. 2011a,b, 2019; Huang et al. 2015, 2017).

Using these methods for interpolation and bias correction, several GST products have been produced and widely used for global climate assessment and monitoring. These products include NOAA global surface temperature (NOAAGlobalTemp) v4 and v5 (Vose et al. 2012; Zhang et al. 2019), the global surface temperature v4 (HadCRUT4) generated by the Met Office Hadley Centre and the Climatic Research Unit (CRU) at the University of East Anglia (Morice et al. 2012), Goddard Institute for Space Studies (GISS) Surface Temperature (GISTEMP; Hansen et al. 2010), Berkeley Earth Surface Temperature (BEST; Rohde et al. 2013a,b), and Japan Meteorological Agency (JMA) global surface temperature (Ishihara 2006).

The evolution of globally averaged GSTs described in those studies is qualitatively similar for periods since the 1880s even though different methods are used in data interpolation and bias correction. For example, all analyses indicated that GST experienced a cooling over the 1880s–1910s, a warming over the 1910s to 1940s, a slight cooling over the 1940s to 1970s, and an enhanced warming over the 1970s to 2010s. Note that in the 1940s there was a switch in the source of available observations and a corresponding rapid shift in observational bias. Quantitatively, there are some differences in trend magnitude in the above periods among these products, particularly for the period 1998–2012 (IPCC 2013; Karl et al. 2015; Fyfe et al. 2016; Lewandowsky et al. 2016; Medhaug et al. 2017; Rahmstorf et al. 2017), which led to questions regarding a change, slowdown, or cessation of the general warming trend observed since the 1970s.

The debates regarding recent GST warming trends are partly caused by the selection of starting and ending year of the hiatus (Medhaug et al. 2017). However, the uncertainty in the GST products may also impact the significance of the warming trends. For example, the GST difference between NOAAGlobalTemp v5 and HadCRUT4 is mostly associated with differences in SSTs that are associated with the

uncertainty of SST bias correction (Huang et al. 2015). Therefore, it is important to quantify the GST uncertainty when a GST product is generated.

The focus of this paper is to estimate the uncertainty in NOAAGlobalTemp v5 and to compare its uncertainty with other products. Total uncertainty in NOAAGlobalTemp v5 consists of parametric and reconstruction uncertainties. The reconstruction uncertainty represents the residual errors due to using a limited number of empirical orthogonal teleconnection (van den Dool et al. 2000) modes in SST and LSAT reconstructions, and how well the retained EOTs span the variations at any given time. The parametric uncertainty represents the dependence of SST and LSAT reconstructions on randomly selecting the optional values of the internal parameters of SST and LSAT.

The rest of the paper is arranged as follows. The datasets used for uncertainty estimation and comparison are described in section 2. The methodology and estimation of SST and LSAT uncertainties are described in sections 3 and 4 respectively. The GST uncertainty in NOAAGlobalTemp v5 is quantified based on SST and LSAT uncertainties and compared with the GST uncertainties in the other products in section 5. A summary and conclusions are given in section 6.

2. Datasets used for uncertainty estimation and comparison

a. Data used for uncertainty estimation

The monthly $2^\circ \times 2^\circ$ Optimum Interpolation SST (OISST) data (Table 1) are used to train SST EOTs (Table 2, 23rd row) and used as pseudo-observations to derive SST reconstruction uncertainty (section 3d). The monthly OISST is derived from the weekly $1^\circ \times 1^\circ$ OISST (wOISST) from 1982 to 2011 (Reynolds et al. 2002), which is consistent with those used in the previous versions of ERSST (v4 and v3b). To better represent the interannual variations of SST, different SST data combinations are selected to train the EOTs as described in parameter 23 in Table 2 and appendix A. The weekly data are first interpolated to the daily data time scale, and then the daily data on $1^\circ \times 1^\circ$ grids are averaged to monthly data on $2^\circ \times 2^\circ$ grids. The OISST data are observation-based estimates that are methodologically independent of ERSSTv5. Both OISST and ERSSTv5 include in situ data. The OISST includes satellite-derived data while ERSSTv5 does not.

The monthly $5^\circ \times 5^\circ$ LSAT data, derived from the European Centre for Medium-Range Weather Forecasts (ECMWF) interim reanalysis (ERA-Interim; Table 1, second row) from 1982 to 2011 (Dee et al. 2011),

TABLE 1. Datasets used to assess uncertainties of ERSSTv5, LSAT, and NOAAGlobalTemp reconstructions.

SST products	Spatial resolution	Temporal resolution	Data ingest	Analysis method	External forcing
wOISST	$1^\circ \times 1^\circ$ Global	Weekly 1982–2013	In situ SST, satellite-based SST	Optimum interpolation	N/A
ERA-Interim	$0.75^\circ \times 0.75^\circ$	Daily	Various atmospheric observations	Assimilation	SST from HadISST
GFDL-ESM2G	$1^\circ \times 0.9^\circ$ Global	Daily 1861–2005	N/A	Coupled model simulation	Greenhouse gases, trace gases, aerosols, ozone, land use
HadGEM2-AO	$1^\circ \times 0.8^\circ$ Global	Monthly 1860–2005	N/A	Coupled model simulation	Greenhouse gases, aerosols
CanESM2	$2.8^\circ \times 2.8^\circ$	1850–2005	N/A	Coupled model simulation	Greenhouse gases, aerosols, cloud microphysics

are used to train LSAT EOTs. Original resolutions of ERA-Interim are daily, approximately 0.75° in longitude and latitude, and 60 levels. The air temperature at the lowest level (2 m) over the land is used as LSAT. ERA-Interim is an observation-based reanalysis forced by the SST from HadISST1 (Rayner et al. 2003) and the Operational SST and Sea Ice Analysis (OSTIA) system (Stark et al. 2007; Donlon et al. 2015) as a low boundary condition, which is methodologically independent of the LSAT from NOAAGlobalTemp.

LSATs from coupled model simulations are used as pseudo-observations to assess reconstruction uncertainty (section 4d). The model simulations are independent and provide spatially complete analyses. These model simulations with different resolutions are box-averaged to the monthly $5^\circ \times 5^\circ$ grids of NOAAGlobalTemp and used to estimate the reconstruction uncertainty. To assure that the estimated reconstruction uncertainty is not sensitive to the selection of model simulations, three model simulations are tested (Table 1):

- 1) Geophysical Fluid Dynamics Laboratory (GFDL) Earth System Model version 2G (ESM2G; Dunne et al. 2012). This coupled model has a resolution of approximately 2.5° in longitude, near 2° in latitude, and daily from 1861 to 2005.
- 2) Met Office (UKMO) Hadley Centre Global Environmental Model version 2-AO (HadGEM2-AO; Collins et al. 2008). This coupled model has a resolution of 1.9° in longitude, near 1.3° in latitude, and daily from 1860 to 2005.
- 3) Second Generation Canadian Earth System Model (CanESM2; Arora et al. 2011). This coupled model has a resolution of approximately 2.8° in longitude and latitude, and daily from 1850 to 2005.

The reconstruction uncertainty derived using HadGEM2-AO is eventually added with the parametric uncertainty to estimate the total uncertainty of LSAT in section 5b.

b. Data used for comparisons

GST and its uncertainty in HadCRUT4 (Morice et al. 2012), BEST (Rohde et al. 2013a,b), and GISTEMP (Hansen et al. 2010; Lenssen et al. 2019) are compared with NOAAGlobalTemp v5. The GST anomaly in HadCRUT4 is on monthly $5^\circ \times 5^\circ$ grids from 1850 to 2017, which is relative to its climatological mean over 1961–90 and is rescaled here to 1971–2000 for comparison purposes. HadCRUT4 combines LSAT from CRUTEM4 (Jones et al. 2012) and SST from HadSST3 (Kennedy et al. 2011a,b). In HadCRUT4, the total uncertainty of local GST consists of uncorrelated and supplemental uncertainties, and the total uncertainty of globally averaged GST consists of bias, sampling, and coverage uncertainties. For comparison purposes, the spread (one standard deviation or 1σ) from the 100-member ensemble of HadCRUT4 is calculated as an uncertainty measure and compared with that in NOAAGlobalTemp v5 (section 5c).

The GST anomaly in BEST is on monthly $1^\circ \times 1^\circ$ grids from 1850 to 2017. The anomaly is relative to the 1951–80 climatological mean, which is rescaled to 1971–2000 and then box-averaged to $5^\circ \times 5^\circ$ grids. BEST uses LSAT from the Global Historical Climatology Network–monthly (GHCNm) v3 (Lawrimore et al. 2011) and SST from HadSST3 (Kennedy et al. 2011a,b). The LSAT observations are homogenized using the “scalpel” method and averaged and interpolated using “Berkeley average” and kriging (or Gaussian regression) algorithms. Overall, the GST in BEST is closest to HadCRUT4 because the SST component of BEST is the same as that in HadCRUT4, although their difference is notable in the recent decade and in the nineteenth century. The uncertainty of BEST includes observational biases and undersampling effects.

The GST anomaly in GISTEMP (Hansen et al. 2010) is on monthly $2^\circ \times 2^\circ$ grids from 1880 to 2017. The anomaly is relative to a climatological mean over 1951–80, which is rescaled to 1971–2000 then

TABLE 2. ERSSTv5 parameters and their operational and alternative options. In parameter number 23, the “even years” are 1982, 1984, . . . , 2010; and the “odd” years are 1983, 1985, . . . , 2011. These parameters are explained in [appendix A](#).

Parameter	Operational option	Alternative options
1. First guess	Unadjusted ERSSTv4	Unadjusted and adjusted ERSSTv4
2. SST STD for QC	OISST v2 (1982–2011)	COADS (1950–79); OISST v2
3. Minimum SST STD	1.0°C	0.5°, 1.0°, 1.5°C
4. Maximum SST STD	4.5°C	3.5°, 4.5°, 5.5°C
5. SST STD multiplier	4.5	3.5, 4.5, 5.5
6. SST observation random error	0.0°C	1.3°C for ship SSTs and 0.5°C for buoy SSTs
7. Ship SST error	1.3°C	1.2°, 1.3°, 1.4°C
8. Buoy SST error	0.5°C	0.4°, 0.5°, 0.6°C
9. Argo SST error	0.5°C	0.4°, 0.5°, 0.6°C
10. SSTA calculation	In situ basis	Gridbox basis; in situ basis
11. NMAT for SST bias	HadNMAT2	UKMO NMAT; HadNMAT2; HadNMAT2 in three latitudinal belts: 90°–30°S, 30°S–30°N, 30°–90°N; HadNMAT2 in 25° × 25° running domain
12. Ship SST bias smoothing	Lowess $f = 0.10$	Annual; lowess $f = 0.10$; linear; lowess–linear
13. Ship SST bias readjustment based on buoy SST	0.077°C	0.062°, 0.077°, 0.092°C
14. Argo SST adjustment based on buoy SST	0.03°C	0°, 0.03°, 0.06°C
15. Buoy SST weighting	6.8	5.8, 6.8, 7.8
16. Argo SST weighting	6.8	5.8, 6.8, 7.8
17. Max number of observations	10	5, 10, 15
18. Min number of months for annual average	2	1, 2, 3
19. Min ratio of superobs	0.03	0.02, 0.03, 0.04
20. Min number of years for LF filter	2 yr	1, 2, 3 yr
21. LF filter period	15 yr	11, 15, 19 yr
22. HF filter period	3 month	1, 3 month
23. EOT training periods and spatial scales	1982–2011 $(L_x, L_y) = (5000, 3000)$ km	1982–2011, $(L_x, L_y) = (5000, 3000)$ km; 1982–2011, $(L_x, L_y) = (6000, 4000)$ km; 1982–2011, $(L_x, L_y) = (4000, 2000)$ km; 1982–2005, $(L_x, L_y) = (5000, 3000)$ km; 1988–2011, $(L_x, L_y) = (5000, 3000)$ km; Even years from 1982 to 2012, $(L_x, L_y) = (5000, 3000)$ km; Odd years from 1983 to 2013, $(L_x, L_y) = (5000, 3000)$ km
24. EOT weighting	$W = N/(N + \xi^2)\cos(\varphi)$	$W = \cos(\varphi)$; $W = N/(N + \xi^2)\cos(\varphi)$
25. EOT acceptance value	0.10	0.05, 0.10, 0.20
26. Ice concentration factor	1.0	0.9, 1.0, 1.1
27. Min ice for SST adjustment	0.6	0.5, 0.6, 0.7
28. Max ice for SST adjustment	0.9	0.8, 0.9, 1.0

box-averaged to $5^\circ \times 5^\circ$ grids. GISTEMP combines LSAT from GHCNm v3 ([Lawrimore et al. 2011](#)) with SST from ERSSTv5 ([Huang et al. 2017](#)). Regions without observations over the land are filled by the weighted average with weights decaying linearly to zero at 1200 km. Overall, the GST in GISTEMP is closest to NOAA GlobalTemp v5 because the SST component of GISTEMP is the same as that of NOAA GlobalTemp v5.

All uncertainties are quantified and compared using 1σ threshold except for the uncertainty interval (1.96σ) for a temperature trend at the 95% confidence.

3. SST and its uncertainty

a. ERSSTv5, its internal 28 parameters and 1000-member ensemble

SSTs in ERSSTv5 are produced on monthly $2^\circ \times 2^\circ$ grids from 1854 to 2017 ([Huang et al. 2017](#)) and used as the oceanic component of NOAA GlobalTemp v5. ERSSTv5 includes available in situ SST observations from ships, buoys, and Argo floats. Satellite-based observations are not included. The ship and buoy SSTs are from the International Comprehensive Ocean–Atmosphere

Dataset (ICOADS) Release 3.0 (R3.0) (Freeman et al. 2017). The Argo temperature data of 0–5-m depth are defined as SSTs, and the data are from the Argo Global Data Assembly Centre (GDAC; <https://www.seanoe.org/data/00311/42182/>). A recent study (Huang et al. 2019) suggested that Argo and buoy SSTs are playing an equally important role in SST analyses in the global oceans. Huang et al. (2017, 2018, 2019) demonstrated that spatial and temporal variabilities of SSTs are more realistic and more reliable in ERSSTv5 compared with previous versions. Using these observations from ships, buoys, and Argo floats, ERSSTv5 is further processed as follows.

1) QUALITY CONTROL

Observations from ships, buoys, and Argo floats are subject to quality control (QC) before being ingested into the ERSSTv5 system because random and systematic errors may be present. The random errors are larger in ship observations and smaller in buoy and Argo observations (Reynolds et al. 2002; Smith and Reynolds 2003, 2004, 2005; Kent and Challenor 2006; Kennedy et al. 2011a,b), which have been taken into account in parameters 7–9 (Table 2 and appendix A). QC is performed by computing the difference between observations and a first guess at regular $2^\circ \times 2^\circ$ grids or in situ locations using an SST standard deviation (STD; 1σ) and a multiplication factor, while SST STD itself is limited by a minimum and a maximum. The final number of observations ingested into ERSSTv5 depends on 10 QC parameters (parameters 1–10; Table 2 and appendix A).

2) BIAS CORRECTION

Ship SSTs may contain biases for a variety reasons (Smith and Reynolds 2003, 2004, 2005; Kennedy et al. 2011a,b; Huang et al. 2015). Biases in ship SST measurements are corrected in ERSSTv5 using bias-corrected nighttime marine air temperatures (NMATs) from the Hadley Centre (HadNMAT2; Kent et al. 2013) as a large-scale measurement frame of reference before 1985. The quantification of the ship biases depends on the region of interest and variance of SST and NMAT. However, ship biases after 1985 are quantified using more accurate and precise buoy SSTs from ICOADS R3.0, which are used to adjust ship biases determined from NMAT. Similarly, biases in Argo SSTs resulting from slight differences in observing depth are corrected according to buoy SSTs based on statistics over 1990–2010. Corrections to ship and Argo SSTs are designed to maximize compatibility with buoy SSTs at a nominal depth of 0.2 m. However, the compatibility may vary when different options of parameters

11–14 (Table 2 and appendix A) are selected, particularly in the selection of different NMAT and different bias fitting domains.

3) SUPEROBSERVATION AND ITS LOW- AND HIGH-FREQUENCY COMPONENTS

The observations from ships, buoys, and Argo floats are merged into superobservations (superobs) on $2^\circ \times 2^\circ$ grids using a weight determined by their signal-to-noise ratios and a maximum number (5–15) of observations (Reynolds and Smith 1994; Reynolds et al. 2002). Annually averaged superobs are first calculated with a minimum of 1–3 months within a year. The reconstruction itself consists of low- and high-frequency components. The low-frequency (LF) component of annually averaged superobs is determined by a running window of 11–19 yr and a spatial window of $25^\circ \times 25^\circ$. The missing values in superobs fields are filtered out by filling the running average when the ratio of superobs coverage within $25^\circ \times 25^\circ$ reaches a minimum criterion. The high-frequency (HF) component for each grid box is set as the difference between superobs and LF component, and is further filtered by a 1–3-month filter. Therefore, the separation of LF and HF components involves parameters 15–22 (Table 2 and appendix A).

4) HF DECOMPOSITION

The HF component is decomposed using a maximum of 140 EOTs calculated with different sets of training data that are sensitive to the EOTs (Huang et al. 2017). These EOTs are fitted using different weighting methods. Not all 140 EOTs are used to reconstruct the HF component of SSTs. The final selection of EOTs is based on a minimum criterion of EOT variance supported by superobs. The HF decomposition is sensitive to the selections of parameters 23–25 (Table 2 and appendix A).

5) ICE CONCENTRATION CONSTRAINT

In the high latitudes, the reconstructed SST should be consistent with the freezing point of seawater in regions covered with sea ice. Sea ice concentration is derived from satellite observations and may differ around 10% among available ice concentration products. In the area of mixed open water and sea ice, the final SST is interpolated between reconstructed SST and seawater freezing point of -1.8°C . The range of ice concentration is given by a minimum and maximum concentration (Reynolds et al. 2002; Smith et al. 2008). Therefore, options for parameters 26–28 (Table 2 and appendix A) may impact the final SST in the region covered with sea ice.

Most of these 28 parameters in processing ERSSTv5 in sections 3a(1)–3a(5) are the same as those used in

estimating the uncertainty of ERSSTv4 (Huang et al. 2016). New parameters are added here for ERSSTv5 in association with Argo SSTs (Table 2, rows 9, 14, and 16) and the readjustment of ship SSTs according to buoy SSTs (Table 2, row 13). The NMAT selection parameter (Table 2, row 11) used in correcting ship SSTs is revised by adding a $25^\circ \times 25^\circ$ running domain in ERSSTv5. In addition, more parameter options are considered here for the EOT modes (Table 2, row 23).

Options of these 28 parameters in ERSSTv5 are first determined by perturbing parameter values by 10%–100% (Table 2). These options are then randomly selected in generating a 1000-member ensemble of SST anomalies (SSTAs; referenced to the 1970–2000 climatological mean). The ensemble is finally used to quantify the parametric uncertainty.

b. SST uncertainty methods

The uncertainties in the early versions of ERSSTv3b and NOAAGlobalTemp v3 (Smith et al. 2008; Vose et al. 2012) consist of low-frequency uncertainty, high-frequency uncertainty, and bias uncertainty. In the ocean component, the low-frequency uncertainty was estimated using SST variances from a coupled model simulation. The high-frequency uncertainty was estimated using SST variance difference between OISST and ERSST. The bias uncertainty was estimated using the absolute difference of SST biases between ERSSTv3b and HadSST3. These uncertainty estimations in ERSSTv3b and NOAAGlobalTemp v3 was reasonable by applying ERSSTv3b bias correction, which generally decrease with time. However, when the updated SST bias correction of ERSST v4 and v5 was applied, the bias uncertainty and therefore the total uncertainty was large between the 1920s and 1960s and smaller before the 1920s and after the 1960s. In particular, there are no clear reasons to explain why the uncertainty increased from the 1880s to the 1920s. In addition, the estimation of the uncertainties in ERSSTv3b and NOAAGlobalTemp v3 was very much dependent on OISST and model simulations, but not much dependent on ERSST or NOAAGlobalTemp themselves. Therefore, the uncertainty algorithms are updated in ERSSTv4, ERSSTv5, and NOAAGlobalTemp v5 as in the following subsections.

1) PARAMETRIC UNCERTAINTY

Following Huang et al. (2016), SST uncertainty is separated into parametric and reconstruction uncertainties. The 1000-member ensemble is used to assess parametric uncertainty (ε_p) at gridbox level (i.e., local SST uncertainty):

$$\varepsilon_p^2(x, y, t) = \frac{1}{M} \sum_{m=1}^M [A_m(x, y, t) - \bar{A}(x, y, t)]^2, \quad (1)$$

$$\bar{A} = \frac{1}{M} \sum_{m=1}^M A_m(x, y, t), \quad (2)$$

where A_m represents one of an M -member ensemble ERSSTv5 analysis using in situ observations from ships, buoys, and Argo floats; \bar{A} represents the ensemble mean, and $M = 1000$. Symbols x , y , and t represent longitude, latitude, and time, respectively.

Local SST uncertainties from Eq. (1) can be area-averaged into regional- and global-scale uncertainties, which allows typical grid box uncertainties in different regions to be compared easily. However, these area-averaged uncertainties are generally much larger than the uncertainty associated with an area-averaged SST such as the global oceans:

$$\varepsilon_p^2(t) = \frac{1}{M} \sum_{m=1}^M [A_m^g(t) - \bar{A}^g(t)]^2, \quad (3)$$

$$\bar{A}^g = \frac{1}{M} \sum_{m=1}^M A_m^g(t), \quad (4)$$

where superscript g represents the average over the global oceans.

2) RECONSTRUCTION UNCERTAINTY

The intention of applying EOT decomposition in SST reconstruction is to filter out potential noise or random errors in observations and to interpolate available observations to data-void areas using the spatial covariance spanned by the set of EOTs. However, the EOT decomposition with a limited number of modes can bring about residual errors even if observations were perfect (free of noise or random errors and 100% area covered over the entire global oceans). The residual between perfect observations and their EOT decomposition is defined here as reconstruction uncertainty (ε_r), which is the variance not spanned by the set of EOT modes used. As in Huang et al. (2016), pseudo-observational datasets (e.g., gridded products from model simulations, reconstruction analysis, or reanalysis) are used to generate another set of 1000-member ensemble and ε_r of local SST is assessed as follows:

$$\varepsilon_r^2(x, y, t) = \frac{1}{N} \sum_{n=1}^N [A_n(x, y, t) - D(x, y, t)]^2, \quad (5)$$

where A_n represents one member of an N -member ($N = 1000$) ensemble reconstruction using pseudo-observation

dataset $D(x, y, t)$ as described in section 2a. Similar to Eq. (3), ε_r of the globally averaged SST is calculated as

$$\varepsilon_r^2(t) = \frac{1}{N} \sum_{n=1}^N [A_n^g(t) - D^g(t)]^2, \quad (6)$$

where superscript g represents a global average. In section 3d, ε_r is quantified using OISST described in section 2a, which is very similar when different model simulated data are used as shown in Huang et al. (2016).

We want to clarify that the differences between ε_r and ε_p in Eqs. (1) and (5) are in two aspects. First, the estimation of ε_r uses pseudo-observations that have a complete spatial and time coverage, while the estimation of ε_p uses in situ observations that do not have a complete spatial and time coverage. Second, ε_p is quantified as a root-mean-square difference (RMSD) between ensemble members and ensemble average, while ε_r is quantified as a RMSD between ensemble members and pseudo-observations. In principle, ε_p is associated with the uncertainty of temperature when temperature sampling changes with space and time, while ε_r is associated with the residual error that cannot be resolved by a set of limited EOTs.

3) TOTAL UNCERTAINTY

The total uncertainty (ε_t) in ERSSTv5 consists of parametric uncertainty [Eqs. (1) and (5)] and reconstruction uncertainty [Eqs. (3) and (6)]:

$$\varepsilon_t^2(x, y, t) = \varepsilon_p^2(x, y, t) + \varepsilon_r^2(x, y, t_0), \quad (7a)$$

$$\varepsilon_t^2(t) = \varepsilon_p^2(t) + \varepsilon_r^2(t_0), \quad (7b)$$

where the time variable t in ε_p and ε_t represents time in month and year, and time variable t_0 in ε_r represents monthly mean from January to December. The averaged (1982–2017) ε_r is used in Eq. (7) since its variability over time is very small.

It should be noted that the total uncertainty in Eq. (7) does not explicitly include the sampling uncertainty since it has implicitly been included in parameter uncertainty as discussed in Huang et al. (2016). When the sampling of observations is complete in space and time and the quality of observations is perfect, the parametric uncertainty will approach zero and the total uncertainty will approach the reconstruction uncertainty.

c. SST parametric uncertainty

The parametric uncertainty (ε_p) in ERSSTv5 quantified in Eqs. (1) and (2) is generally larger in the earlier period (say 1854–1900; Fig. 1a) than the later periods of 1900–50 (Fig. 1b) and 1950–2010 (Fig. 1c). This is

because with denser sampling in the more recent decades the analysis is less sensitive to the details in the parameter settings (Huang et al. 2017). In the earlier period (Fig. 1a), ε_p is 0.6°–0.8°C in the northwestern North Pacific, the northwestern North Atlantic, the eastern equatorial Pacific, and the eastern equatorial Atlantic, which is largely associated with low observation coverage and/or strong SST variability in these areas. In contrast, ε_p is less than 0.4°C in other regions, particularly in the Arctic and the Southern Ocean. The smaller uncertainty in the Arctic and the Southern Ocean does not necessarily mean that the analysis is accurate, but only implies that the analysis is less sensitive to the changes of the 28 internal parameters in those regions. Dominant factors for the small uncertainty are that the areas are often covered with sea ice and therefore SSTs are less variable. Further, the observations over these regions are extremely sparse, leaving the reconstructed SSTA persistently near zero.

As observation coverage increases, ε_p decreases in the northwestern North Pacific, the northwestern North Atlantic, the eastern equatorial Pacific, and the eastern equatorial Atlantic (Figs. 1b,c). In the later period (Fig. 1c), ε_p is relatively large (approximately 0.2°C) in the Southern Ocean due to sparse observational coverage over the area. With no observations ε_p is low because the analyzed anomaly is always near zero, which implies that there may be an uncertainty component that is not fully accounted for and can be explained by the structural uncertainty (Kennedy 2014). With dense observations ε_p is low because the dense sampling makes parameter details less important. With few observations ε_p can be larger since details of the parameter settings have more impact on the analysis. Averaged over the global oceans (Fig. 2a; solid red), ε_p is approximately 0.4°C before 1880 and decreases gradually to less than 0.2°C after 1960. There are two spikes of ε_p (0.4°C) in the short periods of the two world wars in the later 1910s and early 1940s, respectively, due to low observation coverage.

In contrast, ε_p of the globally averaged SST in ERSSTv5 quantified in Eqs. (3) and (4) is much smaller than that of local SST, which is approximately 0.08°C before 1880 and decreases to approximately 0.04°C in the period of 1950–80 and less than 0.02°C after 1980 (Fig. 2b; red solid line). The ε_p of the globally averaged SST is much smaller than that of the local SST over the global oceans, because the uncertainty of the local SST largely cancels when averaged over the global ocean.

The term ε_p of the local SST in ERSSTv5 (Figs. 1a–c) is mostly consistent with that in ERSSTv4 (Figs. 1d–f) over the global oceans, because internal parameters and their selections are mostly the same in ERSSTv5

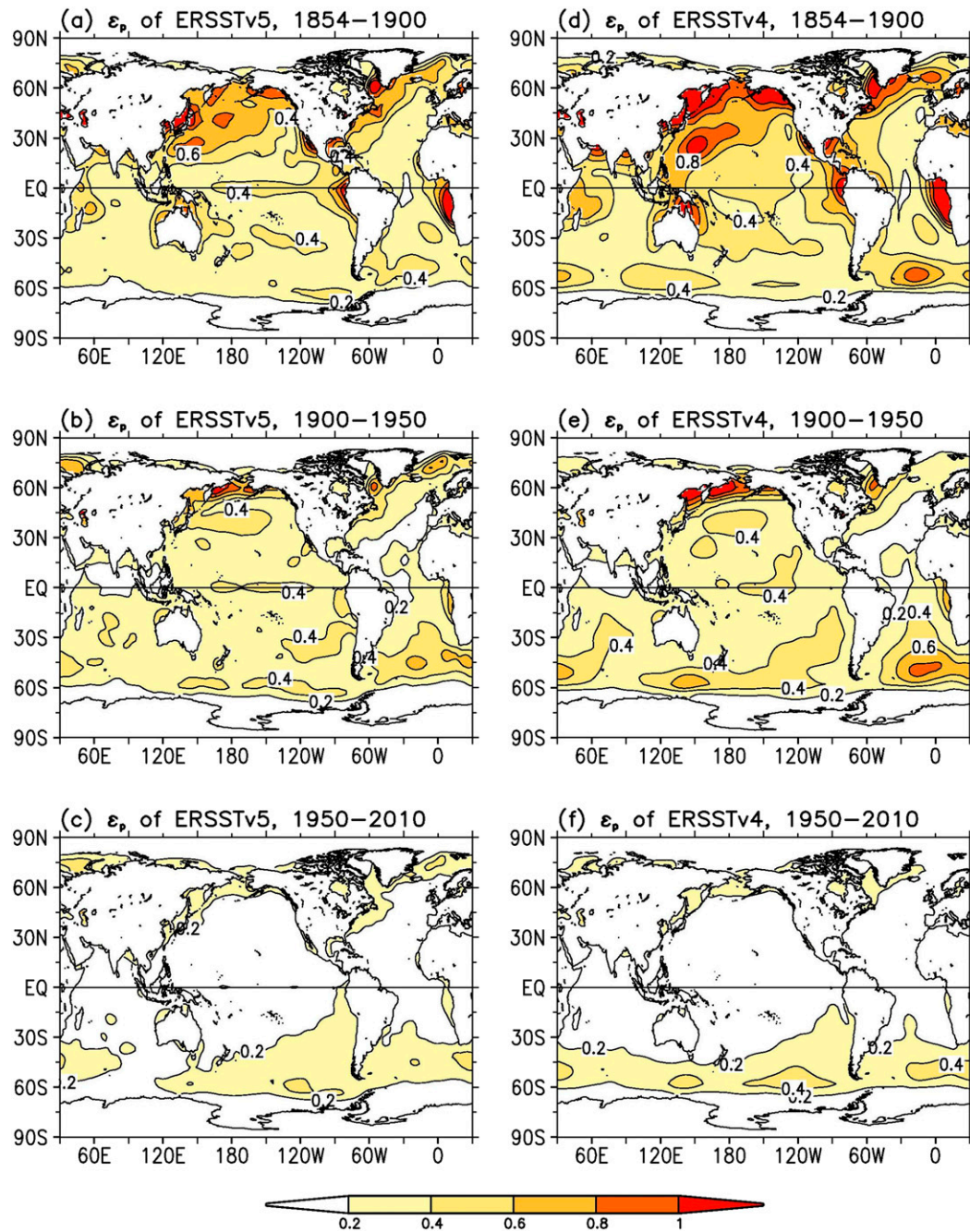


FIG. 1. Parametric uncertainty (1σ) of ERSSTv5 in (a) 1854–1900, (b) 1900–50, and (c) 1950–2010. (d)–(f) As in (a)–(c), but for ERSSTv4. Contour intervals are 0.2°C.

and ERSSTv4. The difference is that the magnitude of ε_p is approximately 0.1°C smaller in ERSSTv5 (Fig. 2a, solid red) than in ERSSTv4 (solid black) before the 1900s and in the late 1910s and early 1940s. Similarly, ε_p of the globally averaged SST is 0.02°–0.04°C smaller in ERSSTv5 (Fig. 2b, solid red) than in ERSSTv4 (solid black) before the 1940s and after the 2000s, although their temporal evolutions are very consistent.

Tests show that these differences of ε_p between ERSSTv5 and ERSSTv4 are largely associated with the changes in EOTs. EOTs in ERSSTv4 are damped to zero north of 65°N and south of 60°S (Huang et al. 2015), while EOTs in ERSSTv5 are not (Huang et al. 2017). The purpose of damping the EOTs in the high latitudes was to avoid a potential overshooting of observed SSTs from lower latitudes to high latitudes. Tests showed that

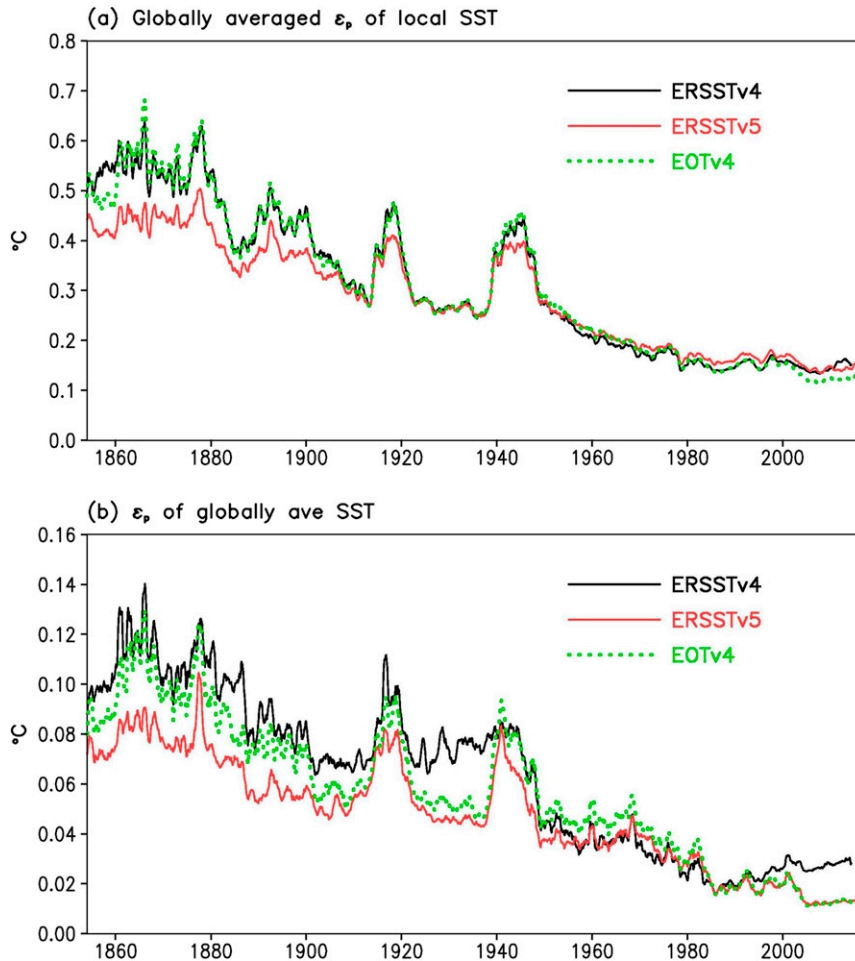


FIG. 2. (a) Globally averaged parametric uncertainty (1σ) of local SST, and (b) parametric uncertainty (1σ) of globally averaged SST in ERSSTv5 (solid red), ERSSTv4 (solid black), and EOTv4 (same as ERSSTv5 except for using EOTs from ERSSTv4; dotted green). A 12-month running filter is applied in plotting.

the damping completely removes the impact of observations in high latitudes, which should be avoided as observations in high latitudes increase rapidly in the modern time period (Huang et al. 2017). The use of nondamped EOTs enables a more reliable analysis that is not sensitive to the selections of the other parameters in ERSSTv5. This is why ε_p is lower in ERSSTv5 than in ERSSTv4. When the EOTs in ERSSTv4 (EOTv4) are used in the ERSSTv5 system while other parameter selections in ERSSTv5 are held constant (Fig. 2), ε_p values of both local and globally averaged SSTs in EOTv4 (dotted green) become close to that in ERSSTv4 (solid black).

However, the change of EOTs cannot explain why ε_p of the globally averaged SST is lower in ERSSTv5 in the periods of 1920–40 and 1990–2017 (Fig. 2b, dotted green and solid black). The lower ε_p over 1920–40 and

1990–2017 in ERSSTv5 is more consistent with overall decreasing uncertainty due to increasing observation coverage, while ε_p in ERSSTv4 increases in these two periods. To detect the causes for lower ε_p in these periods in ERSSTv5, the 1000 ensemble members were grouped according to the selections of a specific parameter. There are typically three potential options for each parameter value, and therefore the size of each group is approximately 333. Uncertainties within each group were calculated for every parameter in both ERSSTv5 and ERSSTv4 as shown in Eq. (3).

Based on this analysis, the higher uncertainty in ERSSTv4 over 1920–40 is associated with selections of the adjusted SSTs in the first-guess options from ERSSTv3b (refer to row 1 of Table 1 in Huang et al. 2016) and the lower value (0.5°C) in minimum STD (Table 2, row 3). In contrast, the uncertainty in ERSSTv5

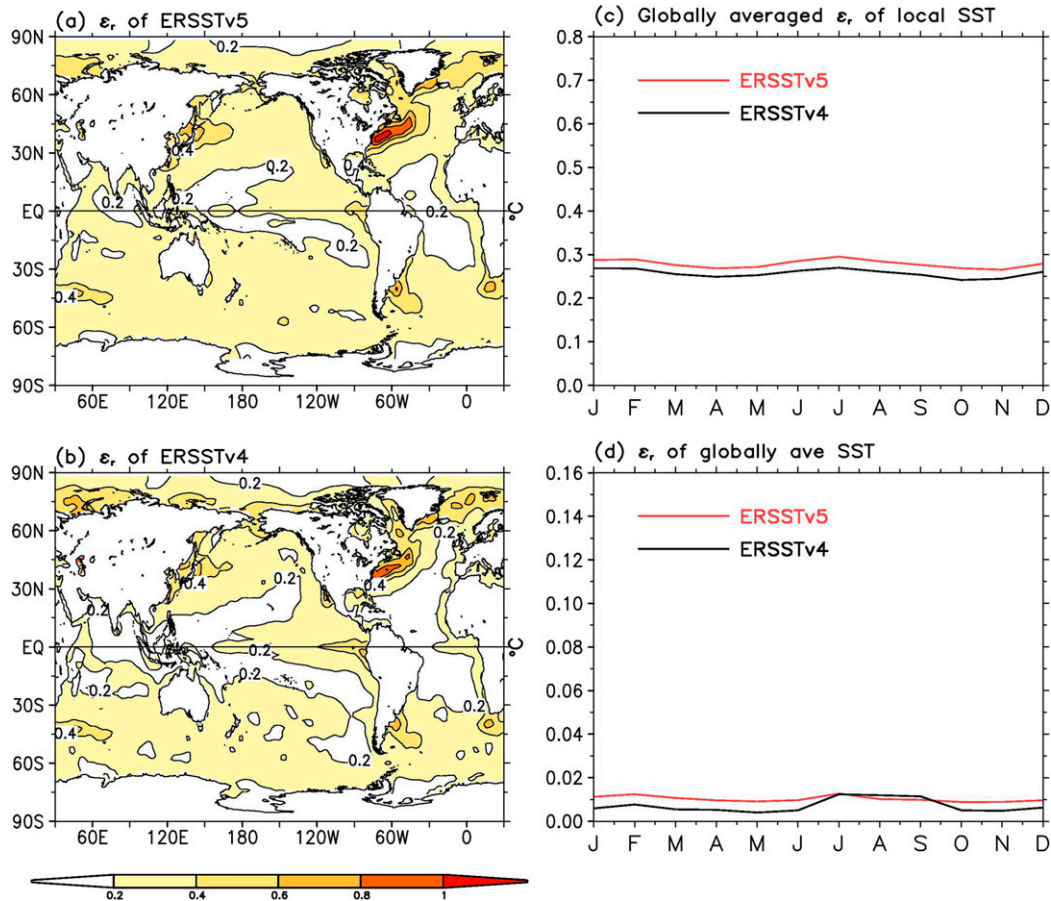


FIG. 3. Averaged (1982–2017) reconstruction uncertainty (1σ) of (a) ERSSTv5 and (b) ERSSTv4, and seasonal variation of (c) globally averaged 1σ reconstruction uncertainty of local SST and (d) reconstructed uncertainty (1σ) of globally averaged SST in ERSSTv5 (solid red) and ERSSTv4 (solid black). Contour intervals are 0.2°C in (a) and (b).

is not sensitive to the first-guess options from ERSSTv4 (Table 2, row 1) and minimum STD due to improved ship SST bias correction over ERSSTv4. This indicates that the use of more recent first guess from ERSSTv4 results in a lower uncertainty over 1920–40. Likewise, our detection shows that the higher uncertainty in ERSSTv4 over 1990–2017 is associated with the selection of a large increment of ship-buoy readjustment [0.04°C ; refer to row 9 of Table 1 in Huang et al. (2016)]. In contrast, the uncertainty in ERSSTv5 does not change much over 1990–2017 because a much smaller increment (0.015°C ; Table 2, row 13) of ship-buoy readjustment is used.

d. SST reconstruction uncertainty

The reconstruction uncertainty (ε_r) quantified in Eq. (5) is usually not sensitive to the selection of the pseudo-observational dataset (Huang et al. 2016). Therefore, ε_r is estimated here using the wOISST

(1982–2017) described in section 2a (Fig. 3). Since this product has a complete spatial coverage over the entire period of 1982–2017, ε_r over the global oceans is nearly constant in time and its spatial distribution is very close in ERSSTv5 and ERSSTv4 (Figs. 3a,b). To avoid the dependence of ε_r estimation on wOISST in operational ERSSTv5 uncertainty production, monthly ε_r is calculated from the uncertainty data over 1982–2017. The averaged ε_r is $0.4^{\circ}\text{--}0.8^{\circ}\text{C}$ in the Gulf Stream, the Kuroshio, and the northern North Atlantic where SST variability is much larger than its global average. In the tropical oceans, ε_r is smaller (approximately 0.2°C). On a global average, ε_r of local SST quantified in Eq. (6) is approximately 0.3°C (Fig. 3c) with little seasonal variation. In contrast, ε_r of the globally averaged SST is very small (0.01°C ; Fig. 3d). It should be noted that ε_r in ERSSTv5 (Figs. 3c and 3d; solid red) is slightly higher than that in ERSSTv4 (solid black) because the values of the 28 internal parameters are randomly selected in

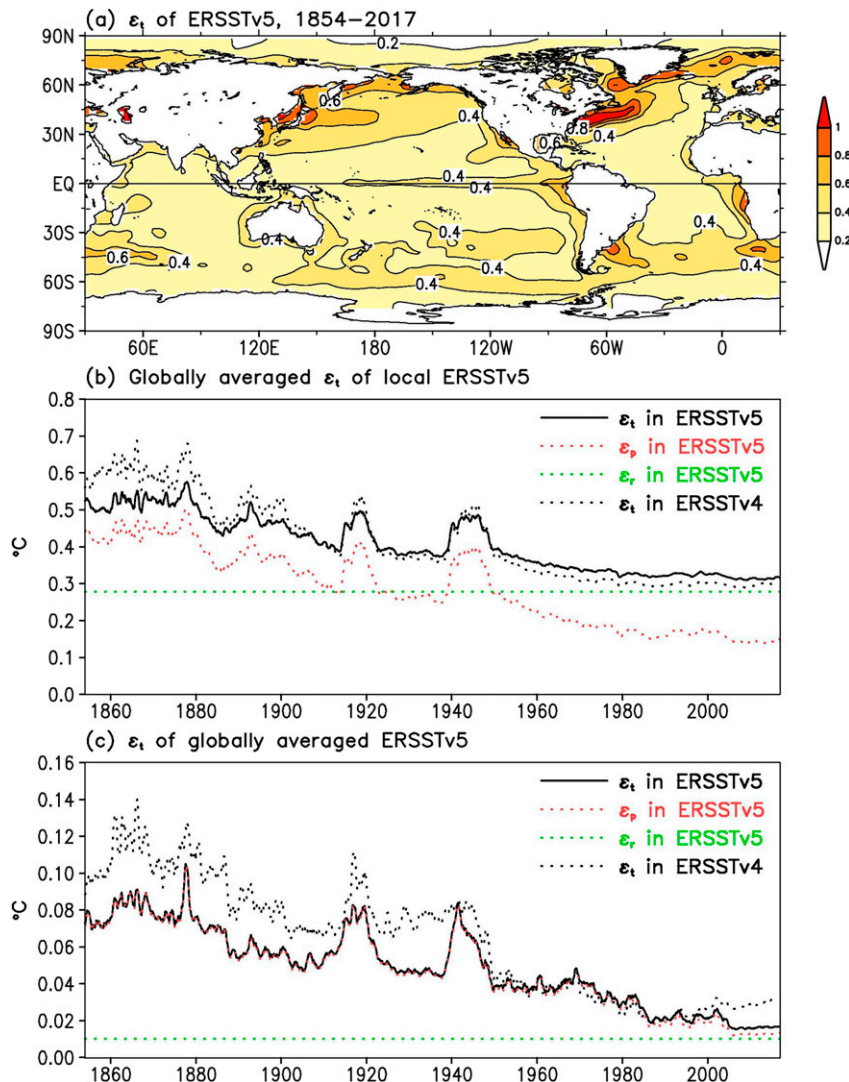


FIG. 4. (a) Averaged (1854–2017) total uncertainty (1σ) of local SST in ERSSTv5, (b) globally averaged total uncertainty (1σ) of local SST, and (c) total uncertainty (1σ) of globally averaged SST in ERSSTv5 (solid black) and ERSSTv4 (dotted black) overlapped with the parametric uncertainty (dotted red) and reconstruction uncertainty (dotted green) in ERSSTv5. Contour intervals are 0.2°C in (a). A 12-month running filter is applied in plotting in (b) and (c).

ERSSTv5 uncertainty estimation whereas the “best” parameter selections are employed in estimating the reconstruction uncertainty in ERSSTv4. The higher ε_r in ERSSTv5 may represent the covariance term of ε_p and ε_r that is ignored in Eq. (7) when total uncertainty is estimated.

e. SST total uncertainty

The total uncertainty (ε_t) is quantified using Eq. (7). The averaged (1854–2017) ε_t (Fig. 4a) is large in the northwestern North Pacific and the northwestern North Atlantic (0.4° – 1.0°C), central-eastern equatorial Pacific,

eastern equatorial Atlantic, and the Southern Ocean (0.4° – 0.6°C) where both ε_p and ε_r are relatively large. Also, ε_t is large on the coasts of the Arctic (0.4° – 0.6°C) due to a large ε_r ; however, overall ε_t in the Arctic is small (approximately 0.2°C) since much of the area is covered by permanent sea ice and therefore both the parametric and reconstruction uncertainties are small. On global average (Fig. 4b), ε_t of local SST is approximately 0.5°C before the 1880s, and decreases gradually to less than 0.4°C after the 1950s. Mostly ε_t (solid black) is attributed to ε_p (dotted red) before the 1910s and during the two world wars. After 1950, ε_t (solid

TABLE 3. LSAT parameters and their operational and alternative options. In parameter number 5, the “even years” are 1982, 1984, . . . , 2010; and the “odd” years are 1983, 1985, . . . , 2011. These parameters are explained in [appendix B](#).

Parameter	Operational option	Alternative options
1. GHCNm v4 data	First of 100 members	Random selection of 100 members
2. Min number of months annual average	2 months	1, 2, 3 months
3. LF filter periods	15 yr	11, 15, 19 yr
4. Min number of years for LF filter	2 yr	1, 2, 3 yr
5. EOTs training periods and spatial scales	1982–2011, $L_x = 4000$ km, $L_y = 2000$ km	1982–2011, $L_x = 4000$ km, $L_y = 2000$ km; 1982–2011, $L_x = 5000$ km, $L_y = 3000$ km; 1982–2011, $L_x = 3000$ km, $L_y = 1000$ km; 1982–2005, $L_x = 4000$ km, $L_y = 2000$ km; 1988–2011, $L_x = 4000$ km, $L_y = 2000$ km; 1982–2011 even years, $L_x = 4000$ km, $L_y = 2000$ km; 1982–2011 odd years, $L_x = 4000$ km, $L_y = 2000$ km
6. EOT acceptance criterion	0.2	0.15, 0.20, 0.25

black) is mostly attributed to ε_r (dotted green). Overall, ε_t in ERSSTv5 (solid black) is consistent with that in ERSSTv4 (dotted black). One exception is that ε_t is lower by 0.1°C in ERSSTv5 before the 1900s due to the nondamped EOTs applied in ERSSTv5. Another exception is that ε_t is slightly higher in ERSSTv5 after the 1950s, which is attributed to the slightly higher ε_r due to random selections of internal parameter values in assessing ε_r in ERSSTv5.

In contrast, ε_t of the globally averaged SST in ERSSTv5 ([Fig. 4c](#); solid black) is dominated by ε_p (dotted red). The contribution from ε_r (dotted green) is much less than that from ε_p until the 2000s. In ERSSTv5 ε_t (solid black) is $0.02^\circ\text{--}0.04^\circ\text{C}$ smaller than that in ERSSTv4 (dotted black) before the 1910s and over 1920–40, and is approximately 0.02°C smaller after 2000. The smaller ε_t is attributed to the use of the nondamped EOTs in ERSSTv5 before the 1910s, to the updated first guess over 1920–40, and to updated ship-buoy adjustment after 1990 as discussed in [section 3c](#).

4. LSAT and its uncertainty

a. GHCNm v4 and its 100-member ensemble

Monthly LSAT anomalies on $5^\circ \times 5^\circ$ grids from GHCNm v4 over 1880–2016 ([Menne et al. 2018](#)) are used as the land component of NOAA GlobalTemp v5. GHCNm v4 includes station data from GHCN-Daily ([Menne et al. 2012](#)), GHCNm v3 ([Lawrimore et al. 2011](#)), and the International Surface Temperature Initiative (ISTI; [Rennie et al. 2014](#)). The number of stations is much larger in GHCNm v4 (approximately 25 000) than in GHCNm v3 (approximately 7000), which is partly attributed to the inclusion of stations with incomplete records over the base period of 1961–90. The area coverage

increases by 3%–15% over 1880–1940, 9%–15% over 1940–90, and approximately 20% over 1990–2016 in GHCNm v4 compared with v3.

GHCNm v4 station data are screened for random errors through spatial and temporal consistency checks ([Lawrimore et al. 2011](#); [Menne et al. 2012, 2018](#)). The station data are homogenized using the pairwise homogenization algorithm (PHA; [Menne and Williams 2009](#)), which detects and minimizes shifts caused by changes in the observing environment surrounding the station, observing instrument replacements, daily observing frequency, and station relocations.

A 100-member ensemble ([Table 3](#), row 1) of GHCNm v4 ([Menne et al. 2018](#)) was used to estimate uncertainties resulting from 1) the methods used to homogenize and grid the station data ([Jones et al. 1997](#); [Morice et al. 2012](#)), 2) nonstandard instrument exposures ([Folland et al. 2001](#); [Brohan et al. 2006](#); [Trewin 2010](#); [Morice et al. 2012](#)), and 3) station distributions within grid boxes ([Jones et al. 1997](#)). Note that only 35 ensemble members are used over Antarctica where the low station density and lack of records prior to the midtwentieth century limits the use of the full range of parameter variations described in [Menne et al. \(2018\)](#).

b. Interpolated LSAT and its 1000-member ensemble

To further explore the uncertainty in LSAT associated with the geographic coverage of the gridbox anomalies, each of the 100-member LSATs from GHCNm v4 is interpolated over the global land and expanded to a 1000-member ensemble as follows ([Smith et al. 2008](#); [Table 3](#) and [appendix B](#)):

- 1) The annual LSAT anomaly is calculated as the mean of the monthly anomalies with a minimum number of months 1–3.

- 2) The annual LSAT anomaly is separated into LF and HF components. The LF component is retrieved by applying a filter of $25^\circ \times 25^\circ$ in space and 11–19 yr in time if a minimum of 1–3-yr data is available.
- 3) The HF component is initially set as the difference between original LSAT anomaly and its LF component, and then decomposed by a maximum of 65 EOTs with different spatial scales and different training periods. However, not all 65 EOTs are used in the decomposition. The acceptance of a specific EOT mode is determined by the acceptance criterion of 0.15–0.25 (Table 3, row 6) that quantifies whether the EOT mode is supported by observations. The acceptance criterion is calculated as a ratio between the EOT variance over the area of observations and the total EOT variance.
- 4) The decomposed HF components are summed and combined with the LF component. By using LF filter and HF decomposition, the GHCNm v4 data are interpolated to the land surface area where no observations are available.

By randomly selecting the values of the parameters listed in Table 3, a 1000-member ensemble of interpolated LSAT ensemble is generated on monthly $5^\circ \times 5^\circ$ grids over the global land surface. This ensemble LSAT is used to assess the parametric uncertainty in the framework of NOAA GlobalTemp v5 in the following section 4c. By using a limited set of 65 EOTs in HF reconstruction, a reconstruction uncertainty (section 4d) is introduced and needs to be included in the total uncertainty (section 4e).

c. LSAT parametric uncertainty

The parametric uncertainty (ε_p) of local and globally averaged LSAT is defined using 1000-member LSAT ensemble in a similar method in Eqs. (1)–(4) in section 3b. Figure 5 shows the averaged ε_p of local LSAT in four time periods. For 1880–1900 (Fig. 5a), ε_p is large (0.8° – 1.5°C) in northern North America, tropical South America between 15°S and 15°N , northern Africa between the equator and 30°N , and northeastern Asia. The large ε_p is associated with sparse data coverage in those regions. Generally ε_p is smaller (0.2° – 0.4°C) in other regions. In Antarctica, there are no observations before the 1950s, and consequently the reconstructed LSAT anomaly is near zero, which may imply an unaccounted for uncertainty component associated with a structural uncertainty (Kennedy 2014). Therefore LSAT variation among ensemble members is not sensitive to the selections of the internal parameters, which results in a small ε_p in Antarctica (0.4° – 0.6°C).

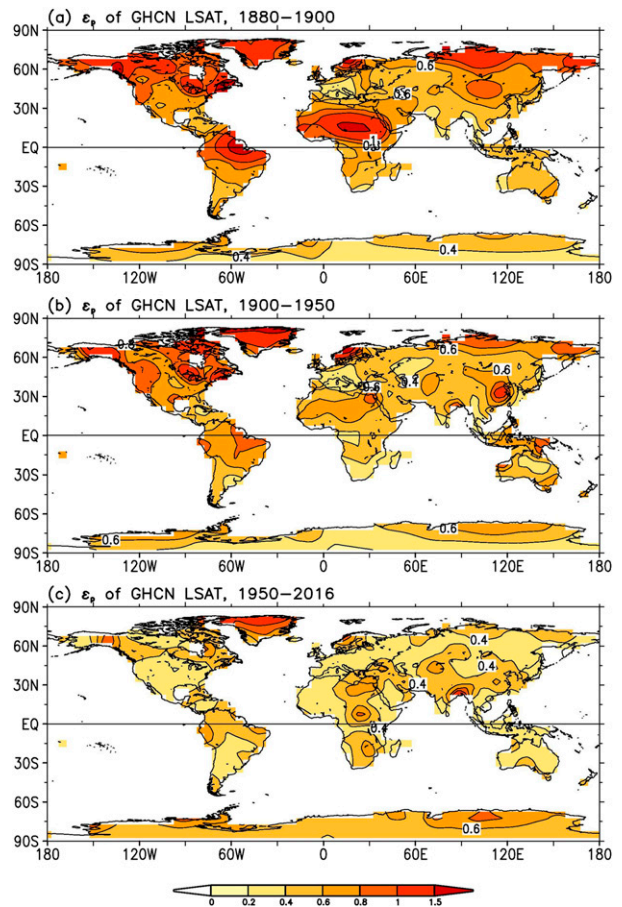


FIG. 5. Parametric uncertainty (1σ) of local LSAT in (a) 1880–1900, (b) 1900–50, and (c) 1950–2016. Contours are 0.2° , 0.4° , 0.6° , 0.8° , 1.0° , and 1.5°C .

For 1900–50 (Fig. 5b), the spatial distributions of ε_p are similar to those over 1880–1900. Note that ε_p remains large (0.8° – 1.5°C) in northern North America. However, the magnitude of ε_p decreases in tropical South America between 30°S and 15°N (0.6° – 0.8°C), northern Africa between the equator and 30°N (0.6°C), and northern Asia north of 60°N (0.6° – 0.8°C). The value of ε_p in Antarctica remains small (0.4° – 0.6°C) due to the absence of observations.

For 1950–2010 (Fig. 5c), ε_p continues to decrease in North and South America, Africa, and Eurasia (0.4° – 0.6°C) as observational coverage increases, while ε_p remains high (0.6° – 1.0°C) in Greenland due to the absence of observations. However, ε_p in Antarctica increases slightly to 0.4° – 0.8°C because observations are collected after the 1950s along the coast of Antarctica, in the islands of the Ross Sea and the Weddell Sea, in the Antarctic Peninsula, and in the interior of Antarctica. However, these observations are sparse and do not cover all of Antarctica, and thus the resulting LSAT

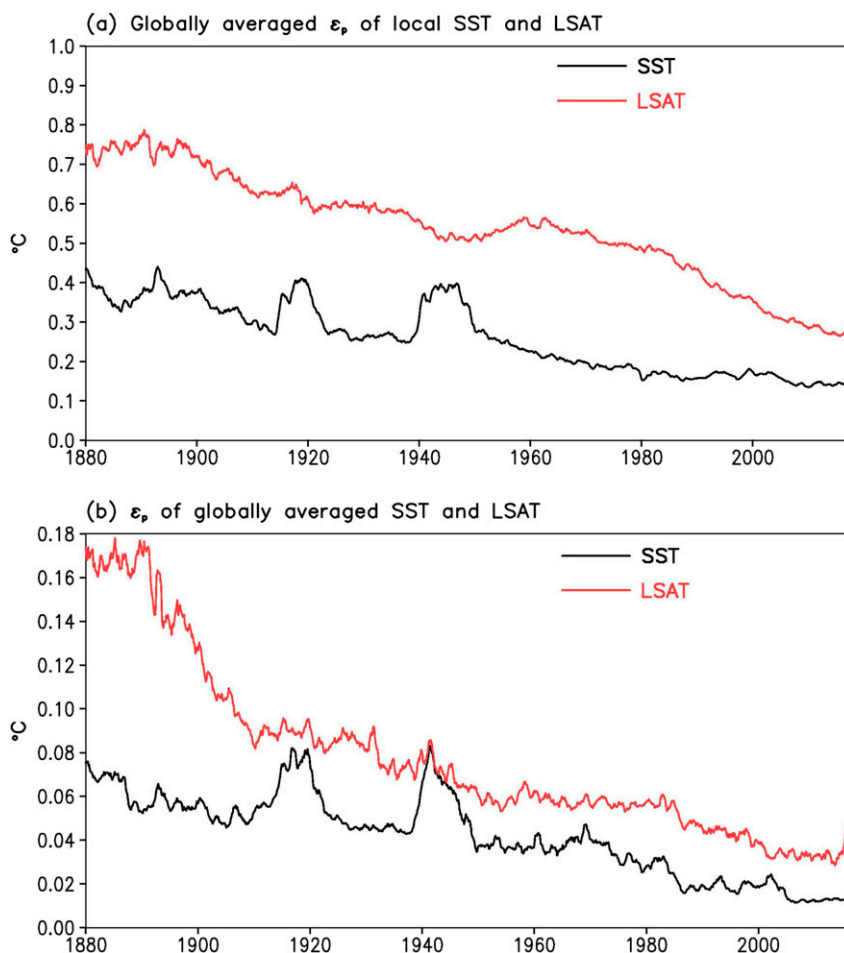


FIG. 6. (a) Globally averaged parametric uncertainty (1σ) of local SST, LSAT, and LSAT without including the Antarctic; (b) parametric uncertainty (1σ) of globally averaged SST, LSAT, and LSAT without including the Antarctic. A 12-month running filter is applied in plotting.

reconstructions are relatively sensitive to the selections of the internal parameters, which leads to a slightly higher ε_p in Antarctica.

Averaged over the global land surface (Fig. 6a, solid red), ε_p of local LSAT is approximately 0.7°C for 1880–1900 and decreases to 0.4°C for the 2000s–2010s except for a slight increase in the 1950s, which is associated mostly with the increased ε_p in Antarctica (Fig. 5c). Overall, ε_p of local LSAT (Fig. 6a, solid red) is approximately 2 times larger than that of local SST (Fig. 6a, dotted black) because of the larger variance in LSAT anomalies compared to SST anomalies.

The value of ε_p of globally averaged LSAT (Fig. 6b, solid red) is much smaller than ε_p of local LSAT. The ε_p of globally averaged LSAT is only 0.09°C – 0.18°C for 1880–1910, decreases to approximately 0.07°C from the 1950s to 1980s, and decreases further to 0.04°C in the 2000s. In comparison to ε_p of globally averaged

SST (Fig. 6b, solid black), ε_p of globally averaged LSAT (Fig. 6b, solid red) is 0.04°C – 0.10°C higher over 1880–1910, about 0.03°C higher over 1920–40, and approximately 0.02°C higher after the 1950s.

d. LSAT reconstruction uncertainty

The reconstruction uncertainty (ε_r) of local and globally averaged LSAT are defined, as for SST, in Eqs. (5) and (6) in section 3b using a 1000-member ensemble. The ε_r of local LSAT is assessed using pseudo-observations of LSATs from coupled model simulations of CanESM2, GFDL-ESM2G, and HadGEM2 over the period of 1861–2007 described in section 2a. Figure 7 shows that the averaged (1861–2007) ε_r is low (approximately 0.4°C) over most of Eurasia, Africa, Australia, North America, and South America, is slightly higher (approximately 0.6°C – 0.8°C) over Greenland, Alaska, western North America, and northern Canada, and is the highest (0.8°C – 1.0°C) over

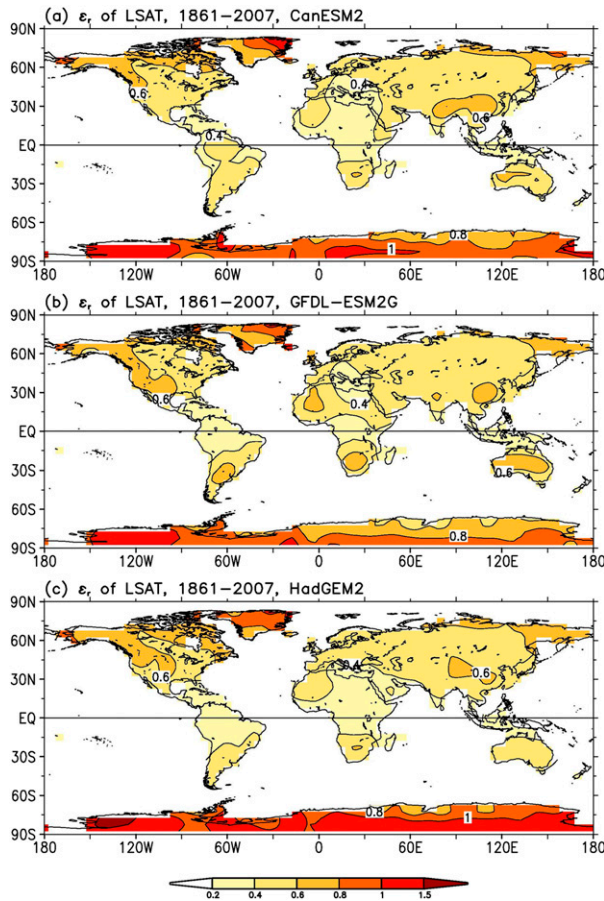


FIG. 7. Averaged (1861–2007) reconstruction uncertainty (1σ) of LSAT using perfect surface air temperature from (a) CanESM2, (b) GFDL-ESM2G, and (c) HadGEM2. Contours are 0.2°, 0.4°, 0.6°, 0.8°, 1.0°, and 1.5°C.

Antarctica. Overall, ε_r is very similar when different LSAT pseudo-observations are used, which suggests that ε_r is stable. Later in section 4e, ε_r derived using HadGEM2 is used to estimate the total uncertainty of LSAT.

There are two factors that result in a high ε_r in Antarctica and Greenland: 1) the higher variability of LSAT and 2) the lower reliability of EOTs in these regions. The variability of LSAT is generally large (2° – 4°C) over Northern Hemisphere land areas north of 40°N and in Antarctica, and is smaller (0.5° – 1.0°C) over the tropical–subtropical land between 60°S and 40°N . Since observations over Antarctica and Greenland are sparse, ERA-Interim and its derived EOTs may be less reliable in these regions. Furthermore, the modes of LSAT variability in ERA-Interim differ from those in the coupled model simulations, which may lower the capability of EOTs derived from ERA-Interim in reconstructing the LSAT from the coupled model.

In Eurasia and North America north of 40°N , EOTs are more reliable due to relatively dense observations. The variability of LSAT is consistent between ERA-Interim and the coupled model simulations, although the variability is large (2° – 4°C). Therefore, ε_r is relatively smaller. In tropical–subtropical regions between 60°S and 40°N , in addition to the reliable EOTs and the consistency of LSAT variability between ERA-Interim and coupled model simulations, the low variability of LSAT (0.5° – 1.0°C) also contributes to the lower ε_r .

Overall, the globally averaged ε_r of local LSAT is nearly constant 0.5°C (Fig. 8a) with a slight seasonal variation (Fig. 8c) using all three coupled model simulations, which is slightly larger than the ε_r of local SST (0.3°C ; Fig. 4b, dotted green). In contrast, the ε_r of the globally averaged LSAT is approximately 0.03°C with little seasonal variation using all three coupled model simulations (Figs. 8b,d), which is slightly larger than that of globally averaged SST (0.01°C ; Fig. 4c, dotted green). The higher ε_r in LSAT than in SST is associated with a smaller number of EOTs in LSAT (65 at maximum) than in SST (140 at maximum), as well as the higher LSAT variance.

e. LSAT total uncertainty

The total uncertainty (ε_t) of LSAT consists of parametric and reconstruction uncertainty as shown in Eq. (7). HadGEM2 is used here to assess the reconstruction uncertainty because it has a slightly higher uncertainty over Antarctica to avoid potential underestimation. The monthly reconstruction uncertainty with seasonal variation is calculated and added with the parametric uncertainty to form ε_r . Figure 9a shows the averaged (1880–2016) ε_t of local LSAT: ε_t is high (0.8° – 1.5°C) over North America, South America near the equator, northern Africa between the equator and 30°N , northeastern Asia north of 60°N , China, and Antarctica. The high ε_t in these regions is mostly attributed to the parametric uncertainty (Fig. 5). However, ε_t in Antarctica is attributed mainly to the reconstruction uncertainty (Fig. 7), while both parametric and reconstruction uncertainties contribute in Greenland, China, and adjacent regions.

Averaged over the global land surface (Fig. 9b, solid black), ε_t of local LSAT is 0.8° – 0.9°C for 1880–1900, decreasing slightly to approximately 0.6°C in the 2010s; this ε_t is mostly attributed to parametric uncertainty (0.5° – 0.7°C) before the 1980s (Fig. 9b, dotted red) and to reconstruction uncertainty (0.5°C) after the 1980s (Fig. 9b, dotted green). In contrast, ε_t of globally averaged LSAT (Fig. 9c, solid black) is mainly due to parametric uncertainty (Fig. 9c, dotted red) over the entire period of the 1880s–2010s. The contribution from reconstruction uncertainty (Fig. 9c, dotted green) is small.

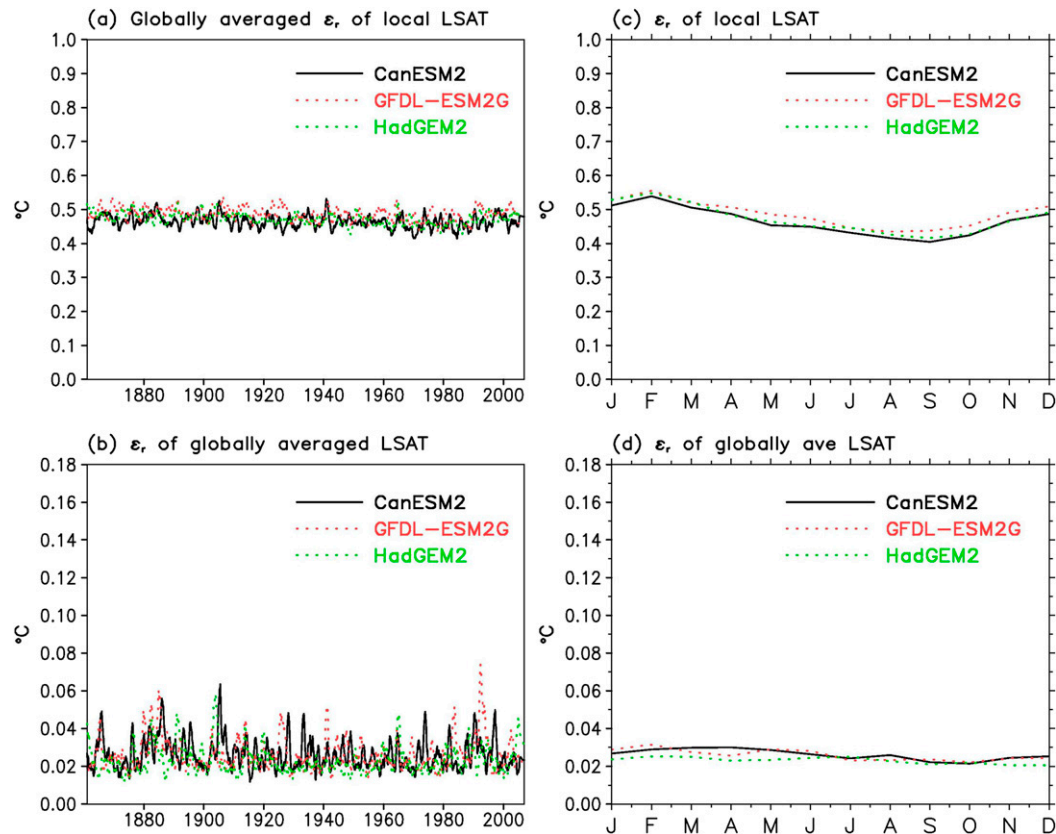


FIG. 8. (a) Globally averaged reconstruction uncertainty (1σ) of local LSAT and (b) reconstruction uncertainty (1σ) of globally averaged LSAT from 1861 to 2007 using perfect data from CanESM2 (solid black), GFDL-ESM2G (dotted red), and HadGEM2 (dotted green). (c),(d) As in (a),(b), but for their seasonal variation.

5. NOAAGlobalTemp v5 and its uncertainty

a. NOAAGlobalTemp v5

NOAAGlobalTemp v5 (Zhang et al. 2019) is a monthly $5^\circ \times 5^\circ$ gridded dataset consisting of LSAT from GHCNm v4 (Menne et al. 2018) and ERSSTv5 (Huang et al. 2017). A 1000-member ensemble of NOAAGlobalTemp v5 is generated by merging ERSSTv5 and LSAT ensembles. For example, a NOAAGlobalTemp ensemble member is produced by merging a randomly selected member of ERSSTv5 ensemble (section 3a) and a randomly selected member of LSAT ensemble (section 4a). The randomly selected members were not removed from the pool and therefore they can be chosen more than once. Tests showed that the uncertainty estimation remained almost the same when the number of NOAAGlobalTemp ensemble members increased from 1000 to 2000. To ensure temporal consistency between ocean and land, the GHCNm v4 anomalies relative to their base period over 1961–90 are adjusted according to a climatological mean over 1971–2000. To ensure gridbox consistency between ocean and land areas, ERSSTv5

anomalies on $2^\circ \times 2^\circ$ grids are interpolated to $1^\circ \times 1^\circ$ grids and then box-averaged to $5^\circ \times 5^\circ$ grids that match GHCNm v4. SSTA and LSAT anomalies are finally merged; boxes with both land and ocean are weighted according to the area ratio of land and ocean within a specific grid box (Smith et al. 2008).

b. GST total uncertainty

The globally averaged uncertainty of local GST is approximated by

$$\bar{\epsilon}_{i,G} = \alpha \bar{\epsilon}_{i,S} + \beta \bar{\epsilon}_{i,L}, \quad (8)$$

where $\bar{\epsilon}_{i,G}$, $\bar{\epsilon}_{i,S}$, and $\bar{\epsilon}_{i,L}$ represent the globally averaged total uncertainty of local GST, SST, and LSAT, respectively; α and β are the ratios of the ocean and land area over the globe, which are approximately 0.71 and 0.29, respectively. The reason for estimating globally averaged uncertainty using Eq. (8) and later Eq. (9) is that the reconstruction uncertainty of GST has to be estimated separately over the land and in the oceans, although the parametric uncertainty of GST can be

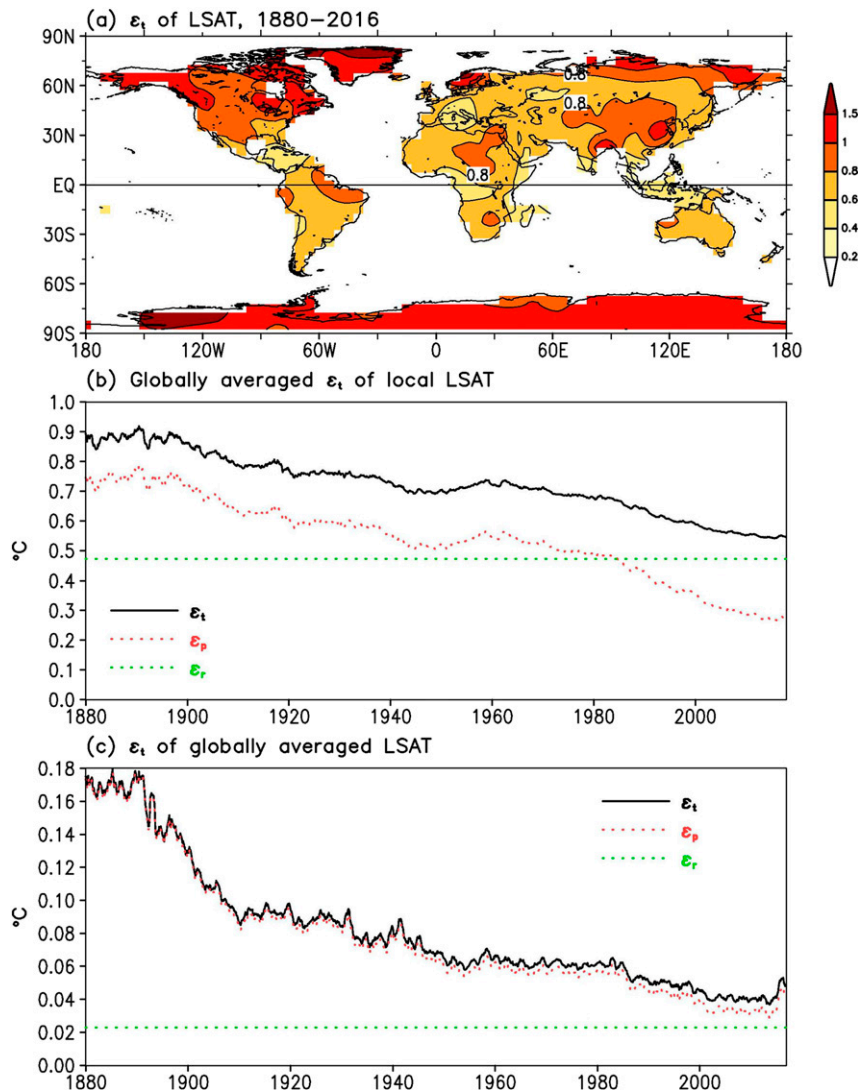


FIG. 9. (a) Averaged (1880–2016) total uncertainty (1σ) of local LSAT, (b) globally averaged total (solid black), parametric (dotted red), and reconstruction (dotted green) uncertainty (1σ) of local LSAT, and (c) total, parametric, and reconstruction uncertainty (1σ) of globally averaged LSAT. Contours are 0.2° , 0.4° , 0.6° , 0.8° , 1.0° , and 1.5° in (a). A 12-month running filter is applied in plotting in (b) and (c).

estimated using merged GST. Since the ocean surface area is more than twice as large as the land surface area, the globally averaged $\bar{\epsilon}_{t,G}$ of local GST (Fig. 10a, solid green) is closer to that of local SST (Fig. 10a, solid black). Overall, $\bar{\epsilon}_{t,G}$ of local GST (Fig. 10a) is approximately 0.6°C for 1880–1900 and decreases to approximately 0.4°C in the 2010s, with two spikes during the two world wars.

The globally averaged $\bar{\epsilon}_{t,G}$ of local GST in NOAA GlobalTemp v5 (Fig. 11a, solid red) is compared with that in HadCRUT4 (Fig. 11a, solid green). The $\bar{\epsilon}_{t,G}$ in HadCRUT4 includes uncorrelated, supplementary, and parametric components over $5^\circ \times 5^\circ$ grids for local

GST, which is attributed to the uncertainties of HadSST3 and CRUTEM4. The uncertainty of CRUTEM4 is further attributed to sampling, station, and bias components (Morice et al. 2012). Comparisons show that $\bar{\epsilon}_{t,G}$ of local GST in NOAA GlobalTemp v5 is $0.1^\circ\text{--}0.2^\circ\text{C}$ higher than that in HadCRUT4. The lower uncertainty in HadCRUT4 results from a lower uncertainty of local LSAT in CRUTEM4 (Fig. 12b, solid green) than in GHCNm v4 (Fig. 12b, solid red), because the uncertainty of local SST in ERSSTv5 is very close to that in HadSST3 (Fig. 12a). The higher uncertainty of local LSAT in reconstructed GHCN v4 is largely a result of its more comprehensive assessment of homogenization

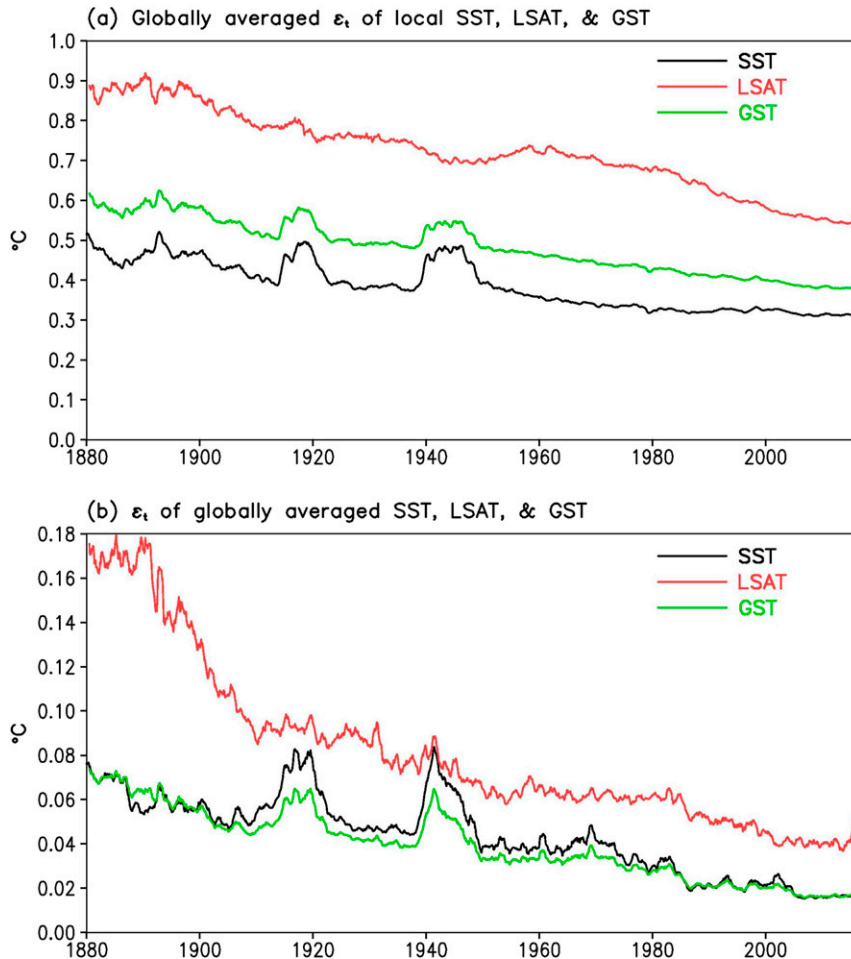


FIG. 10. (a) Globally averaged total uncertainty (1σ) of local SST (black), LSAT (red), and GST (green), and (b) total uncertainty (1σ) of globally averaged SST, LSAT, and GST. A 12-month running filter is applied in plotting.

uncertainty, which leads to more variance in station anomalies within the ensemble, especially for the deeper past, whereas the CRUTEM uncertainty model for this component only accounts for undetected breaks in station data that average to zero. In addition, the inclusion of reconstruction uncertainty may also contribute somewhat to the larger LSAT uncertainty in GHCN v4. When reconstruction uncertainty is excluded in GHCN v4, the uncertainty of local LSAT is very close in GHCN v4 (Fig. 12b, dotted red) and CRUTEM4 (Fig. 12b, solid green) before the 1990s. It should be noted that the uncertainty in CRUTEM4 increases slightly after the 1980s. The reason for this is not immediately clear but is likely associated with reductions in station numbers and spatial coverage. In contrast, the uncertainty in GHCNm v4 decreases gradually after the 1990s.

The total uncertainty ($\varepsilon_{t,G}$) of globally averaged surface temperature (GST) is calculated according to Ku (1966):

$$\varepsilon_{t,G}^2 = \alpha^2 \varepsilon_{t,S}^2 + \beta^2 \varepsilon_{t,L}^2, \quad (9)$$

where $\varepsilon_{t,G}$, $\varepsilon_{t,S}$, and $\varepsilon_{t,L}$ represent the total uncertainty of globally averaged GST, SST, and LSAT, respectively, according to the approximation of $GST = \alpha \times SST + \beta \times LSAT$. The covariance term between SST and LSAT is ignored since reconstructions of SST and LSAT are independent. Calculations show that $\varepsilon_{t,G}$ of globally averaged GST (Fig. 10b, solid green) is $0.05\text{--}0.07^\circ\text{C}$ for 1880–1900 and decreases gradually to approximately 0.02°C in the 2010s except for spikes during the two world wars. Overall, $\varepsilon_{t,G}$ of globally averaged GST is closer to that of globally averaged SST (Fig. 10b, solid black) than to that of globally averaged LSAT (Fig. 10b, solid red), which mostly results from the greater areal weightings of SST.

The value of $\varepsilon_{t,G}$ of globally averaged GST in NOAA GlobalTemp v5 (Fig. 11b, dotted red) is compared with that in HadCRUT4 (Fig. 11b, dotted

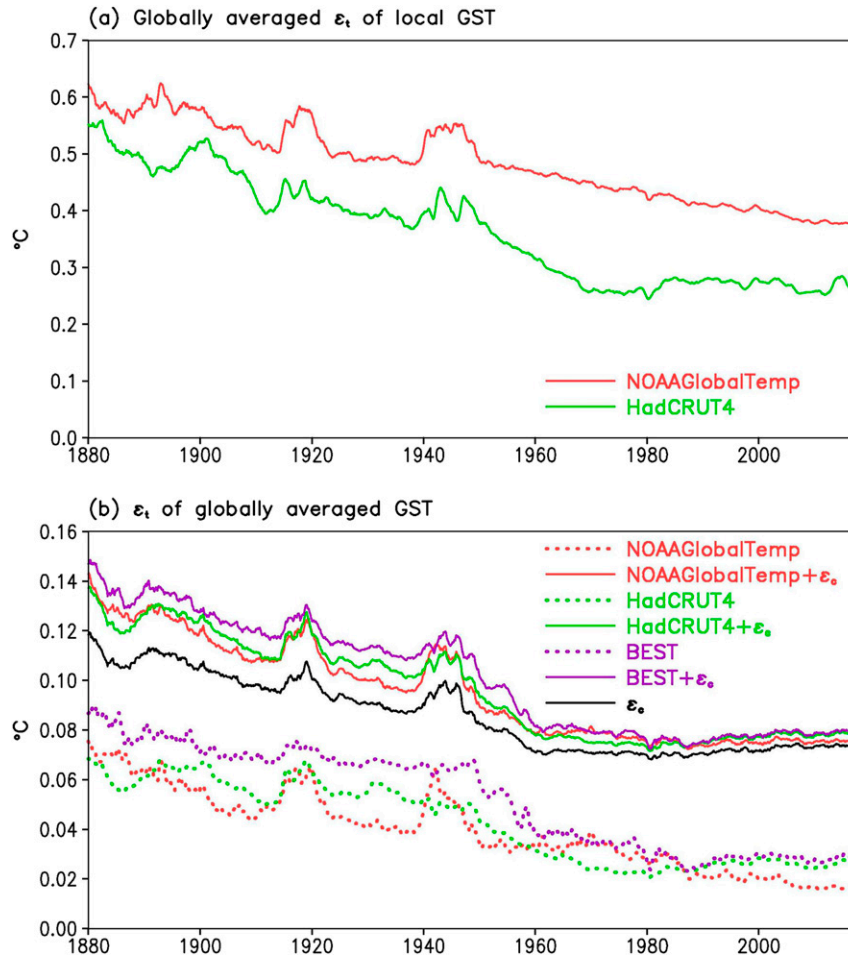


FIG. 11. (a) Globally averaged total uncertainty (1σ) of local GST in NOAAGlobalTemp (solid red) and HadCRUT4 with uncorrelated and supplementary terms (solid green). (b) Uncertainties (1σ) of globally averaged GST in NOAAGlobalTemp with (solid red) and without (dotted red) coverage uncertainty (ϵ_c ; solid black), HadCRUT4 with (solid green) and without (dotted green) ϵ_c , and BEST with (solid purple) and without (dotted purple) ϵ_c . Note that ϵ_c is calculated using HadCRUT4 data mask and the near-surface air temperature from the NCEP–NCAR reanalysis. A 12-month running filter is applied in plotting.

green) and BEST (Fig. 11b, dotted purple). The $\epsilon_{t,G}$ in HadCRUT4 includes components of measurement, sampling, bias, and coverage for globally averaged GST. The $\epsilon_{t,G}$ in BEST includes statistical and spatial under-sampling effects and ocean biases. Overall, $\epsilon_{t,G}$ is consistent among NOAAGlobalTemp v5, HadCRUT4, and BEST, particularly after the 1960s. However, the $\epsilon_{t,G}$ in BEST is slightly higher (by 0.01° – 0.02°C) than in NOAAGlobalTemp v5 and HadCRUT4 for 1880–1960.

Since HadCRUT4 and BEST do not have valid values in every grid box over the global surface, it is necessary to include coverage uncertainty (ϵ_c) (Brohan et al. 2006; Kennedy et al. 2011a,b; Morice et al. 2012) when total uncertainties are compared among products. The term

ϵ_c is associated with the error in estimation of globally averaged surface temperature with nonglobally covered data; ϵ_c can be calculated by combining a selected data mask (e.g., HadCRUT4) and a series of monthly pseudo-observations. First, all monthly pseudo-observations are subsampled with the data mask at a specific month. Second, global averages are calculated for both subsampled and spatially complete pseudo-observations. Finally, the STD between the global averages is defined as the coverage uncertainty for that specific data mask.

A common or collocated data mask among NOAAGlobalTemp v5, HadCRUT, and BEST is used to calculate ϵ_c using near-surface air temperature in the NCEP–NCAR reanalysis (Morice et al. 2012). By

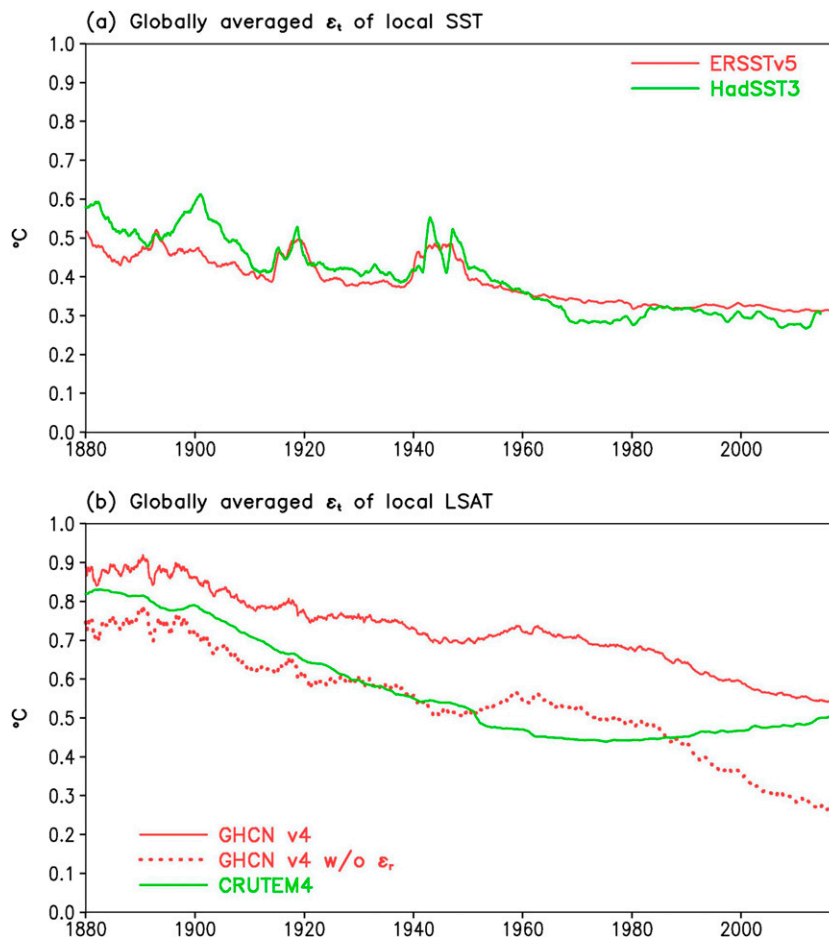


FIG. 12. Globally averaged uncertainties (1σ) of (a) local SST in ERSSTv5 (solid red) and HadSST3 (solid green) and (b) local LSAT in GHCN v4 (solid red), GHCN v4 without including reconstruction uncertainty (ε_r ; dotted red), and CRUTEM4 (solid green).

including ε_c (Fig. 11b, solid black), $\varepsilon_{t,G}$ is very consistent among the three products (Fig. 11b; solid red, green, and purple lines) since ε_c is the dominant term and the same ε_c is used in all three products.

It should be noted that ε_c (Fig. 11b, solid black) is clearly larger than the total uncertainties of the globally averaged GSTs without including ε_c (Fig. 11b, dotted lines), which may imply the importance of the spatial coverage of observations to the uncertainty of GSTs. It should also be noted that ε_c may depend on the spatial variability of datasets for a given data mask. For comparison purposes, ε_c is calculated using the near-surface air temperatures from ERA-Interim, and the LSAT over the land and the SST over the oceans in HadGEM2-AO and NOAAGlobalTemp v5. Comparisons indicate that ε_c deviates very slightly (less than 0.02°C) when different near-surface temperatures are used.

The uncertainties in this study are for the temperatures at monthly time scale, although a 12-month running

average is applied for the clarity of comparisons among different uncertainty components and among different products. The uncertainties at monthly time scale are much larger than those at annual time scale as indicated in Kennedy et al. (2011a).

c. GST comparisons

The globally and ensemble averaged GST of NOAAGlobalTemp v5 is compared with those of HadCRUT4, BEST, and GISTEMP (Fig. 13). To make the comparison fair, the global average of NOAAGlobalTemp v5 is first filtered with the common data mask. Following HadCRUT4 (Morice et al. 2012), the global average is first separated into the averages of the NH and SH, and then the arithmetic average of the NH and SH is calculated [i.e., hemispheric averages (HAs)]. Figure 13 shows that the GST derived from HAs in NOAAGlobalTemp v5 is consistent with that in HadCRUT4. However, the GST in NOAAGlobalTemp

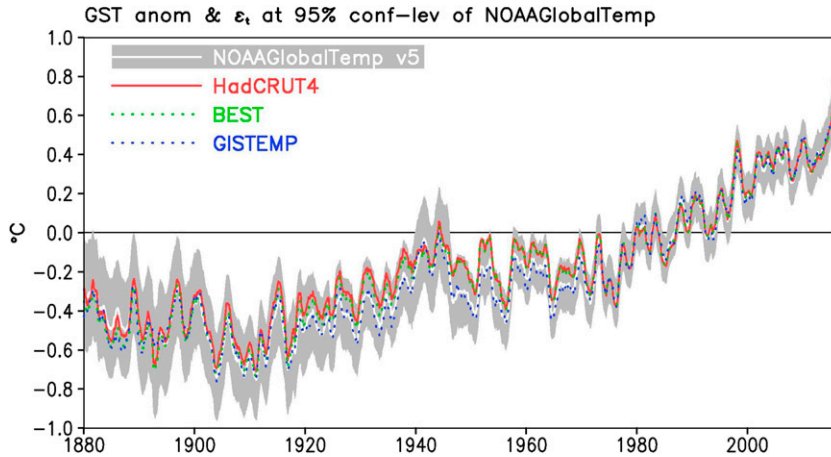


FIG. 13. Globally averaged GST anomaly in NOAAGlobalTemp v5 (solid white), HadCRUT4 (solid red), BEST (dotted green), GISTEMP (dotted blue), and globally averaged uncertainty at the 95% confidence level in NOAAGlobalTemp v5 (gray shading). Global averages are based on hemispheric averages (HAs) using the common data mask where all data have valid data. A 12-month running filter is applied in plotting.

v5 is slightly lower than HadCRUT4 for the 1920s–1960s largely due to colder SST in ERSSTv5 (Huang et al. 2015, 2017). The differences are within the uncertainty at the 95% confidence level (1.96σ) regardless of which uncertainty is used (i.e., NOAAGlobalTemp v5, HadCRUT4, or BEST). It should be noted that the globally averaged GSTs are very close in operational and ensemble averaged NOAAGlobalTemp v5 and therefore the distribution of the ensemble GSTs is nearly symmetric around the operational GST (not shown).

The linear trends of hemispheric and annually averaged GST are calculated over different time periods for each ensemble member. The fitting uncertainty at the 95% confidence level is estimated by considering the effective sampling number scaled by lag-1 autocorrelation (von Storch and Zwiers 1999). For a given time series $T(i)$ of member i at time t , temperature can be fitted linearly as

$$T(i) = a(i) + b(i)t \pm c(i), \quad \text{for } i = 1, N, \quad (10)$$

where $a(i)$ is a constant, $b(i)$ is a fitted linear trend, and $c(i)$ is a fitting uncertainty at 95% percent confidence level. Following Karl et al. (2015), the uncertainty of a fitted trend (ε_t) consists of data uncertainty (ε_d) and fitting uncertainty (ε_f):

$$\varepsilon_t^2 = \varepsilon_d^2 + \varepsilon_f^2. \quad (11)$$

Here ε_d is quantified as 1.96σ of the N -member linear trends $b(i)$; ε_f is quantified as the ensemble average of the N -member fitting uncertainties $c(i)$. Tests using

NOAAGlobalTemp v5 and HadCRUT4 indicate that the uncertainty of a linear trend is mostly attributed to ε_f while the contribution from ε_d is smaller, and that 1σ variation of $\varepsilon_f(i)$ among ensemble members is much smaller than the ensemble average of $\varepsilon_f(i)$. These features indicate that the deviations among the ensemble members of the globally averaged GST time series are mostly systematic for a given set of randomly selected parameter values within a specified reconstruction methodology.

Table 4 displays the ensemble-averaged trends and their uncertainties at the 95% confidence level in NOAAGlobalTemp v5 over different time periods.

TABLE 4. Linear trends ($^{\circ}\text{C decade}^{-1}$) \pm their uncertainty at the 95% confidence level of globally averaged GST. Global averages are derived using HAs on the grids where all NOAAGlobalTemp v5, HadCRUT4, BEST, and GISTEMP have valid data. Linear trends are ensemble averages of 1000 members in NOAAGlobalTemp v5, 100 members in HadCRUT4, and one member in BEST and GISTEMP. Uncertainties include data uncertainty and fitting uncertainty in Eq. (11), and have taken into account the effective sampling number quantified by lag-1 autocorrelation.

	1880–2016	1950–2016	2000–16
NOAAGlobalTemp v5 Ensemble	0.069 ± 0.012	0.134 ± 0.020	0.187 ± 0.110
HadCRUT4	0.066 ± 0.010	0.118 ± 0.024	0.163 ± 0.114
BEST	0.070 ± 0.009	0.119 ± 0.022	0.174 ± 0.113
GISTEMP	0.071 ± 0.011	0.141 ± 0.019	0.198 ± 0.115
NOAAGlobalTemp v5 Operational	0.071 ± 0.011	0.135 ± 0.019	0.197 ± 0.114
NOAAGlobalTemp v4 Operational	0.068 ± 0.011	0.135 ± 0.018	0.194 ± 0.110

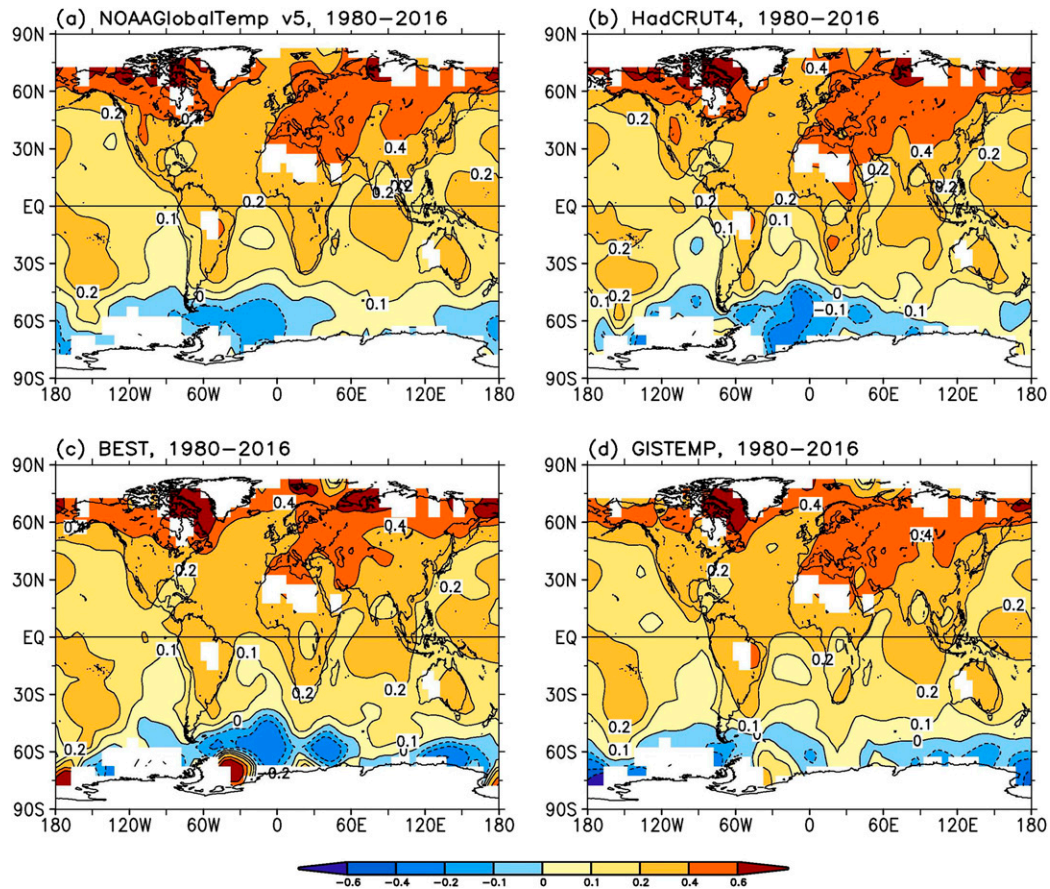


FIG. 14. Averaged (1980–2016) GST in (a) NOAA GlobalTemp v5, (b) HadCRUT4, (c) BEST, and (d) GISTEMP over a common data mask. Contours are 0° , $\pm 0.1^{\circ}$, $\pm 0.2^{\circ}$, $\pm 0.4^{\circ}$, and $\pm 0.6^{\circ}\text{C}$.

Overall, the warming trends of GSTs are statistically significant. The GST ensemble average trends are weak ($0.069^{\circ} \pm 0.012^{\circ}\text{C decade}^{-1}$) in the centennial time scale (1880–2016), moderate ($0.134^{\circ} \pm 0.019^{\circ}\text{C decade}^{-1}$) in the decadal time scale (1950–2016), and highest ($0.188^{\circ} \pm 0.110^{\circ}\text{C decade}^{-1}$) in the twenty-first century (2000–16), suggesting a stronger warming of GST in the recent decades. The warming trend in NOAA GlobalTemp v5 is slightly stronger than in HadCRUT4 and BEST, which is mostly attributed to a stronger SST trend in ERSSTv5 than HadSST3 (Huang et al. 2015, 2017). However, the warming trend is slightly weaker in NOAA GlobalTemp v5 than in GISTEMP, probably due to different interpolation method over the land surface. It should be noted that different time periods of warming are used to measure the changes in warming at different time scales. It may be argued that the higher warming in the twenty-first century is associated with its short time period and stronger external forcing.

The reason for using HAs is to compensate for hemispheric differences in data coverage (i.e., poorer

coverage in the SH). However, comparisons show that the global area-weighted average of GST and its trends are very close to those of HAs in both NOAA GlobalTemp v5 and HadCRUT4 (not shown). The trend differences between those two methods are small (less than 2%). In Fig. 13 and Table 4, ensemble averages of GSTs are calculated as the arithmetic mean of N members. To avoid the impact of extremely low or high GSTs among the ensemble members, the median of N members is selected in some studies to represent an ensemble average (Morice et al. 2012; Kennedy et al. 2011a,b, 2019). However, our comparisons indicate that the difference between the arithmetic mean and median is small in both NOAA GlobalTemp v5 and HadCRUT4.

After the 1980s (Fig. 13), the spatial distribution of GSTs is similar among those products. For example, the averaged GSTs over 1980–2016 (Fig. 14) show a higher temperature anomaly (0.4° to 0.6°C) in NH land areas and a lower temperature anomaly (-0.1° to -0.2°C) in the Southern Ocean. The spatial correlations among

those products are high, with a maximum correlation (0.96) between NOAAGlobalTemp v5 and HadCRUT4 and a minimum correlation (0.89) between BEST and GISTEMP. Over the period of 1920–70 when globally averaged GSTs differ most among those products (Fig. 13), the spatial correlations decrease slightly, with a maximum correlation (0.89) between NOAAGlobalTemp v5 and HadCRUT4 and a minimum correlation (0.70) between BEST and GISTEMP. Over the early period of 1880–1920, the globally averaged GSTs are similar (Fig. 13), but their spatial distributions differ more, with a maximum correlation of 0.76 between HadCRUT4 and BEST and a minimum correlation of 0.26 between HadCRUT4 and GISTEMP.

6. Summary

NOAAGlobalTemp v5 (Zhang et al. 2019) consists of monthly GST on $5^\circ \times 5^\circ$ grids based on SSTs in ERSSTv5 (Huang et al. 2017) and LSAT from GHCNm v4 (Menne et al. 2018). An important aspect of these SST, LSAT, and GST products is their uncertainty, which is a measure of the reliability of a product. Uncertainties of SST and LSAT consist of parametric (ε_p) and reconstruction (ε_r) uncertainties. The uncertainty of local (or globally averaged) GST in NOAAGlobalTemp v5 is quantified by the linear (or squared) summation of the uncertainties of SST and LSAT weighted by the ratio of land (29%) and ocean (71%) surface area over the globe [Eqs. (8) and (9)].

The term ε_p represents the sensitivity of an analysis to its internal parameters; ε_p is quantified by 1σ standard deviation of SST or LSAT among a multimember ensemble. The multimember ensemble is created by perturbing the values of internal parameters. There are 28 and 6 internal parameters in SST and LSAT reconstruction systems, respectively. The number of parameters in ERSSTv5 (28) is more than that in LSAT (6) because the parameters GHCNm v4 already include other parameters such as observation homogenization, biases, and random errors in 100-member GHCNm datasets (Menne et al. 2018). A total of 1000 members of SST, LSAT, and GST are generated to extensively explore the uncertainty space of the parameters.

The term ε_r represents the residual that cannot be resolved by a limited number of statistical modes in an analysis, even if observations are perfect and spatially complete. The maximum number of modes is 140 for SST and 65 for LSAT; ε_r is estimated by 1σ standard deviation of the difference between reconstructed and

original pseudo-observations from model simulations or independent analyses of observations. A total of 1000 members of SST and LSAT are generated to extensively explore the uncertainty space. The most important features of the uncertainty in NOAAGlobalTemp v5 are the following:

- 1) Uncertainties of globally averaged SST, LSAT, and GST ($0.02^\circ\text{--}0.18^\circ\text{C}$; Fig. 10b) are much smaller than those at local grid scale ($0.4^\circ\text{--}0.9^\circ\text{C}$; Fig. 10a), because the errors in SST, LSAT, and GST analyses cancel when averaged globally.
- 2) Uncertainties at local grid scale are larger over the land ($0.7^\circ\text{--}0.9^\circ\text{C}$) than over the oceans ($0.3^\circ\text{--}0.5^\circ\text{C}$), consistent with the larger variability in LSAT than in SST. Similarly, the uncertainty at global scale is larger over the land ($0.05^\circ\text{--}0.18^\circ\text{C}$) than over the oceans ($0.02^\circ\text{--}0.10^\circ\text{C}$). The uncertainty of GST is close to that of SST because the ocean area is more than 2 times larger than the land area.
- 3) Uncertainties of SST, LSAT, and GST are large ($0.4^\circ\text{--}0.9^\circ\text{C}$ at local grid scale and $0.05^\circ\text{--}0.18^\circ\text{C}$ at global scale) in the earlier periods generally decrease with time (except during the two world wars), and are smallest in the modern era ($0.4^\circ\text{--}0.7^\circ\text{C}$ at local grid scale and $0.02^\circ\text{--}0.06^\circ\text{C}$ at global scale). These features clearly indicate that the decreasing uncertainty with time is directly associated with the increasing numbers and spatial coverage of both SST and LSAT observations (e.g., Huang et al. 2017; Menne et al. 2018).
- 4) The values of ε_p and ε_r of SST are large in the areas of the Kuroshio, the Gulf Stream, the eastern equatorial Pacific and Atlantic, and the Southern Ocean where observations are sparse and/or SST variability is large. At the local grid scale, ε_p is dominant over ε_r before the 1910s. In contrast, ε_r is dominant over ε_p after the 1950s. Between the 1910s and 1950s ε_p and ε_r are comparable. At the global scale, ε_p is dominant over ε_r through the entire period until the 2010s, when both are very small.
- 5) The value of ε_p of LSAT is large in northern North America, South America near the equator, northern Africa ($0^\circ\text{--}30^\circ\text{N}$), and northeastern Asia; ε_r of LSAT is large in Greenland and Antarctica. At the local grid scale, ε_p is larger than ε_r before the 1940s, comparable over the 1940s to 1980s, and smaller after the 1990s. At the global scale, ε_p is dominant over ε_r throughout the entire period.
- 6) Comparisons indicate that uncertainties of NOAAGlobalTemp v5 are very close to those independently assessed in HadCRUT4, BEST,

and GISTEMP at both the local grid and the global scale.

Globally and ensemble averaged GST and its uncertainty in NOAAGlobalTemp v5 are compared against those in HadCRUT4, BEST, and GISTEMP. Comparisons show that the GSTs are consistent over the 1880s to 1900s and the 1970s to 2010s. In contrast, GST is slightly warmer in HadCRUT4 than in NOAAGlobalTemp v5 over the 1900s to 1940s and the 1940s to 1970s, which is mostly attributed to the higher SST in HadSST3 than in ERSSTv5 (Huang et al. 2017). Overall, the difference of GSTs is small between NOAAGlobalTemp v5 and GISTEMP, since the same SSTs from ERSSTv5 are used. Similarly, the GST difference is small between HadCRUT4 and BEST because the same SSTs from HadSST3 are used. However, these differences are within the uncertainty ranges at the 95% confidence level, indicating overall consistency among NOAAGlobalTemp v5, HadCRUT4, BEST, and GISTEMP.

All products (NOAAGlobalTemp v5, HadCRUT4, BEST, and GISTEMP) show that the warming over the global surface is stronger in the recent decades than in the past 50–100 years as described in many other studies (e.g., Karl et al. 2015). The difference is that the warming trends in the recent decade in NOAAGlobalTemp v5 are slightly higher than those in HadCRUT4 and BEST, but slightly lower than that in GISTEMP at various time scales.

In conclusion, the global surface temperature and its uncertainty in NOAAGlobalTemp are consistent with other studies. The warming trend of the global surface temperature is persistent over time and stronger in the recent decades.

Acknowledgments. The authors thank the three anonymous reviewers for their helpful comments and suggestions that have greatly improved the manuscript. The data from BEST are downloaded from <http://berkeleyearth.org/data> (16 November 2018). The data from GISTEMP are downloaded from <https://data.giss.nasa.gov/gistemp> (13 November 2018). The data from HadCRUT4 and CRUTEM4 are downloaded from <https://www.metoffice.gov.uk/hadobs> (29 May 2018). The data from monthly OISST are downloaded from <http://ftp.emc.ncep.noaa.gov/cmb/sst> (17 January 2012). The data from ERA-Interim (<https://doi.org/10.5065/D63B5XW1>) are downloaded from <https://climatedataguide.ucar.edu/climate-data/era-interim> (21 September 2015). The data from GFDL-ESM2G, HadGEM-AO, and CanESM are downloaded from <https://esgf-node.llnl.gov/search/cmip5> (30 October 2012). ERSSTv5 ensemble data are available at

<ftp://ftp.ncdc.noaa.gov/pub/data/cmb/ersst/v5/2019.ersstv5.ensemble>. NOAAGlobalTemp v5 ensemble data are available at <ftp://ftp.ncdc.noaa.gov/pub/data/cmb/ersst/v5/2019.ngtv5.ensemble>. NOAAGlobalTemp v5 and v4 operational products are available at <ftp://ftp.ncdc.noaa.gov/pub/data/noaaglobaltemp> (30 December 2018).

APPENDIX A

ERSSTv5 Internal Parameters and Their Options

The number of internal parameters increases from 24 in ERSSTv4 (Huang et al. 2015, 2016) to 28 in ERSSTv5 (Huang et al. 2017). The parameters in ERSSTv5 are assigned 2 to 7 optional values (Table 2). The “best” combination of these options is selected and used in the operational ERSSTv5 production. The other alternative options are used for parametric uncertainty estimation. The details of these parameters and their options in ERSST (v3b, v4, and v5) are described as follows:

a. First guess (FG) used for quality control (QC)

The deviation of observations from FG is assessed to ensure that outlier observations are not included in the analysis. The adjusted SSTs are used as an FG in ERSSTv3b (Smith et al. 2008) and ERSSTv4 (Huang et al. 2016), but the unadjusted SSTs are used in ERSSTv5 (Huang et al. 2017). Since raw observations are not bias adjusted, the use of unadjusted SST in QC is better to filter out true outliers (Huang et al. 2017). The unadjusted and adjusted SSTs from ERSSTv4 are used to assess the contribution of FG to the uncertainty of ERSSTv5.

b. SST standard deviation (STD) used for QC

The observed raw SSTs may be discarded in QC procedure, if they deviate from FG by more than 4.5 times the SST STD. Two sets of SST STDs are used. One is derived from COADS observations from 1950 to 1979 (Woodruff et al. 1998) and applied in ERSSTv3b. The other is from monthly OISST from 1982–2011 (Reynolds et al. 2002) and applied in ERSSTv4 and ERSSTv5. Since SST STD is smaller in OISST than in COADS, fewer SST raw observations may be included when the STD from OISST is applied. The factor of 4.5 is termed the STD multiplier and may vary as described in parameter 5.

c. Minimum SST STD

To maintain a good QC procedure, a minimum STD (1.0°C) is set in ERSST, and its alternative options are 0.5° and 1.5°C.

d. Maximum SST STD

Similar to the minimum SST STD, a maximum STD (4.5°C) is set in ERSST, and its alternative options are 3.5° and 5.5°C .

e. SST STD multiplier

The multiplier to STD in QC procedure is set to 4.5 in ERSST, and its alternative options are 3.5 and 5.5. A larger (smaller) value of maximum (minimum) STD and a larger STD multiplier enables ERSST system to include more (fewer) extreme raw SST observations into subsequent SST processing.

f. Random error of SST observations

The random error of SST observations is added to a single ship, buoy, or Argo measurement in the uncertainty estimation of ERSSTv4 (Huang et al. 2016) and ERSSTv5, while it is not added in ERSST operational production. The mean of the random error is set to 0°C , and the STD of the random error is set to the magnitude of random errors for ships, buoys, and Argo floats as explained in the next subsection.

g. Ship, buoy, and Argo SST errors

Random errors of ship and buoy observations are different, which are approximately 1.3° and 0.5°C (Reynolds et al. 2002; Kent and Challenor 2006; Huang et al. 2017), respectively. The random error of Argo observation is set to be the same as that of buoy observation due to the same type of temperature sensor in buoys and Argo floats (Huang et al. 2017). These empirically derived errors are somewhat uncertain when they are taken into account in weighting EOTs [refer to Eq. (3) in Huang et al. (2015)]. Therefore their values are perturbed by 0.1°C accordingly as their alternative options.

h. SSTA calculation

In an earlier version ERSSTv3b, bin averaged SSTs were calculated first on a regular $2^{\circ} \times 2^{\circ}$ grid, and then SSTAs were calculated as the differences between SST and its climatological mean over 1971–2000. In ERSSTv4 and ERSSTv5, the SSTAs at in situ locations are first calculated between SSTs and SST climatological mean at these locations, and then SSTAs are bin-averaged to a $2^{\circ} \times 2^{\circ}$ grid. The order of operations can have an impact as indicated in Huang et al. (2015). These two options of SSTA methods are used for parametric uncertainty estimation.

i. NMAT for ship SST bias adjustment

In ERSSTv3b and ERSSTv4, the nighttime marine temperature (NMAT) is used to calculate ship SST

biases (Huang et al. 2015). In ERSSTv3b, an earlier version of the UKMO NMAT is used, while HadNMAT2 (Rayner et al. 2003) is used in ERSSTv4 and ERSSTv5. SST biases are calculated by fitting the biases to a global climatological difference between SST and NMAT. However, tests showed SST biases may change if they are fitted to regional climatological modes, say a $25^{\circ} \times 25^{\circ}$ running domain (Kent et al. 2017). Therefore, bias uncertainty is taken into account by including options of using different NMATs and fitting area. In the uncertainty estimation of ERSSTv4, the fitting within different latitudinal belts is assigned as an additional source of uncertainty, which is not considered in ERSSTv5 due to potentially a large meridional gradient near the boundary between two latitudinal belts.

j. Ship SST bias smoothing

To reduce the impacts of noise at short time scales, a low-frequency filter (lowess filter of coefficient $f = 0.10$; equivalent to 16-yr low-pass filter; Cleveland 1981) is applied to the fitting coefficient of ship SST biases in ERSST [see details in Huang et al. (2015)]. In pursuing a full bias uncertainty, additional options of linear fitting and annually averaged filtering are also considered. Alternative filters of $f = 0.05$ and 0.15 are included in ERSSTv4 uncertainty, but not included in ERSSTv5 uncertainty due to their similarity to that of $f = 0.10$. Instead, a combined filter of linear–lowess is incorporated into ERSSTv5 (Huang et al. 2017).

k. Ship SST bias readjustment

The biases of ship SSTs over 1854–2016 are initially calculated using NMAT as a reference (Huang et al. 2015). The biases of ship SSTs are also assessed by more accurate buoy SSTs over 1980–2015 (Huang et al. 2017). It was found that there is a systematic offset of 0.077°C between the biases relative to NMAT and buoy SST over 1990–2010. Therefore, the bias relative to NMAT was readjusted by the offset so that it is consistent with the one derived from the buoy SST over 1985–2015. The offset of 0.062° , 0.077° , and 0.092°C are used in ERSSTv5 uncertainty estimation.

l. Argo SST adjustment

The SSTs from buoys and Argo floats are generally consistent. Their averaged difference between 1990 and 2010 is approximately 0.03°C with a RMSD of 0.03°C over the global oceans (Huang et al. 2017). Therefore, the differences of 0.0° , 0.03° , and 0.06° are used to assess its contribution to ERSSTv5 uncertainty.

m. Buoy and Argo SST weighting

An earlier study (Reynolds and Smith 1994) indicated that the random error variance of buoy observations is about 6.8 times smaller than that of ship observations. Therefore buoy observations are weighted by 6.8 when they are merged with ship observations. The same weighting is assigned to Argo observations that have similar measurement quality to buoys and the same type of temperature sensor (Huang et al. 2017). Alternative weightings are set to 5.8 and 7.8. It should be noted that parameters 15 and 16 may not be completely independent of parameters 7–9, and therefore the uncertainty derived from these parameters may slightly be underestimated.

n. Maximum observation number

The superobs on $2^\circ \times 2^\circ$ grids are calculated by averaging SST observations from ships, buoys, and Argo floats weighted by their number of observations. To protect from the averaged superobs being overwhelmed by a single densely observed grid box, a maximum number of observations is set to 10 in ERSST. Its impact on parametric uncertainty is considered by alternative numbers of 5 and 15.

o. Minimum number of months for annual average

In constructing LF anomaly, an annual average is calculated first. The minimum number of months with available monthly SST data is set to 2 months to calculate an annual average in ERSST. Alternative numbers are set to 1 and 3 months.

p. Minimum ratio of superobs

In reconstructing LF anomaly, a $26^\circ \times 26^\circ$ spatial running mean filter is applied to the superobs merged from ships, buoys, and Argo floats on $2^\circ \times 2^\circ$ grids. In the grids where the superobs are labeled as missing, the missing value is replaced by the averaged superobs within a $26^\circ \times 26^\circ$ subdomain, if the area coverage of the superobs within the subdomain is greater than 0.03 (five valid superobs versus a maximum of 169 grids). In estimating parametric uncertainty, alternative coverages are set to 0.02 and 0.04. The area of $26^\circ \times 26^\circ$ in the LF filter is not perturbed, since tests showed that the changes in LF anomaly are very slight as discussed in Huang et al. (2016).

q. Minimum number of years for LF filter

In constructing LF component of annually averaged SSTA, a median filter of 11–19 years is applied in ERSST. The LF component of SSTA is only valid if the number of annually averaged SSTA is more than

two years within the LF period window. Alternative minimum numbers of 1, 2, and 3 years are used to include its contribution to ERSSTv5 uncertainty.

r. LF filter period

In ERSST, SSTAs are decomposed into LF and HF components. The LF component is constructed by applying a median 15-yr filter to annually averaged SSTAs. The LF period is perturbed among 11, 15, and 19 years to include its potential contribution to SST uncertainty. The reason to define the LF component in ERSST reconstruction is to reasonably retrieve the interannual variations (HF) so that they can be reconstructed by EOT modes. Therefore a 15-yr period has been used in ERSST reconstruction. The LF period can be perturbed but its low bound should be longer than a decade. Therefore 11 years is selected as its low bound, and 19 years is selected as its high bound to make 15 years in the middle of the low and high bounds.

s. HF filter period

In ERSST, the HF component of SSTA is filtered using a 3-month running filter to account for missing superobs. An alternative option without the filter (i.e., 1-month filter only) is added to quantify its impact on SST uncertainty. The reason to use a 3-month filter is to take advantage of large heat capacity of water and therefore a high lag-1 autocorrelation of SST (approximately 0.77 in the global ocean). Therefore, we may reasonably interpolate (average) the current month SST when it is missing using SSTs in the previous and next months. Strictly speaking, the interpolated superobs are not actual superobs, and therefore a 1-month filter is used to keep the superobs in $2^\circ \times 2^\circ$ grids as is.

t. EOTs

In ERSST, HF SSTAs are decomposed with EOTs to filter out high-order noise. EOTs were calculated using monthly OISST derived from weekly OISST v2 from 1982 and 2005 in ERSSTv3b, but from 1982 to 2011 in ERSSTv4 and ERSSTv5. The maximum number of EOTs is 130 in ERSSTv3b and 140 in ERSSTv4 and ERSSTv5. As shown by Huang et al. (2015), the selection of EOT training periods leads to a slightly different SSTA reconstruction, particularly in the tropical oceans. Therefore, several groups of EOTs are derived: 1) EOTs from three alternative training periods (1982–2005, 1988–2011, 1982–2011) and 2) EOTs from even-year data (1982, 1984, . . . , 2012) and odd-year data (1983, 1984, . . . , 2013).

The EOTs in ERSST are localized empirical orthogonal functions (EOFs) by damping the modes to zero 4000 (3000) km in longitude (latitude) away from the

center of a mode. Damping scales of 5000, 4000, and 3000 km in longitude, and 4000, 3000, and 2000 km in latitude are used to explore the effects of domain truncation in EOTs.

u. EOT weighting

In fitting HF SSTAs, an EOT mode is weighted by grid box area in ERSSTv3b. Additional weighting of observation number and its associated error is considered in ERSSTv4 and ERSSTv5 (Huang et al. 2015). Therefore, these two weighting options are used in parametric uncertainty estimation.

v. EOT acceptance value

Not all 140 EOT modes are actually accepted to reconstruct HF component of SSTAs. An EOT mode is accepted if EOT acceptance value (Huang et al. 2015) is higher than a certain criterion. The EOT acceptance value assesses whether a particular EOT mode is supported by observations or is potentially an artifact. Huang et al. (2015) showed that the acceptance value is sensitive in determining the resulting SSTA reconstruction. The acceptance value is set to 0.2 in ERSSTv3b and is set to 0.1 in ERSSTv4 and ERSSTv5. Three alternative options of 0.05, 0.1, and 0.2 are set for parametric uncertainty estimation.

w. Ice concentration factor

The ice concentration from HadISST2 (1870–2015; Titchner and Rayner 2014) is used in ERSSTv5, which is close to HadISST1 ice (Rayner et al. 2003) in the Northern Hemisphere and 5%–10% higher in the Southern Hemisphere (Huang et al. 2017). The difference between these two versions of ice concentration data may imply a measure of uncertainty in observing ice concentration. Therefore, ice concentration is alternated by multiplying a factor of 0.9, 1.0, and 1.1.

x. Minimum/maximum ice for SST adjustment

In ERSST, the combined SST from LF and HF components is adjusted in the ice-covered area when the ice concentration falls between a minimum and maximum of 0.6 and 0.9, respectively (Reynolds et al. 2002; Smith et al. 2008). These minimum and maximum values are perturbed by 0.1 as their alternative options.

APPENDIX B

LSAT Internal Parameters and Their Options

There are five explicit parameters in reconstructing LSAT over the global land surface (Table 3). Other parameters such as observation homogenization, biases,

and random errors are included implicitly in 100-member GHCNm v4 datasets (Menne et al. 2018). Explicit parameters in LSAT are assigned by 3 to 7 optional values (Table 3). These parameters and GHCNm options are used for parametric uncertainty estimation. The details of these parameters and their options in LSAT are described as follows.

a. GHCNm data

The 100-member ensemble of GHCNm (Menne et al. 2018) is randomly selected as an internal parameter to assess the parametric uncertainty of LSAT reconstruction.

b. Minimum number of months for annual average

In constructing LF anomaly, an annual average is calculated with a minimum number of months of available monthly LSAT. The minimum number is set to 2 months with an alternative numbers of 1 and 3 months.

c. LF filter periods

In LSAT reconstruction, anomalies are decomposed into LF and HF components. The LF component is constructed by applying a median 15-yr filter to annually averaged LSAT anomalies. LF periods are perturbed among 11, 15, and 19 years to include the potential contribution to LSAT uncertainty.

d. Minimum number of years for the LF filter

In constructing the LF component of annually averaged LSAT anomalies, a median filter of 11–19 years is applied. The LF component of LSAT is only valid if the number of annually averaged LSAT anomalies is more than two years within the LF period. Alternative minimum number of years of 1, 2, and 3 are used to include its contribution to LSAT uncertainty.

e. EOT training periods and spatial scales

In reconstructing LSAT, HF LSAT anomalies are decomposed with EOTs to filter out small-scale noise. EOTs were calculated using monthly ERA-40 from 1971 and 2000 in ERSSTv3b, but using monthly ERA-Interim from 1982 to 2011. The maximum number of EOTs is 65. Several groups of EOTs are derived and randomly selected to assess the parametric uncertainty of LSAT: 1) EOTs from three alternative training periods (1982–2005, 1988–2011, 1982–2011) and 2) EOTs from even-year data (1982, 1984, . . . , 2012) and odd-year data (1983, 1984, . . . , 2013).

The EOTs in reconstructing LSAT are localized EOFs by damping the modes to zero 4000 (2000) km in longitude (latitude) away from the center of a mode. Damping scales of 5000, 4000, and 3000 km in longitude,

and 3000, 2000, and 1000 km in latitude are used to explore the effects of domain truncation in EOTs.

f. EOT acceptance value

Not all 65 EOT modes are actually accepted to reconstruct the HF component of LSAT anomalies. The acceptance value (refer to parameter 25 in appendix A) to select an EOT is set to 0.2. Three alternative options of 0.15, 0.20, and 0.25 are set for parametric uncertainty estimation.

REFERENCES

- Arora, V. K., and Coauthors, 2011: Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophys. Res. Lett.*, **38**, L05805, <https://doi.org/10.1029/2010GL046270>.
- Blunden, J., D. S. Arndt, and G. Hartfield, Eds., 2018: State of the Climate in 2017. *Bull. Amer. Meteor. Soc.*, **99** (8), S1–S310, <https://doi.org/10.1175/2018BAMSStateoftheClimate.1>.
- Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.*, **111**, D12106, <https://doi.org/10.1029/2005JD006548>.
- Cleveland, W. S., 1981: LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *Amer. Stat.*, **35**, 54–65, <https://doi.org/10.2307/2683591>.
- Collins, W. J., and Coauthors, 2008: Evaluation of the HadGEM2 model. Hadley Centre Tech. Note 74, Met Office, 47 pp.
- Cowan, K., and R. G. Way, 2014: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quart. J. Roy. Meteor. Soc.*, **140**, 1935–1944, <https://doi.org/10.1002/qj.2297>.
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, <https://doi.org/10.1002/qj.828>.
- Donlon, C. J., M. Martin, J. Stark, J. Roberts-Jones, E. Fiedler, and W. Wimmer, 2015: The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sens. Environ.*, **115**, 140–158, <https://doi.org/10.1016/j.rse.2010.10.017>.
- Dunne, J. P., and Coauthors, 2012: GFDL's ESM2 global coupled climate-carbon Earth system models. Part I: Physical formulation and baseline simulation characteristics. *J. Climate*, **25**, 6646–6665, <https://doi.org/10.1175/JCLI-D-11-00560.1>.
- Folland, C. K., and D. E. Parker, 1995: Correction of instrumental biases in historical sea surface temperature data. *Quart. J. Roy. Meteor. Soc.*, **121**, 319–367, <https://doi.org/10.1002/qj.49712152206>.
- , and Coauthors, 2001: Global temperature change and its uncertainties since 1861. *Geophys. Res. Lett.*, **28**, 2621–2624, <https://doi.org/10.1029/2001GL012877>.
- Freeman, E., and Coauthors, 2017: ICOADS Release 3.0: A major update to the historical marine climate record. *Int. J. Climatol.*, **37**, 2211–2232, <https://doi.org/10.1002/joc.4775>.
- Fyfe, J. C., and Coauthors, 2016: Making sense of the early-2000s warming slowdown. *Nat. Climate Change*, **6**, 224–228, <https://doi.org/10.1038/nclimate2938>.
- Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, <https://doi.org/10.1029/2010RG000345>.
- Huang, B., and Coauthors, 2015: Extended Reconstructed Sea Surface Temperature version 4 (ERSST.v4), Part I. Upgrades and intercomparisons. *J. Climate*, **28**, 911–930, <https://doi.org/10.1175/JCLI-D-14-00006.1>.
- , and Coauthors, 2016: Further exploring and quantifying uncertainties for Extended Reconstructed Sea Surface Temperature (ERSST) version 4 (v4). *J. Climate*, **29**, 3119–3142, <https://doi.org/10.1175/JCLI-D-15-0430.1>.
- , and Coauthors, 2017: Extended Reconstructed Sea Surface Temperature version 5 (ERSSTv5), Upgrades, validations, and intercomparisons. *J. Climate*, **30**, 8179–8205, <https://doi.org/10.1175/JCLI-D-16-0836.1>.
- , W. Angel, T. Boyer, L. Cheng, G. Chepurin, E. Freeman, C. Liu, and H.-M. Zhang, 2018: Evaluating SST analyses with independent ocean profile observations. *J. Climate*, **31**, 5015–5030, <https://doi.org/10.1175/JCLI-D-17-0824.1>.
- , C. Liu, G. Ren, H.-M. Zhang, and L. Zhang, 2019: The role of buoy and Argo observations in two SST analyses in the global and tropical Pacific oceans. *J. Climate*, **32**, 2517–2535, <https://doi.org/10.1175/JCLI-D-18-0368.1>.
- IPCC, 2013: *Climate Change 2013: The Physical Science Basis*. T. F. Stocker et al., Eds., Cambridge University Press, 1535 pp., <https://doi.org/10.1017/CBO9781107415324>.
- Ishihara, K., 2006: Calculation of global surface temperature anomalies with COBE-SST (in Japanese). *Wea. Serv. Bull.*, **73** (special issue), S19–S25.
- Jones, P. D., T. J. Osborn, and K. R. Briffa, 1997: Estimating sampling errors in large-scale temperature averages. *J. Climate*, **10**, 2548–2568, [https://doi.org/10.1175/1520-0442\(1997\)010<2548:ESEILS>2.0.CO;2](https://doi.org/10.1175/1520-0442(1997)010<2548:ESEILS>2.0.CO;2).
- , D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice, 2012: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *J. Geophys. Res.*, **117**, D05127, <https://doi.org/10.1029/2011JD017139>.
- Karl, T. R., and Coauthors, 2015: Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, **348**, 1469–1472, <https://doi.org/10.1126/science.aaa5632>.
- Kennedy, J. J., 2014: A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Rev. Geophys.*, **52**, 1–32, <https://doi.org/10.1002/2013RG000434>.
- , N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011a: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *J. Geophys. Res.*, **116**, D14103, <https://doi.org/10.1029/2010JD015218>.
- , —, —, —, and —, 2011b: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *J. Geophys. Res.*, **116**, D14104, <https://doi.org/10.1029/2010JD015220>.
- , —, C. P. Atkinson, and R. E. Killick, 2019: An ensemble data set of sea surface temperature change from 1850: The Met Office Hadley Centre HadSST.4.0.0.0 data set. *J. Geophys. Res.*, **124**, 7719–7763, <https://doi.org/10.1029/2018JD029867>.
- Kent, E. C., and P. G. Challenor, 2006: Toward estimating climatic trends in SST. Part II: Random errors. *J. Atmos. Oceanic Technol.*, **23**, 476–486, <https://doi.org/10.1175/JTECH1844.1>.
- , N. A. Rayner, D. I. Berry, M. Saunby, B. I. Moat, J. J. Kennedy, and D. E. Parker, 2013: Global analysis of night marine air temperature and its uncertainty since 1880: The HadNMAT2 data set. *J. Geophys. Res.*, **118**, 1281–1298, <https://doi.org/10.1002/JGRD.50152>.

- , and Coauthors, 2017: A call for new approaches to quantifying biases in observations of sea surface temperature. *Bull. Amer. Meteor. Soc.*, **98**, 1601–1616, <https://doi.org/10.1175/BAMS-D-15-00251.1>.
- Ku, H. H., 1966: Notes on the use of propagation of error formulas. *J. Res. Nat. Bureau Standards*, **70C**, 263–273, <https://doi.org/10.6028/jres.070c.025>.
- Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie, 2011: An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *J. Geophys. Res.*, **116**, D19121, <https://doi.org/10.1029/2011JD016187>.
- Lenssen, N. J. L., G. A. Schmidt, J. E. Hansen, M. J. Menne, A. Persin, R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP uncertainty model. *J. Geophys. Res. Atmos.*, **124**, 6307–6326, <https://doi.org/10.1029/2018JD029522>.
- Lewandowsky, S., J. Risbey, and N. Oreskes, 2016: The “pause” in global warming: Turning a routine fluctuation into a problem for science. *Bull. Amer. Meteor. Soc.*, **97**, 723–733, <https://doi.org/10.1175/BAMS-D-14-00106.1>.
- Medhaug, I., M. Stolpe, E. Fischer, and R. Knutti, 2017: Reconciling controversies about the ‘global warming hiatus.’ *Nature*, **545**, 41–47, <https://doi.org/10.1038/nature22315>.
- Menne, M. J., and C. N. Williams, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate*, **22**, 1700–1717, <https://doi.org/10.1175/2008JCLI2263.1>.
- , I. Durre, R. S. Vose, B. E. Gleason, and T. G. Houston, 2012: An overview of the Global Historical Climatology Network–Daily Database. *J. Atmos. Oceanic Technol.*, **29**, 897–910, <https://doi.org/10.1175/JTECH-D-11-00103.1>.
- , C. N. Williams, B. E. Gleason, J. J. Rennie, and J. H. Lawrimore, 2018: The Global Historical Climatology Network monthly temperature dataset, version 4. *J. Climate*, **31**, 9835–9854, <https://doi.org/10.1175/JCLI-D-18-0094.1>.
- Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, <https://doi.org/10.1029/2011JD017187>.
- Rahmstorf, S., G. Foster, and N. Cahill, 2017: Global temperature evolution: Recent trends and some pitfalls. *Environ. Res. Lett.*, **12**, 054001, <https://doi.org/10.1088/1748-9326/aa6825>.
- Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, <https://doi.org/10.1029/2002JD002670>.
- Rennie, J. J., and Coauthors, 2014: The international surface temperature initiative global land surface databank: Monthly temperature data release description and methods. *Geosci. Data J.*, **1**, 75–102, <https://doi.org/10.1002/gdj3.8>.
- Reynolds, R. W., and T. M. Smith, 1994: Improved global sea surface temperature analyses using optimum interpolation. *J. Climate*, **7**, 929–948, [https://doi.org/10.1175/1520-0442\(1994\)007<0929:IGSSTA>2.0.CO;2](https://doi.org/10.1175/1520-0442(1994)007<0929:IGSSTA>2.0.CO;2).
- , N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis. *J. Climate*, **15**, 1609–1625, [https://doi.org/10.1175/1520-0442\(2002\)015<1609:AIISSAS>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<1609:AIISSAS>2.0.CO;2).
- Rohde, R., and Coauthors, 2013a: A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinfo. Geostat. Overview*, **1**, 1, <https://doi.org/10.4172/2327-4581.1000101>.
- , and Coauthors, 2013b: Berkeley Earth Temperature averaging process. *Geoinfo. Geostat. Overview*, **1**, 2, <https://doi.org/10.4172/GIGS.1000103>.
- Smith, T. M., and R. W. Reynolds, 2003: Extended reconstruction of global sea surface temperature based on COADS data (1854–1997). *J. Climate*, **16**, 1495–1510, <https://doi.org/10.1175/1520-0442-16.10.1495>.
- , and —, 2004: Improved extended reconstruction of SST (1854–1997). *J. Climate*, **17**, 2466–2477, [https://doi.org/10.1175/1520-0442\(2004\)017<2466:IEROS>2.0.CO;2](https://doi.org/10.1175/1520-0442(2004)017<2466:IEROS>2.0.CO;2).
- , and —, 2005: A global merged land–air–sea surface temperature reconstruction based on historical observations (1880–1997). *J. Climate*, **18**, 2021–2036, <https://doi.org/10.1175/JCLI3362.1>.
- , —, T. C. Peterson, and J. Lawrimore, 2008: Improvements to NOAA’s historical merged land–ocean surface temperature analysis (1880–2006). *J. Climate*, **21**, 2283–2296, <https://doi.org/10.1175/2007JCLI2100.1>.
- Stark, J. D., C. J. Donlon, M. J. Martin, and M. E. McCulloch, 2007: OSTIA: An operational, high resolution, real time, global sea surface temperature analysis system. *Proc. Oceans ‘07 IEEE Conf. (MarineChallenges: Coastline to Deep Sea)*, Aberdeen, United Kingdom, IEEE, <https://doi.org/10.1109/OCEANSE.2007.4302251>.
- Titchner, H. A., and N. A. Rayner, 2014: The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. *J. Geophys. Res. Atmos.*, **119**, 2864–2889, <https://doi.org/10.1002/2013JD020316>.
- Trewin, B., 2010: Exposure, instrumentation, and observing practice effects on land temperature measurements. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 490–506, <https://doi.org/10.1002/WCC.46>.
- USGCRP, 2017: Climate Science Special Report: Fourth National Climate Assessment, Volume I, D. J. Wuebbles et al., Eds., U.S. Global Change Research Program, 470 pp., <https://doi.org/10.7930/J0J964J6>.
- van den Dool, H. M., S. Saha, and A. Johansson, 2000: Empirical orthogonal teleconnections. *J. Climate*, **13**, 1421–1435, [https://doi.org/10.1175/1520-0442\(2000\)013<1421:EOT>2.0.CO;2](https://doi.org/10.1175/1520-0442(2000)013<1421:EOT>2.0.CO;2).
- von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.
- Vose, R., and Coauthors, 2012: NOAA’s merged land–ocean surface temperature analysis. *Bull. Amer. Meteor. Soc.*, **93**, 1677–1685, <https://doi.org/10.1175/BAMS-D-11-00241.1>.
- Woodruff, S. D., H. F. Diaz, J. D. Elms, and S. J. Worley, 1998: COADS Release 2 data and metadata enhancements for improvements of marine surface flux fields. *Phys. Chem. Earth*, **23**, 517–526, [https://doi.org/10.1016/S0079-1946\(98\)00064-0](https://doi.org/10.1016/S0079-1946(98)00064-0).
- Zhang, H.-M., and Coauthors, 2019: Updated temperature data give a sharper view of climate trends. *Eos, Trans. Amer. Geophys. Union*, **100**, <https://doi.org/10.1029/2019EO128229>.