Further Exploring and Quantifying Uncertainties for Extended Reconstructed Sea Surface Temperature (ERSST) Version 4 (v4)

BOYIN HUANG

National Centers for Environmental Information, Asheville, North Carolina

PETER W. THORNE

Department of Geography, Maynooth University, Maynooth, Ireland

THOMAS M. SMITH

NOAA/STAR/SCSB and CICS/ESSIC University of Maryland, College Park, Maryland

WEI LIU

Scripps Institute of Oceanography, University of California, San Diego, La Jolla, California

JAY LAWRIMORE, VIVA F. BANZON, HUAI-MIN ZHANG, THOMAS C. PETERSON, AND
MATTHEW MENNE

National Centers for Environmental Information, Asheville, North Carolina

ABSTRACT

The uncertainty in Extended Reconstructed SST (ERSST) version 4 (v4) is reassessed based upon 1) reconstruction uncertainties and 2) an extended exploration of parametric uncertainties. The reconstruction uncertainty (U_r) results from using a truncated (130) set of empirical orthogonal teleconnection functions (EOTs), which yields an inevitable loss of information content, primarily at a local level. The U_r is assessed based upon 32 ensemble ERSST.v4 analyses with the spatially complete monthly Optimum Interpolation SST product. The parametric uncertainty (U_p) results from using different parameter values in quality control, bias adjustments, and EOT definition etc. The U_p is assessed using a 1000-member ensemble ERSST.v4 analysis with different combinations of plausible settings of 24 identified internal parameter values. At the scale of an individual grid box, the SST uncertainty varies between 0.3° and 0.7°C and arises from both U_r and U_p. On the global scale, the SST uncertainty is substantially smaller (0.03°–0.14°C) and predominantly arises from U_p. The SST uncertainties are greatest in periods and locales of data sparseness in the nineteenth century and relatively small after the 1950s. The global uncertainty estimates in ERSST.v4 are broadly consistent with independent estimates arising from the Hadley Centre SST dataset version 3 (HadSST3) and Centennial Observation-Based Estimates of SST version 2 (COBE-SST2). The uncertainty in the internal parameter values in quality control and bias adjustments can impact the SST trends in both the long-term (1901–2014) and "hiatus" (2000–14) periods.

1. Introduction

Sea surface temperature (SST) is an essential climate variable (Bojinski et al. 2014) and plays an important role in climate change monitoring and assessment

(Hartmann et al. 2014). Several SST products have been created over the past several decades and used to quantify the historical SST changes over the world's oceans. These products include the Extended Reconstructed SST (ERSST) version 4 (ERSST.v4) (Huang et al. 2015a; Liu et al. 2015) and its earlier versions (Smith et al. 2008; Smith and Reynolds 2003, 2004), the Centennial Observation-Based Estimates of SST version 2 (COBE-SST2; Hirahara et al. 2014), the Hadley

Centre SST dataset version 3 (HadSST3; Kennedy et al. 2011a,b), the Hadley Centre Sea Ice and SST dataset (HadISST; Rayner et al. 2003), the Kaplan SST (Kaplan et al. 1998), and the weekly Optimum Interpolation SST (OISST) version 2 (v2) (Reynolds et al. 2002) and daily OISST v2 (DOISST; Reynolds et al. 2007; Reynolds 2009). These SST products employ in situ observations primarily from ships and increasingly from buoys in recent decades. Some of them also include satellite-based observations from infrared and/or microwave sensors on polar-orbiting platforms for the period since about 1979.

Various intercomparisons have highlighted key differences between these independently produced products, although their long-term linear trends are broadly similar (Huang et al. 2015a; Liu et al. 2015; Hirahara et al. 2014; Kennedy 2014; Kennedy et al. 2011b). SST producers are often asked which product best represents the "true" historical SST for use in a given application (Huang et al. 2013, 2015b). This question cannot be easily answered since all these products contain errors owing to data and metadata limitations, which serve to preclude definitive analyses (Shen et al. 2007, 1998). In particular, SST analyses exhibit uncertainties caused by incomplete and changing sampling in space and time as well as by errors in the SST observations. Errors in SST values may be caused by occasional human mistakes such as misreading the instrument, as well as by shifts in systematic biases resulting from differences and changes in the types of instruments and measurement protocols. Therefore, SST analyses are affected by the chosen data quality control procedures, bias adjustments, gridding, interpolation, and other analysis methodologies. To understand the resultant datasets and the practical significance of any differences, uncertainty estimates for each product analysis are needed (Kennedy 2014).

The SST uncertainties are usually quantified on each grid box, or for a regional average, or a global average (e.g., Shen et al. 1998; Folland et al. 2001; Smith and Reynolds 2004; Kennedy et al. 2011a; Morice et al. 2012; Shen et al. 2014; Hirahara et al. 2014; Liu et al. 2015). The uncertainty in globally averaged SST benefits from the cancellation of random or quasi-random sources of error by spatial averaging. Therefore, the uncertainty in globally averaged SST is considerably smaller than the uncertainty at most locations on the grid. For example, the 1-sigma uncertainty owing to random errors of a single ship SST observation is as high as 1.3°C (Reynolds et al. 2002; Kent and Challenor 2006), but the globally averaged SST uncertainty owing to the random errors is substantially less than 0.01°C (see section 3c).

In this study, uncertainty assessments for both local and globally or regionally averaged SSTs are based on an ensemble analysis that substantially extends the initial analysis undertaken by Liu et al. (2015) in two key ways. First, Liu et al. (2015) restricted the parametric uncertainty by considering only the subset of ERSST system parameters modified in going from v3b to v4, whereas the present analysis includes a far greater number of internal parameter choices and their possible values in deriving the expanded parametric uncertainty estimate. Second, the reconstruction uncertainty is included in the present study, whereas this uncertainty was not included in Liu et al. (2015). This uncertainty arises due to the local information content loss that inevitably arises from using a finite number of empirical orthogonal teleconnection (EOT) functions (van den Dool et al. 2000; Smith et al. 2008) to reconstruct the globally complete fields. This source of uncertainty differs from and is independent of those additional uncertainties explored within the parametric ensemble.

The remainder of this paper is structured as follows. The ERSST.v4 (Huang et al. 2015a) analysis system is briefly described in section 2, and its internal parameters and their selected values used to derive the parametric uncertainty estimates are listed in the appendix. The datasets and methodology used in our uncertainty estimation is described in section 3. The uncertainties and their impacts on SST trends are assessed in section 4. Subsequently, comparisons with uncertainties in other SST products are undertaken in section 5. Finally, a summary, conclusions, and discussion are given in section 6.

2. ERSST analysis system

Huang et al. (2015a) developed the monthly ERSST.v4 dataset from 1854 to 2014 based on the eigenfunction expansion methods used in Smith et al. (1996), Smith and Reynolds (2003), and Smith et al. (2008). Readers requiring more in-depth methodological details are encouraged to refer to these precursor papers. The spatial resolution is 2° in longitude and latitude over the global oceans, and the temporal resolution is monthly from 1854 to 2014 in this study. In ERSST.v4, the historical observations are decomposed into low- and high-frequency SST anomalies (SSTAs) relative to the 1971–2000 climatology. The low-frequency (LF) SSTA is constructed as follows: 1) the grid boxes without any historical SSTAs are filled with nearby available SSTAs, and 2) a moving filter of 26° × 26° and then a median filter of 15 yr are applied to the monthly 2° × 2° bin-averaged SSTAs. The filters are designed to filter out variations of high frequencies in time and of small scales in space under the assumption that these constitute small-scale noise. The high-frequency (HF) SSTA, defined as the difference between the original and LF

TABLE 1. ERSST.v4 parameters and their operational and alternative options. In parameter number 17, the even years are 1982, 1984, . . . , 2012; the odd years are 1983, 1985, . . . , 2013.

Parameter	Operational option	Alternative options
1. First-guess	Adjusted ERSST.v3b	Adjusted-; Unadjusted-ERSST.v3b
2. SST STD for QC	OISST v2 (1982–2011)	COADS (1950–79); OISST v2
3. Min SST STD	1.0°C	0.5°C; 1.0°C; 1.5°C
4. Max SST STD	4.5°C	3.5°C; 4.5°C; 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 ships and 0.5°C for buoys
7. Ship SST error	1.3°C	1.2°C; 1.3°C; 1.4°C
8. Buoy SST error	0.5°C	0.4°C; 0.5°C; 0.6°C
9. Ship-buoy SST adj	0.12°C	0.08°C; 0.12°C; 0.16°C
10. Buoy SST weighting	6.8	5.8; 6.8; 7.8
11. SSTA calculation	in situ basis	Grid box basis; in situ basis
12. NMAT for SST bias	HadNMAT2	UKMO NMAT; HadNMAT2; regional HadNMAT2
13. SST bias smoothing	$f = 0.10$	Annual; $f = 0.05$; 0.10; 0.20; linear
14. Min number of months for annual average	2	1; 2; 3
15. Min rate of superobservation	0.03	0.02; 0.03; 0.04
16. Max number of observations	10	5; 10; 15
17. EOT training period and spatial scales	1982–2011 $L_x = 5000$ km and $L_y = 3000$ km;	1982–2005; 1988–2011; 1982–2011; 1982–2011 nondamped in high latitudes; even years from 1982 to 2012; odd years from 1983 to 2013; $L_x = 6000$ km and $L_y = 4000$ km; $L_x = 4000$ km and $L_y = 2000$ km
18. EOT weighting	$W = N/(N + \xi^2)\cos(\varphi)$	$W = \cos(\varphi)$; $W = N/(N + \xi^2)\cos(\varphi)$
19. EOT critical value	0.10	0.05; 0.10; 0.20
20. Ice concentration factor	1.0	0.9; 1.0; 1.1
21. Min ice for SST adj	0.6	0.5; 0.6; 0.7
22. Max ice for SST adj	0.9	0.8; 0.9; 1.0
23. LF filter period	15 yr	11 yr; 15 yr; 19 yr
24. HF filter period	3 month	0; 3 month

SSTAs, is reconstructed by fitting SSTAs on the global domain to the 130 leading EOTs. The EOTs are similar to empirical orthogonal functions, except that the EOTs are restricted in domain to a spatial scale of 5000 and 3000 km in longitude and latitude, respectively. The HF SSTA is then merged with the LF SSTA. SSTs are retrieved by adding the monthly climatology to the SSTA fields. The merged SSTs are adjusted toward the freezing point of −1.8°C (Smith and Reynolds 2004) in proximity to sea ice according to the observed ice concentrations from HadISST (1870–2010; Rayner et al. 2003) and the National Centers for Environmental Prediction (NCEP; 2011–14; Grumbine 2014).

The historical ocean observations used for ERSST.v4 analyses arise from the in situ International Comprehensive Ocean–Atmosphere Dataset (ICOADS) Release 2.5 (R2.5; Woodruff et al. 2011) from 1854 to 2007, and from the Global Telecommunication System (GTS) receipts from NCEP after 2007.

The ICOADS and GTS observations exhibit both random errors and systematic biases (Kennedy et al. 2011a,b). This is why filters and EOT decompositions are used to reduce the effect of the random errors, and bias adjustments are applied to remove the systematic biases in the ERSST.v4 analysis. These processing steps act to smooth out the field under the reasonable assumption that much of the high-frequency/local structure leading to a marked ''spottiness'' in the basic data is likely suspicious given the broad spatial and temporal SST correlation structures in most of the global domain. The use of filters and EOT decomposition, however, will lead to an inevitable loss of information content even if the input data are sound. Their use therefore introduces other potential errors into the SST analysis even if all other methodological aspects of the ERSST processing suite are perfect. These smoothing effects are termed herein the reconstruction uncertainty (see details in section 3b). The SSTs estimated by ERSST.v4 may also vary when different but plausible values of the processing system's internal parameters such as for data quality control and bias adjustments are selected (Table 1; also see the appendix). The SST variations associated with the selection of the internal parameters are referred to herein as the parametric uncertainty (see details in section 3c). A total of 24 internal parameters are identified as a result of uncertain methodological choices and

TABLE 2. Test datasets used for ERSST.v4 reconstruction uncertainty assessment.

SST products	Spatial resolution	Temporal resolution	Data ingest	Analysis method	External forcing
GFDL-ESM2G	1° × 0.9° Global	Daily 1861–2005	N/A	Coupled model simulation	Greenhouse gases, trace gases, aerosols, ozone, land use
HadGEM2-AO	1° × 0.8° Global	Monthly 1860–2006	N/A	Coupled model simulation	Greenhouse gases, aerosols
HadISST	1° × 1° Global	Monthly 1871–2013	In situ SST Satellite SST	EOF-based reduced space optimal interpolation	N/A
MOISST	1° × 1° Global	Monthly 1982–2013	In situ SST Satellite SST	Optimum interpolation	N/A
DOISST	0.25° × 0.25° Global	Daily 1982–2013	In situ SST Satellite SST	Optimum interpolation	N/A

hence potentially contribute to the parametric uncertainty. This is considerably more than that in Liu et al. (2015) where only those nine parameters that were modified in upgrading from ERSST.v3b to ERSST.v4 were considered.

3. Data and methods

a. The test data used to derive uncertainty estimates

The test SST datasets are selected from coupled model simulations and observations. The model estimates are independent and spatially complete analyses of SSTs consistent with the model physics. The observationally based estimates are methodologically independent of ERSST.v4 and make use of satellite data, which are not considered in ERSST.v4. The use of a suite of possible test datasets is necessary for ascertaining whether the estimated uncertainties are sensitive to the selection of test datasets. These selected datasets (Table 2) are the following:

1) The SST data from the coupled simulation of Geophysical Fluid Dynamics Laboratory (GFDL) Earth System Model version 2G (ESM2G; Dunne et al. 2012). The resolution of the SST data is 1° in longitude, near 0.9° in latitude, and daily from 1861 to 2005.
2) The SST data from coupled simulation of the United Kingdom Met Office (UKMO) Hadley Centre Global Environmental Model version 2-AO (HadGEM2-AO; Collins et al. 2008). The resolution of HadGEM2-AO SST data is 1° in longitude, near 0.8° in latitude, and monthly from 1860 to 2006.
3) The SST data from the HadISST analysis (Rayner et al. 2003). The resolution of HadISST is 1° × 1° in space and monthly from 1871 to 2013.
4) The monthly OISST (MOISST) data from 1982 to 2013. The MOISST is derived from weekly OISST v2 (Reynolds et al. 2002) data from NCEP. The weekly data are first interpolated to daily data; and the daily

data are then averaged to monthly data. The spatial resolution is 1° × 1°.
5) The daily SST data from DOISST from 1982 to 2013 (Reynolds et al. 2007). The spatial resolution is 0.25° × 0.25°.

b. Reconstruction uncertainty

Following Shen et al. (2004), the reconstruction uncertainty $U_r(x, y, t)$ for the grid box (x, y) and month t is defined as

$$U_r^2(x,y,t) = [A_f(x,y,t) - D(x,y,t)]^2, \qquad (1)$$

where $D(x, y, t)$ is a spatiotemporally complete test dataset (e.g., a dataset from a climate model or a global analysis), and $A_f(x, y, t)$ is the reconstructed data by using the ERSST.v4 reconstruction method but with the data from $D(x,y,t)$. Since the ERSST.v4 reconstruction system is a smoothing procedure and both $A_f(x, y, t)$ and $D(x, y, t)$ are defined for every grid box, $U_r(x, y, t)$ may be considered to represent a smoothing error. This study used those five test datasets listed in Table 2 as a measure of $D(x, y, t)$.

The EOT decomposition acts to damp out small-scale SST variations and will therefore result in an inevitable loss of information if the ICOADS and GTS data were complete and error free. The reconstruction uncertainty arises within the ERSST analysis because a maximum number (130) of SST EOT modes are used to reconstruct high-frequency component SSTs (Huang et al. 2015a; Smith et al. 2008; Smith and Reynolds 2004). When the input data are sparse, the number of EOTs used in reconstruction may be as low as 80 EOTs. However, the SSTA component explained by 81st to 130th EOTs should have been captured within the parametric uncertainty term, since the lower bound of the acceptance criterion parameter therein is very low (0.05).

The test datasets were regridded to 2° × 2° grids where necessary and used to determine the U_r. We determine

U_r by using MOISST (Table 2) as the "perfect" input and combine with the parametric uncertainty to form the total uncertainty of ERSST.v4 in section 3d. MOISST is selected because 1) it is derived from both in situ and satellite measurements and 2) the U_r using MOISST is similar to those using model test datasets, and higher than that using HadISST (see section 4a).

Alternatively, the U_r may be assessed as $U_r^2(x, y, t) = [A_s(x,y,t) - D(x,y,t)]^2$, where A_s represents the analysis using spatially noncomplete subsampled "observations." However, the U_r using the alternative method may interact with the spatially noncomplete subsampling, and it cannot account for the uncertainty of the SST reconstructed in those areas without observations.

c. Parametric uncertainty

The parametric uncertainty U_p is defined as the standard deviation of reconstructed SSTs due to using different values of parameters in the ERSST.v4:

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

$$\overline{A} = \frac{1}{M} \sum_{m=1}^{M} A_m(x,y,t), \quad (3)$$

where $A_m(x, y, t)$ is a member of reconstruction based upon the mth group of parameters used in the ERSST.v4 (see Table 1 and the appendix for the 24 parameters and their ranges), and $\overline{A}$ is the mean of the all the M reconstructions corresponding to the M groups of parameters. We choose $M = 1000$. These 1000 combinations are randomly picked from among the much larger population of possible parameter combinations ($>2^{24}$ given that several parameters have three or more options).

This is an advance in the completeness of consideration of the term from that in Liu et al. (2015). The major differences between present study and Liu et al. (2015) are in the following four aspects: (a) the total number of parameters considered has been increased from nine to 24, (b) the ranges for those nine original parameters have been increased, (c) all parameter values are selected randomly with equal likelihood without preference, and (d) the number of ensemble members has been increased from 100 to 1000. The parameter options are predefined by perturbing each of 24 parameters by 10%–50% of their operational settings, based on our understanding of these parameters and what constitutes methodologically reasonable perturbations to them. These changes are intended to more fully explore this uncertainty component in ERSST.v4 than in Liu et al. (2015).

The parametric uncertainty in Eq. (2) is associated with the choice of internal parameter settings in the ERSST.v4 analysis system. The analyzed SST deviates slightly when a different value is assigned to a specific parameter (Huang et al. 2015a; Liu et al. 2015). For example, an El Niño event in the tropical Pacific may be better represented when more EOTs are accepted for the analysis. Liu et al. (2015) further showed that most parameters interacted in a nonlinear manner such that the effects of changing two parameters independently tended to differ from the effect of changing them concurrently. This points to the need for the creation of ensemble realizations as has also been done for the HadSST3 product (Kennedy et al. 2011b).

The random errors of the input observations were not considered in ERSST.v4 and its previous versions (Smith et al. 2008; Smith and Reynolds 2003, 2004, 2005). To account for the uncertainty resultant from the random error in the input data, the random error is simulated using a Gaussian random number (GRN) generator. The mean of the random error is set to 0°C, while the standard deviation (STD) of the random error is set to 1.3° and 0.5°C (Reynolds et al. 2002; Kent and Challenor 2006) for a single ship and buoy observation in the analyses of every ensemble member, respectively. Additional testing analyses of the operational ERSST.v4 version with and without including the random error showed that the globally averaged difference of local SST is less than 0.2°C before the 1900s and is less than 0.1°C after the 1960s (not shown). The difference for the global averaged SST is near zero in all times. It is included here because 1) some users require local and not global information and 2) this term may interact with parameters varied within the parametric uncertainty ensemble, and so the resulting ensemble may be underdispersive if it is not included. It needs noting that the correlation of random error among observations by the same ship or buoy is not assessed in the current uncertainty assessment in ERSST.v4 because the ship call signs are incomplete.

d. Total uncertainty

The total uncertainty (U_t) is defined as the standard combination of uncertainty terms U_r and U_p under the assumption of independence (which is trivially true given their respective derivations in sections 3b and 3c):

$$U_t^2 = U_r^2 + U_p^2. \quad (4)$$

The definition of the total uncertainty in Eq. (4) is different from that of ERSST.v3b and other SST products (Table 3). In ERSST.v3b, the total uncertainty consists of sampling uncertainty and bias uncertainty (Smith and

TABLE 3. Uncertainty components in ERSST.v4, ERSST.v3b, HadSST3, and COBE-SST2.

SST products	Uncertainty components explored and quantified	Uncertainty quantification methods and references
ERSST.v4	Parametric	Ensemble analyses using perturbed parameter settings for parametric,
	Reconstruction	fully sampled test data for reconstruction (described herein)
ERSST.v3b	Sampling	Difference between fully and subsampled analyses; difference between ship
		SST bias adjustments in ERSST.v3b and Folland and Parker (1995)
	Bias	Smith and Reynolds (2005)
HadSST3	Measurement with correlation	Ensemble analyses using perturbed parameters (for bias) and additional
		spatial and temporal analyses
	Measurement without correlation	Kennedy et al. (2011a)
	Sampling	
	Bias	
	Coverage	
COBE-SST2	Sampling	Difference between fully and subsampled analyses using observations from
		data abundant period
		Hirahara et al. (2014)

Reynolds 2003, 2004, 2005). The bias uncertainty has now been included as part of the parametric uncertainty as described in Table 1 and the appendix. The sampling uncertainty is not explicitly included in Eq. (4) in the present study because it is accounted for within the parametric term. Further reasoning and justification for this choice is given in section 6.

The uncertainties in Eqs. (1)–(4) are defined for the monthly local SSTs on a grid box basis in space and time (x, y, t). The uncertainties of any regionally averaged SST (e.g., globally averaged SST) are defined in the similar ways shown in Eqs. (1)–(3) except that the analyses $A_f(x, y, t)$, $A_m(x, y, t)$, and the test data $D(x, y, t)$ are first averaged over the regional domain of interest before assessing an uncertainty (Shen et al. 1998):

$$U_r^2(t) = [A_f^g(t) - D^g(t)]^2, \quad (5)$$

$$U_p^2(t) = \frac{1}{M} \sum_{m=1}^{M} [A_m^g(t) - \overline{A^g}(t)]^2, \quad (6)$$

$$\overline{A^g}(t) = \frac{1}{M} \sum_{m=1}^{M} A_m^g(t), \quad (7)$$

where the superscript g represents the global average.

4. Results of quantified uncertainties

a. Reconstruction uncertainty

The reconstruction uncertainty (U_r) associated with the ERSST.v4 analysis is assessed using SST data from MOISST. A set of 32 ERSST.v4 uncertainty analyses is created using 32 years (1982–2013) of MOISST data. These data are likely more faithfully reflecting the true seasonal cycle than model based estimates. Each of the

32 analyses uses 12 months (January–December) of periodic SSTs for each of 32 years of MOISST (1982–2013). The test data of each of 32 ensemble members are ingested to the fully sampled ERSST.v4 analysis [A_f in Eqs. (1) and (5)].

Figure 1a shows the averaged (1871–2005) U_r for local SSTs. The U_r is 0.1°–0.2°C in the tropical Indian Ocean, tropical western Pacific, and tropical Atlantic; 0.2°–0.4°C in the eastern tropical Pacific, northwestern North Pacific, and North Atlantic; and 0.2°–0.3°C in the Southern Ocean south of 30°S, as well as in the Arctic Ocean. The global mean U_r is near 0.27°C (Fig. 2a; dotted orange line). Despite a large U_r for local SSTs, the U_r for globally averaged SST is less than 0.01°C (Fig. 2c, dotted orange line) due to the cancellation of errors by global averaging. This cancellation would be expected if the reconstruction procedure employed were adequate—the information loss during the reconstruction should primarily be small-scale structure and therefore its impact on the large-scale average is expected to be small.

By construction, the U_r of using MOISST is constant in time (seasonal cycle included in data but filtered out in the figure). The reasons are that the periodic January–December MOISST for each of 32 years is used as test datasets for the ensemble analyses, and the test data are taken as the fully sampled ''observations'' over the entire analysis period. However, U_r varies in space as indicated in Fig. 1a. The spatial variations of U_r can further be quantified by the difference among its 10th percentile (0.03°C; Fig. 2b, green dashed line), 50th percentile (0.20°C; Fig. 2b; green solid line), and 90th percentile (0.57°C; Fig. 2b, green dotted line) of individual grid box values.

We selected the U_r using MOISST for estimating the total uncertainty in ERSST.v4 because (a) the MOISST
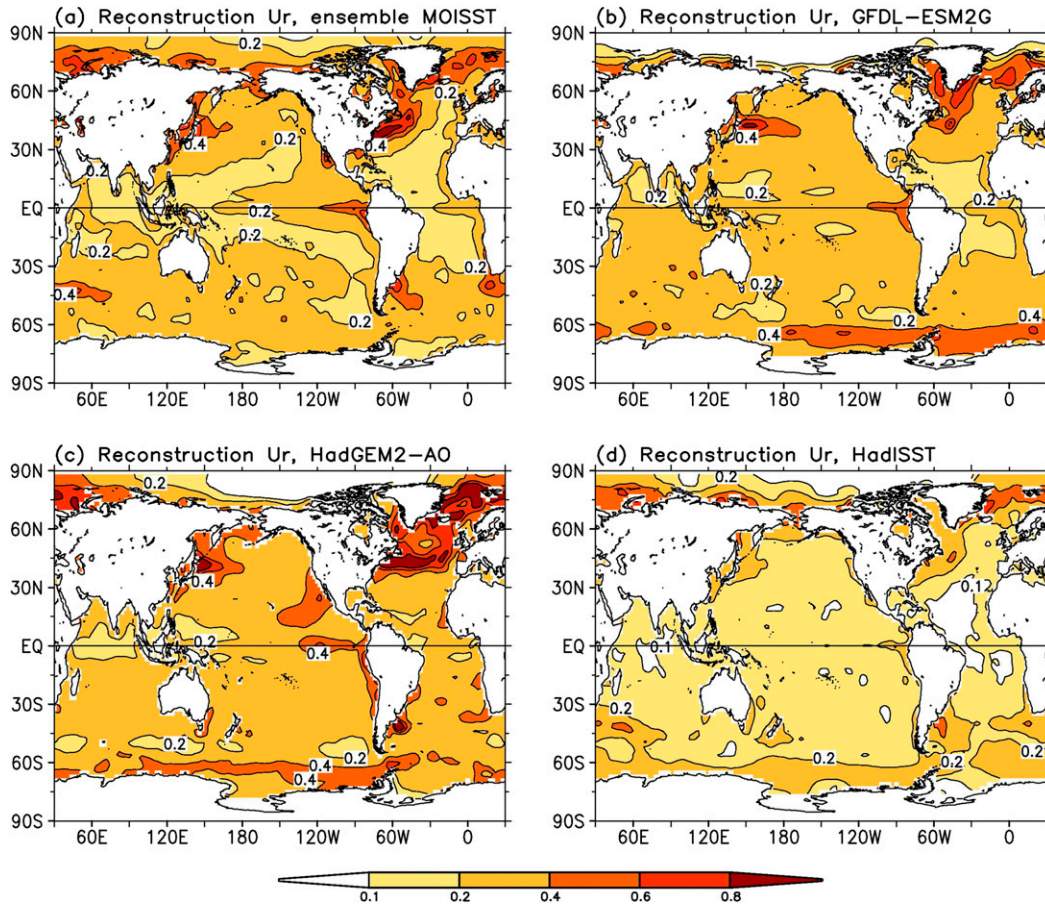
FIG. 1. Averaged (1871–2005) reconstruction uncertainty (1σ) using test data from (a) ensemble MOISST, (b) GFDL-ESM2G, (c) HadGEM2-AO, and (d) HadISST. The averaging period of 1871–2005 is based on the common data period of GFDL-ESM2G, HadGEM2-AO, and HadISST. Contours are 0.1°, 0.2°, 0.4°, 0.6°, and 0.8°C.

data are based on both in situ and satellite observations and may better reproduce the true seasonal cycle and (b) comparisons show that the spatial and temporal variations of U_r using MOISST are similar to those using three other test datasets from GFDL-ESM2G (Figs. 1b, 2a, and 2c), HadGEM2-AO (Figs. 1c, 2a, and 2c), and HadISST (Figs. 1d, 2a, and 2c). The magnitude of U_r using MOISST is similar to that using GFDL-ESM2G and HadGEM2-AO, and approximately 0.1°C larger than that using HadISST. The low U_r using HadISST is possibly because the HadISST is using EOF reconstructions such that the reconstructed SSTs are artificially smooth. This suggests that the U_r is associated with the spatial variability in the SST that cannot easily be resolved under interpolation methods. For example, when the variability of test data from GFDL-ESM2G is reduced by applying a nine-point latitude/longitude smoothing, the U_r reduces by approximately 0.1°C over the world oceans (not shown).

All test cases considered here showed a negligible contribution to the uncertainty of globally averaged SST (Fig. 1c). Thus there is high confidence that this term will make at most a very minor contribution to the global-mean SST uncertainty budget for ERSST.v4. In theory, the U_r of local SST over the global oceans can be reduced by better resolving small-scale variabilities of SSTs if a larger number of EOTs is used. For example, when the number of EOT modes increases from 130 to 260, the globally averaged U_r decreases slightly, from 0.27° to 0.23°C (not shown). But as more EOTs are included there is a risk that EOTs become increasingly driven by residual random and systematic errors in the underlying data and hence that false structures are imparted to the data. This is why U_r is an important aspect of the comprehensive uncertainty budget for local SST analyses. There is a limit to how accurately we can estimate the local SST variations using ERSST.v4 or arguably any other method.
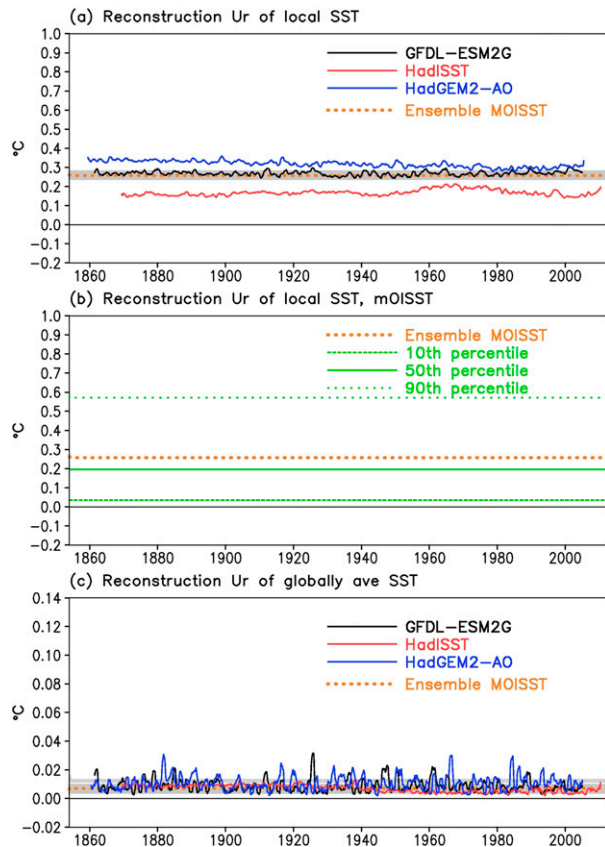
FIG. 2. (a) Globally averaged reconstruction uncertainties of local SST using test data from GFDL-ESM2G (black line), HadISST (red line), HadGEM2-AO (blue line), and ensemble MOISST (dotted orange line). Shaded lines represent uncertainties of ensemble members of MOISST. (b) Ensemble and globally averaged reconstruction uncertainty using test data from MOISST (dotted orange line) in (a) and its 10th (dashed green line), 50th (solid green line), and 90th (dotted green line) percentile. (c) Reconstruction uncertainties of globally averaged SST. The uncertainties are scaled in 1σ. A 12-month running mean is applied in (a)–(c).

b. Parametric uncertainty

1) QUANTIFICATION

The parametric uncertainty (U_p) is defined as the SST STD of the 1000 members from their ensemble average [Eqs. (2) and (3)]. The ensemble and time (1871–2005) averaged U_p (Fig. 3a) is less than 0.2°C in the Arctic Ocean, most of the tropical and North Atlantic, and the Indian Ocean, and is 0.2°–0.4°C in most of the tropical Pacific. The U_p is higher (0.4°–0.8°C) in the North Pacific, northwestern North Atlantic, Southern Ocean between 30° and 60°S, and equatorial and South Atlantic near the coast of Africa. The U_p in these regions is dominated by a small number of parameters related principally to SST bias adjustment, SST quality control, the selection of base-function EOTs, and their acceptance

criterion (Huang et al. 2015a; Liu et al. 2015). This points to the effects of either sparse sampling or strong year-to-year variability (particularly in regional boundary currents) being dominant in determining regional U_p estimation.

The globally averaged U_p for local SSTs (Fig. 3b; red line) is 0.5°–0.6°C before 1880, peaks during the two World Wars, and decreases to approximately 0.2°C after the 1980s. The median U_p (green solid line) is lower than the globally averaged U_p, which is associated with the fact that the grid box distribution is highly skewed with higher grid box U_p mostly confined to the limited regions with sparse observations or strong variability. To assess the spatial variation of U_p, the 90th percentile of U_p is plotted in Fig. 3b (green dotted line). The high values of the 90th percentile indicate that the uncertainty in those regions could be as large as 1°C before the 1880s. In contrast, the 10th percentile U_p is less than 0.1°C (Fig. 3b; green dashed line), which appears mostly in the Arctic, Ross Sea, and Weddell Sea (Fig. 3a). The low U_p in the polar regions may in large part be associated with the lower variability of the SST of water near the freezing point of −1.8°C. We note that there are few active parameters within the ensemble directly or indirectly associated with sea ice and hence the ensemble may be underdispersive here. Ongoing work is considering fundamentally new approaches to the consideration of SSTs in polar regions for future ERSST versions, which may permit better quantification of uncertainty in these regions in future. But at present these are still under development.

Similar to the reconstruction uncertainty, many of the uncertainties in the U_p estimate cancel with regional averaging and hence the globally averaged U_p is smaller (Fig. 3c; red line), being near 0.11°C before 1880, with peaks during the World Wars, and decreasing to less than 0.03°C after the 1950s. The U_p for globally averaged SST is considerably less than the underlying long-term trend of 0.67°C century^{-1} [refer to section 4b(2)] and suggests that the globally averaged SST reconstructed in ERSST.v4 is not overly sensitive to the selection of internal parameter values.

The local U_p assessed in this study (Fig. 3b; red solid line) is approximately 2 times larger than that in the work of Liu et al. (2015), who produced a 100-member ensemble by varying the nine parameters that were changed specifically in upgrading from ERSST.v3b to ERSST.v4 (Fig. 3b; dotted black line). The U_p in globally averaged SST is also approximately 2 times larger in this study (Fig. 3c; red line) than in Liu et al. (2015; Fig. 3c; dotted black line). The potential reasons for the larger U_p in the present study include the following: 1) The number of internal parameters was increased to
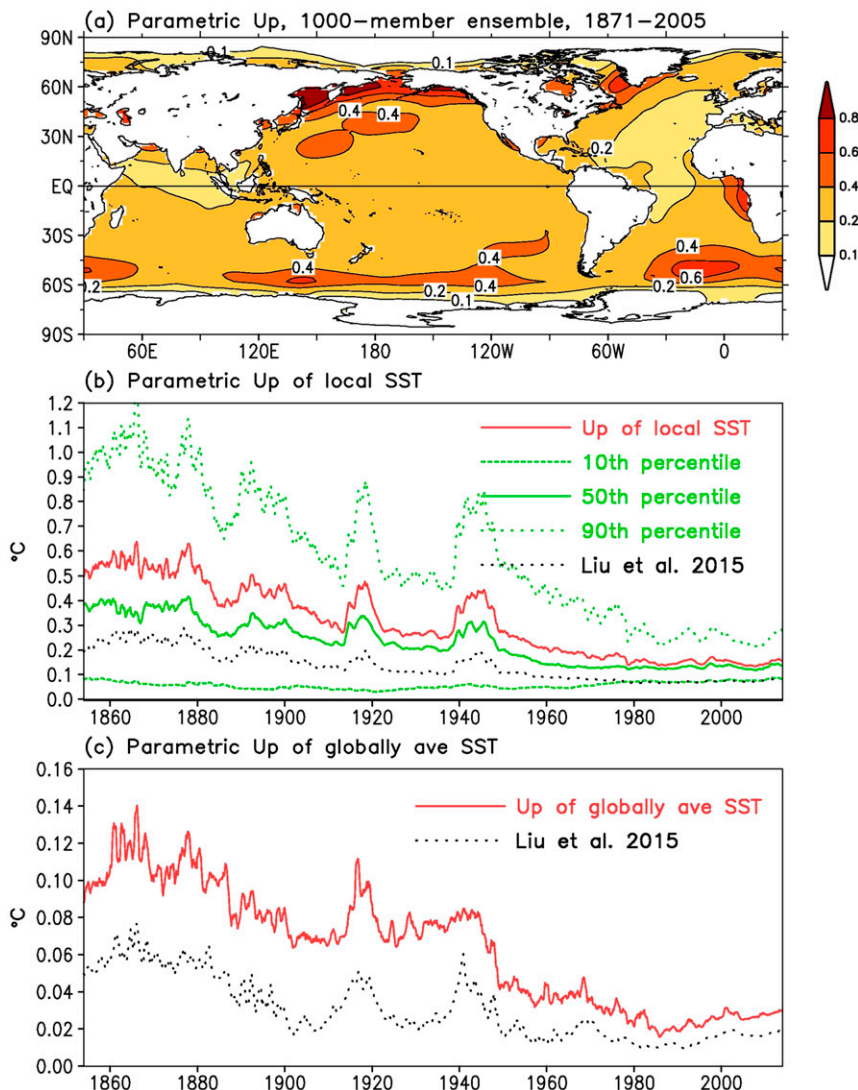
FIG. 3. Parametric uncertainty (1σ) of 1000 ensemble ERSST.v4 analyses. (a) Averaged (1871–2005) uncertainty in contours of 0.1°, 0.2°, 0.4°, 0.6°, and 0.8°C. (b) Globally averaged parametric uncertainty (red line) and its 10th (dashed green line), 50th (solid green line), and 90th (dotted green line) percentiles. (c) Parametric uncertainty (red line) of globally averaged SST. Parametric uncertainty of Liu et al. (2015) is overlapped in (b) and (c) (dotted black lines). A 12-month running mean is applied in (b) and (c).

24 from 9 in Liu et al. (2015), which enables the analysis system to represent the potential uncertainty more completely. 2) The likelihood to select parameter values is equal in this study, while the likelihood is higher to select "operational" parameter values in Liu et al. (2015). 3) The ranges of some of those nine parameter values of Liu et al. (2015) have been increased as further inspection yielded arguments that other, broader, choices for these parameters may be valid. Finally, 4) the ensemble size is increased to 1000 from 100 in Liu et al. (2015), which might not be a dominant contributor to the larger U_p as indicated in the following subsection.

However, the larger ensemble size will, all else being equal, represent the possible solution space more completely.

2) IMPACTS OF ENSEMBLE NUMBERS ON PARAMETRIC UNCERTAINTY AND LONG-TERM SST TRENDS

As shown in Table 1, each of the 24 parameters has more than two options. This implies that at least 2^{24} (approximately 10^7) ensemble analyses are possible. The logical question is how many randomly selected ensemble members are sufficient to get a representative
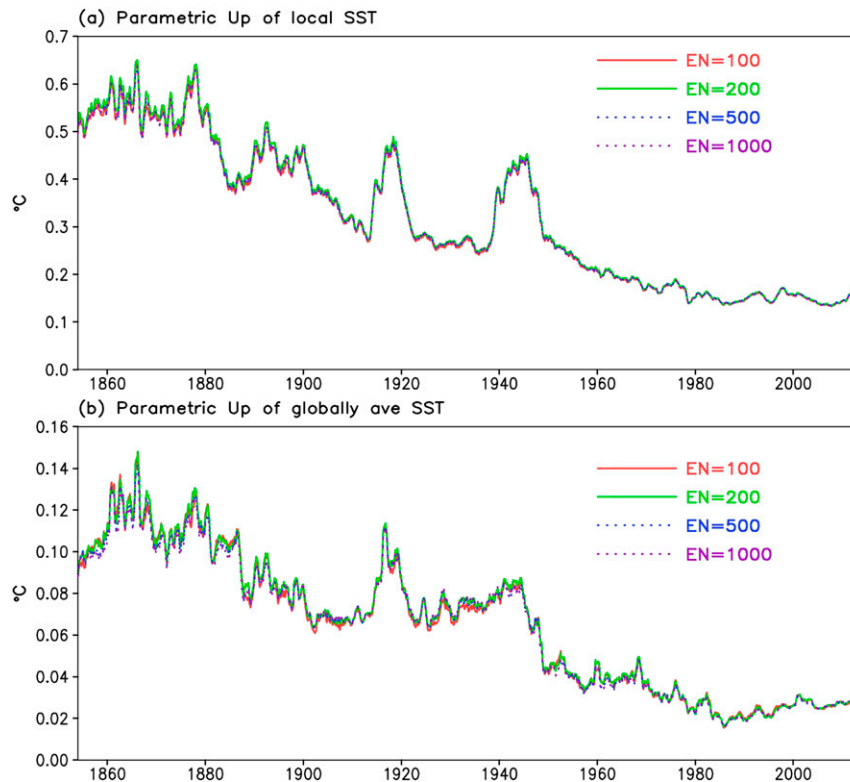
FIG. 4. (a) Globally averaged parametric uncertainty of local SST and (b) parametric uncertainty of globally averaged SST when EN is 100 (red line), 200 (green line), 500 (dotted blue line), and 1000 (dotted purple line).

sample, recognizing that $>2^{24}$ solutions cannot practically be realized. Figure 4 shows, however, that both global averaged U_p of local SST (Fig. 4a) and U_p of globally averaged SST (Fig. 4b) are not very sensitive to the ensemble number (EN) if it is reasonably large. The U_p is almost identical when EN is set to 100, 200, 500, or 1000. In particular, there is virtually no change apparent in going from 500 to 1000 members. This suggests that the U_p estimate is quasi-saturated when EN of 1000 is used in the uncertainty estimate and that further estimates will not serve to greatly alter the findings. If instead the distribution of 1000 members was substantially distinct from that for 500 members, this would imply that a 1000-member ensemble was likely still insufficient to sample fully the plausible U_p and that we could require yet more ensemble members.

The U_p may directly impact the long-term SST trend, which is one of the most important climate change indicators (Karl et al. 2015). Figure 5a shows the ensemble averaged (EN = 1000) SST trend (1901–2014) over the global oceans. The trend is 0.8°–1.0°C century^{-1} in the Southern Ocean between 30° and 60°S, northern Indian Ocean, eastern North Atlantic, and tropical Atlantic; 0.4°–0.6°C century^{-1} in most of the tropical and North

Pacific, northwestern North Atlantic; and less than 0.2°C century^{-1} in the Arctic, North Atlantic south of the Greenland, and along the Antarctic. These trends are mostly significant at the 95% confidence level. The reason for the high confidence level is that the STD of the trends (Fig. 5b) is much smaller than the trend itself, and the degrees of freedom are high (arguably near 1000 since the parameter options of the ensembles were randomly drawn). The STD of SST trends is higher in the northwestern North Pacific, northwestern North Atlantic, and Southern Ocean south of 30°S; and is lower in the Indian Ocean, tropical Atlantic, and Pacific. The spatial distribution of the STD is consistent with the spatial distribution of the U_p shown in Fig. 3a, which represents the impacts of U_p on the SST trends. Furthermore, the SST trends are not very sensitive to the selection of EN. The difference between the SST trends when EN is set to 1000 and 100 is very small (Fig. 5c).

Likewise, the impact of EN on the trend of globally averaged SST is small (Fig. 6a). The ensemble averaged trends of globally averaged SSTs are approximately 0.67°C century^{-1}, which is slightly lower than that in the operational ERSST.v4 (0.69°C century^{-1}); and the range of the SST trends is 0.41°–0.78°C century^{-1} regardless of
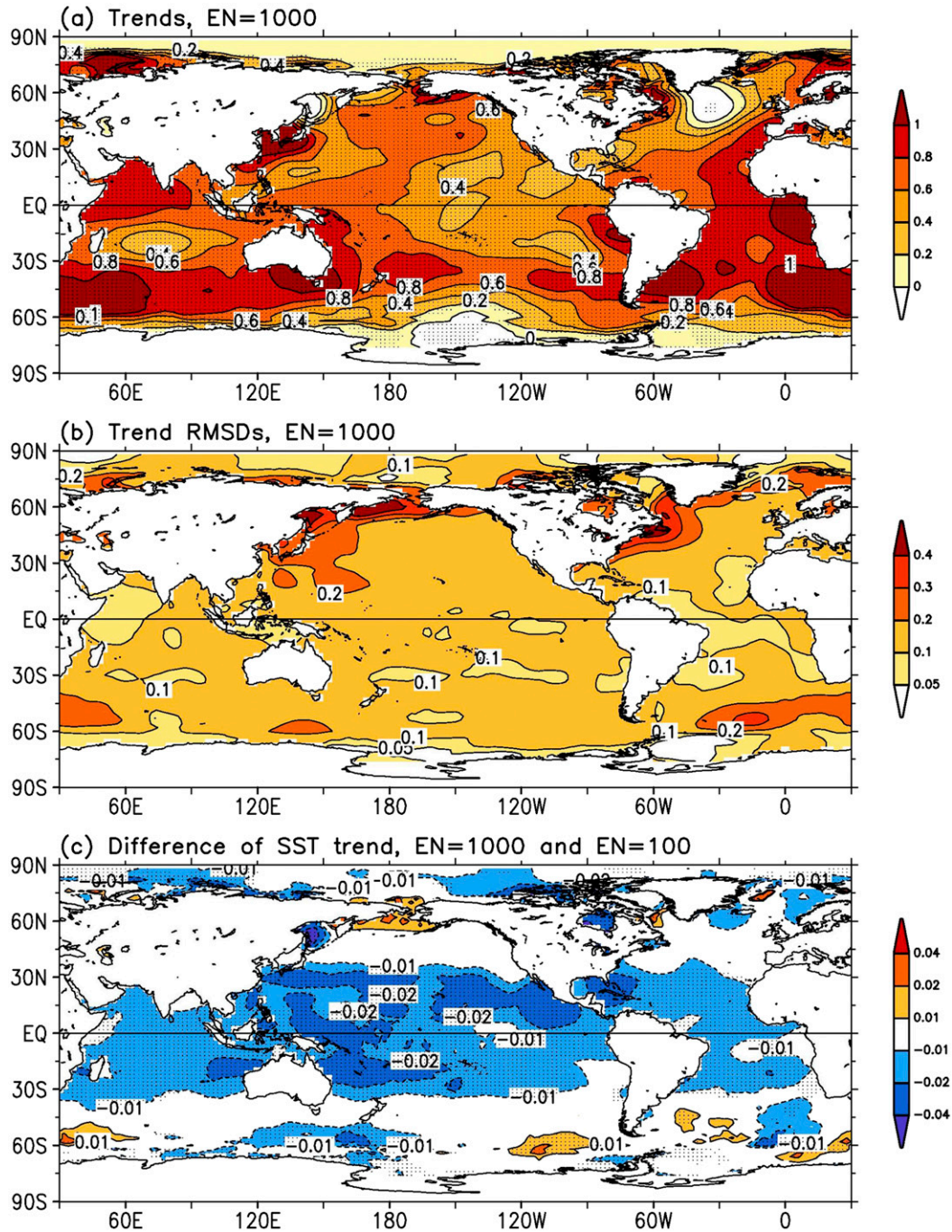
FIG. 5. (a) Ensemble (EN = 1000) averaged SST trends (1901–2014), (b) STD relative to the ensemble average in (a), and (c) the difference of averaged SST trends between EN = 1000 and EN = 100. Contour intervals are 0.2°C century^{-1} in (a); contours are 0.05°, 0.1°, 0.2°, 0.3°, and 0.4°C century^{-1} in (b), and ±0.01°, ±0.02°, and ±0.04°C century^{-1} in (c). The stippled area in (a) and (c) indicates that the trend or trend difference is significant at 95% confidence level.

whether EN is 100, 200, 500, or 1000. The histograms of the SST trends are very similar when EN is larger than 500, indicating that the U_p in global-mean SST trends can be well described using a 1000 ensemble analyses.

The uncertainty estimates herein and in other efforts consider different sources of uncertainty in distinct manners. Given that we do not know the true SSTs they are all relative rather than absolute estimates. That is,
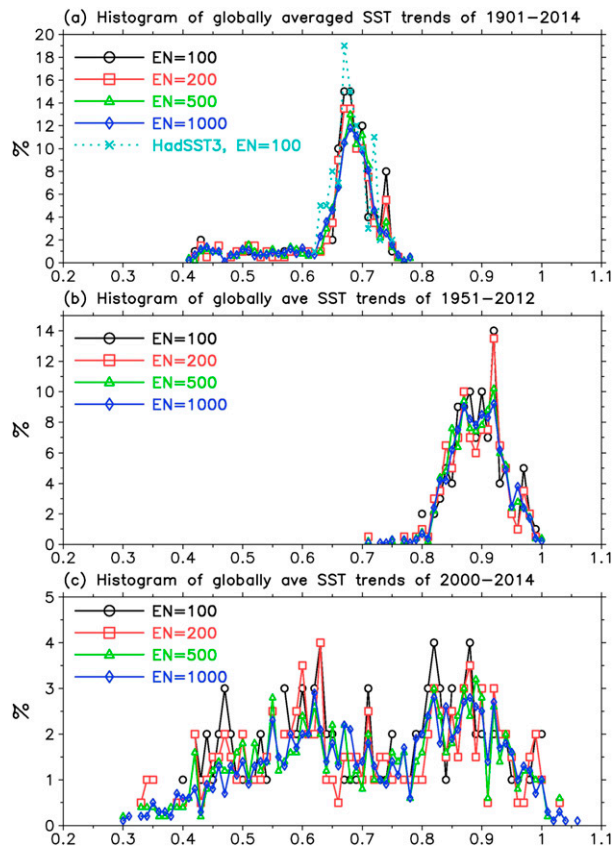
FIG. 6. (a) Histograms of trends (°C century^{-1}; 1901–2014) of globally averaged SST. The histograms are evaluated using parametric uncertainty analyses of ENs of 100 (black line with open circle), 200 (red line with open square), 500 (green line with open triangle), and 1000 (blue line with open diamond). The ensemble averaged trend is approximately 0.676°, 0.670°, 0.668°, and 0.666°C century^{-1} when EN is 100, 200, 500, and 1000, respectively; and is approximately 0.680°C century^{-1} in HadSST3 (light blue line with cross). (b) As in (a), but for trends of 1951–2012. The ensemble averaged trend is approximately 0.903°, 0.901°, 0.902°, and 0.902°C century^{-1} when EN is 100, 200, 500, and 1000, respectively. (c) As in (a), but for trends of 2000–14. The ensemble averaged trend is approximately 0.715°, 0.717°, 0.723°, and 0.730°C century^{-1} when EN is 100, 200, 500, and 1000, respectively.

the estimates are conditional upon the assumptions underlying the analysis and the assumptions regarding sources of uncertainty and their appropriate quantification. This makes their comparison nontrivial in that it is far harder to compare the resulting estimated confidence intervals than the best estimates in a fair and balanced way (Hartmann et al. 2014, box 2.1). For example, the range of SST trends in HadSST3 (Fig. 6a; dotted light-blue line with cross) is narrower (0.62°–0.76°C century^{-1}) than that in ERSST.v4 (0.41°–0.78°C century^{-1}), which is in turn different from (and broader than) the conclusion of the preceding study of Liu at al. (2015). The minimum plausible SST trend is substantially higher in HadSST3

(0.62°C century^{-1}) than in ERSST.v4 (0.41°C century^{-1}). The maximum plausible SST trend is slightly lower in HadSST3 (0.76°C century^{-1}) than in ERSST.v4 (0.78°C century^{-1}). Furthermore, the ensemble averaged SST trend is slightly higher in HadSST3 (0.68°C century^{-1}) than in ERSST.v4 (0.67°C century^{-1}). These differences contribute to the structural uncertainties (refer to section 5b) in SST analyses among different SST products.

3) PRINCIPAL CAUSES OF PARAMETRIC
UNCERTAINTY IN LONG-TERM SST TREND

As seen in Fig. 6a, the long-term (1901–2014) trends of globally averaged SST range from 0.41° to 0.78°C century^{-1} in ERSST.v4, which may primarily be associated with particular selections of some subset of parameters that exert primary control on the outcome. To determine the role of each parameter in the SST trend dispersion, the 1000 ensemble members are first separated into different subensembles according to the chosen options of a particular target parameter from the 24 parameters varied. The number of the subensembles per parameter is the same as the number of that parameter's options (maximum of 8; Table 1). The subensemble averaged trends are then calculated and their deviations relative to the ensemble average (0.67°C century^{-1}) are factorized by that particular parameter in Fig. 7a. The above factor analysis procedure has been repeated for all 24 parameters to ascertain which particular parameters are dominant in determining the trend behavior.

Figure 7a shows that the dominant parameter in affecting the trend dispersion is the third parameter: Min SST STD (Table 1). When Min SST STD is set to be low (0.5°C), more extreme cold observations particularly in the wintertime are excluded from the analysis system during the quality control (QC) procedure. This is particularly true in the earlier period of the analysis before the 1950s. Therefore the SST trend decreases by approximately 0.07°C century^{-1}. In contrast, when Min SST STD is set to be higher (1.0°–1.5°C), the SST trend increases by 0.02°–0.04°C century^{-1}. Similarly, when Max SST STD (the fourth parameter) is set to be low (3.5°C) or higher (5.5°C), the SST trend decreases or increases. These results indicate that the QC criteria play a dominant role in the uncertainty of long-term SST trends. The role of QC is also indicated by the first-guess (the first parameter) selection, which provides the expected value around which the cutoff criteria is applied. When the first guess uses the adjusted (unadjusted) v3b SST field, the SST trend decreases (increases) by 0.03°C century^{-1} respectively; note that for any parameter with only two possible options such as first-guess by construction the effect will be equal and opposite for the
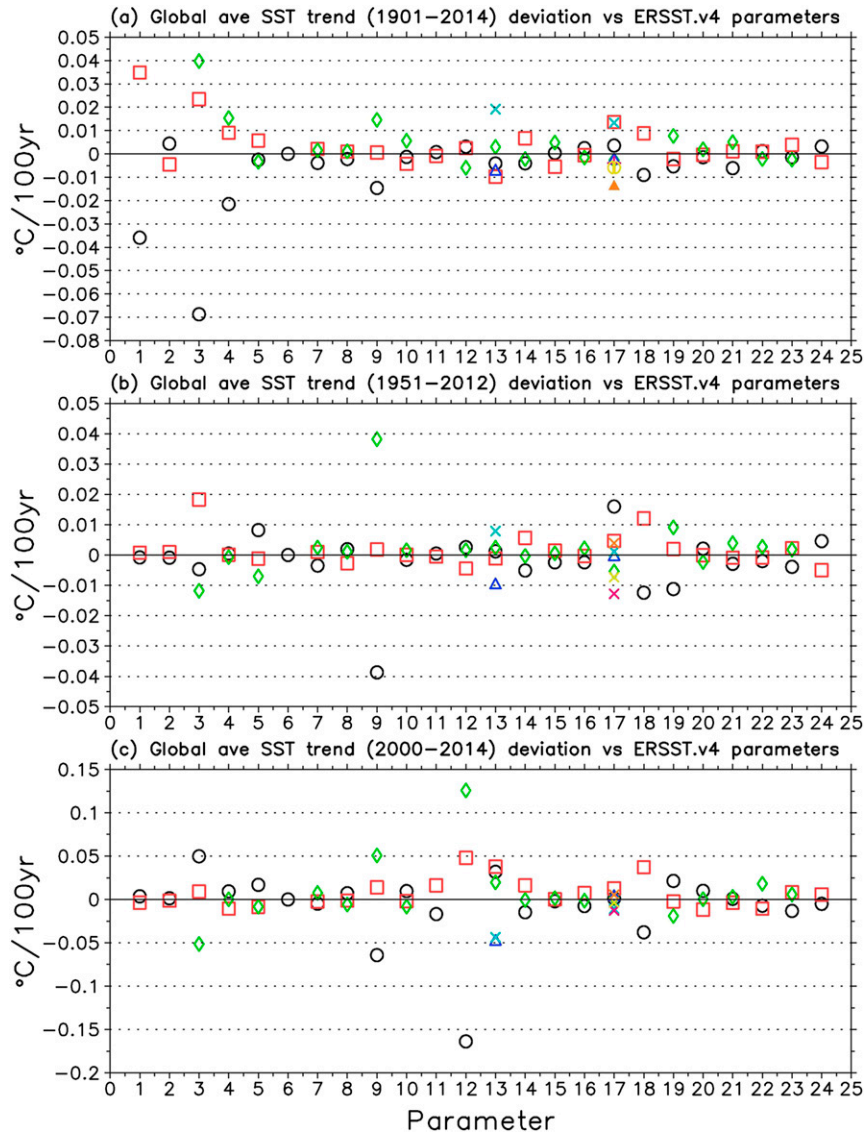
FIG. 7. (a) Factor analysis in the deviations of SST trend of globally averaged SST in (a) 1901–2013, (b) 1951–2012, and (c) 2000–14. Unit in y axis is in °C century^{-1}. The x axis represents each of 24 parameters in Table 1. The symbols represent the ensemble options in second column of Table 1. The ensemble options from 1st to 8th, respectively, are represented by black circle, red square, green diamond, blue triangle, light-blue cross, magenta plus sign, yellow open cycle with vertical bar, and orange closed triangle. The deviations are relative to the ensemble averaged SST trend of 0.67°C, 0.90°C, and 0.73°C century^{-1} in (a)–(c), respectively.

two options. The QC procedures occur early within the sequential processing algorithm so that they can interact with a number of subsequent parameters. In particular QC may impact the weights given to particular EOTs as well as the degree of bias correction to be applied.

The dominant role of quality control may be relatively easily understood in post hoc analysis. But it was far from obvious a priori that QC choices would have any demonstrable impact on long-term trends given that

these steps relate to inclusion or exclusion of a relatively small subset of individual input observations. It appeared more likely that steps associated with the calculation of the bias adjustments would be dominant in determining the trend dispersion of the parametric ensemble of solutions. This points to the importance of holistically assessing U_p by varying all uncertain parameters within the algorithm rather than a restricted subset thereof, because the parameters that actually turn out to be important may

not equate to those highlighted as potentially important based on intuition.

That said, Fig. 7a also shows that the 9th, 13th, and 17th parameters listed in Table 1 play a somewhat important but less dominant role in determining the SST trend dispersion. When the adjustment between ship and buoy observations (the 9th parameter) increases, the SST anomalies increase in the modern period after the 1980s, which serves to increase the long-term SST trend. When the coefficients for the bias adjustment (the 13th parameter) are linearly fitted, the bias adjustment becomes lower before the 1920s and higher between the 1920s and the 1940s [see Fig. 6a in Huang et al. (2015a)]. Therefore, the SST trend increases by approximately $0.02°C\,century^{-1}$. This is consistent with the conclusion in Liu et al. (2015) for this parameter (the two other parameters were not considered by Liu et al.). The SST trend may also change when a particular set of EOTs (the 17th parameter) is selected. The impacts from other parameters are much smaller.

These results serve to highlight which steps are most important in determining the outcome. They therefore naturally highlight potential areas for further innovation and refinement in developing ERSST further to yield better estimates and/or better explore the range of plausible SST histories.

4) UNCERTAINTY OF SST TREND IN THE "HIATUS" PERIODS

A recent study (Karl et al. 2015) indicated that the trend of globally averaged SST in ERSST.v4 in the most recent decades ($0.99°C\,century^{-1}$; 2000–14) is as large as in the longer period of 1951–2012 ($0.88°C\,century^{-1}$). Figure 6b shows the histogram of the trend during the longer period. The trend ranges from 0.7° to $1.0°C\,century^{-1}$, which is higher than the long-term trend shown in Fig. 6a, indicating stronger oceanic warming since the middle of the twentieth century. Factor analyses indicate that the major contributor to this trend uncertainty is the ship-buoy adjustment (the ninth parameter; Fig. 7b).

Relative to the 1951–2012 period, the range of possible SST trends ($0.3°–1.1°C\,century^{-1}$) in the recent "hiatus" decade (2000–14; Fig. 6c) is nearly twice as wide. This implies a larger uncertainty of the trend over the hiatus period than in the 1951–2012 period that may arise in part due to less cancellation of terms over a shorter than a longer segment. The large uncertainty in the recent decade is mostly associated with the selection of the ship SST bias adjustment (the 12th parameter; Fig. 7c) derived from the Nighttime Marine Air Temperature (NMAT) dataset (Huang et al. 2015a), which indicates the important role of surface air temperature

and its uncertainty in assessing the SST uncertainty (Cowtan et al. 2015). The trend difference is as large as $0.3°C\,century^{-1}$ when different ship SST bias adjustments are applied, which also results in a second peak in the histogram shown in Fig. 6c. The SST trend in the recent decade ($0.99°C\,century^{-1}$) in operational ERSST.v4 locates at the high end of the histogram shown in Fig. 7c due to the asymmetric selection of parameters in the 1000-member ensemble. For example, the trend is high when ship SST bias adjusted using the latest and regional NMAT modes (option 1 of the 12th parameter).

The large uncertainty of the globally averaged SST trend is not a unique feature in the recent 15-yr hiatus period. For example (not shown in figure), the SST trend between 1980 and 1994 ranges from 0.15° to $1.0°C\,century^{-1}$, which is mostly associated with the selection of ship SST bias adjustment (12th parameter in Table 1). In the earlier period of 1930–44, the SST trend ranges much wider from 0.2° to $4.5°C\,century^{-1}$ due to higher SST uncertainty (see Fig. 10), which is mostly associated with the selection of QC parameter (Min SST STD; third parameter in Table 1) and ship SST bias adjustment (12th parameter in Table 1). These results suggest that the uncertainty of SST trend depends on 1) the length of time period being considered and 2) the particular observational characteristics of the SST record in the epoch of interest including both sampling completeness and stability of observational techniques. Specifically, for periods as short as 15 years, the uncertainty of SST is driven by the selection of QC parameter values in the earlier period and by ship SST bias adjustment in the modern period.

5. Total quantified uncertainty and intercomparisons to independent estimates

a. Total quantified uncertainty in ERSST.v4

The averaged (1871–2005) U_t for local SSTs in ERSST.v4 (Fig. 8a) is 0.4° to 0.8°C in the northern North Pacific, the northwestern North Atlantic, the Southern Ocean between 30° and 60°S, the eastern equatorial Pacific, and the South Atlantic along the coasts of southern Africa. The large U_t in these regions is associated with both reconstruction uncertainty (0.2°–0.4°C; Fig. 1a) and the parametric uncertainty (0.4°–0.8°C; Fig. 3a). In the lower-latitude oceans between 30°S and 30°N, the U_t is approximately 0.2°–0.4°C (Fig. 8a), which is also attributed to both the reconstruction uncertainty (0.2°–0.4°C; Fig. 1a) and the parametric uncertainty (0.2°–0.4°C; Fig. 3a).

Figure 9 shows that when globally averaged across all grid boxes U_t in ERSST.v4 (solid black line) is as high as
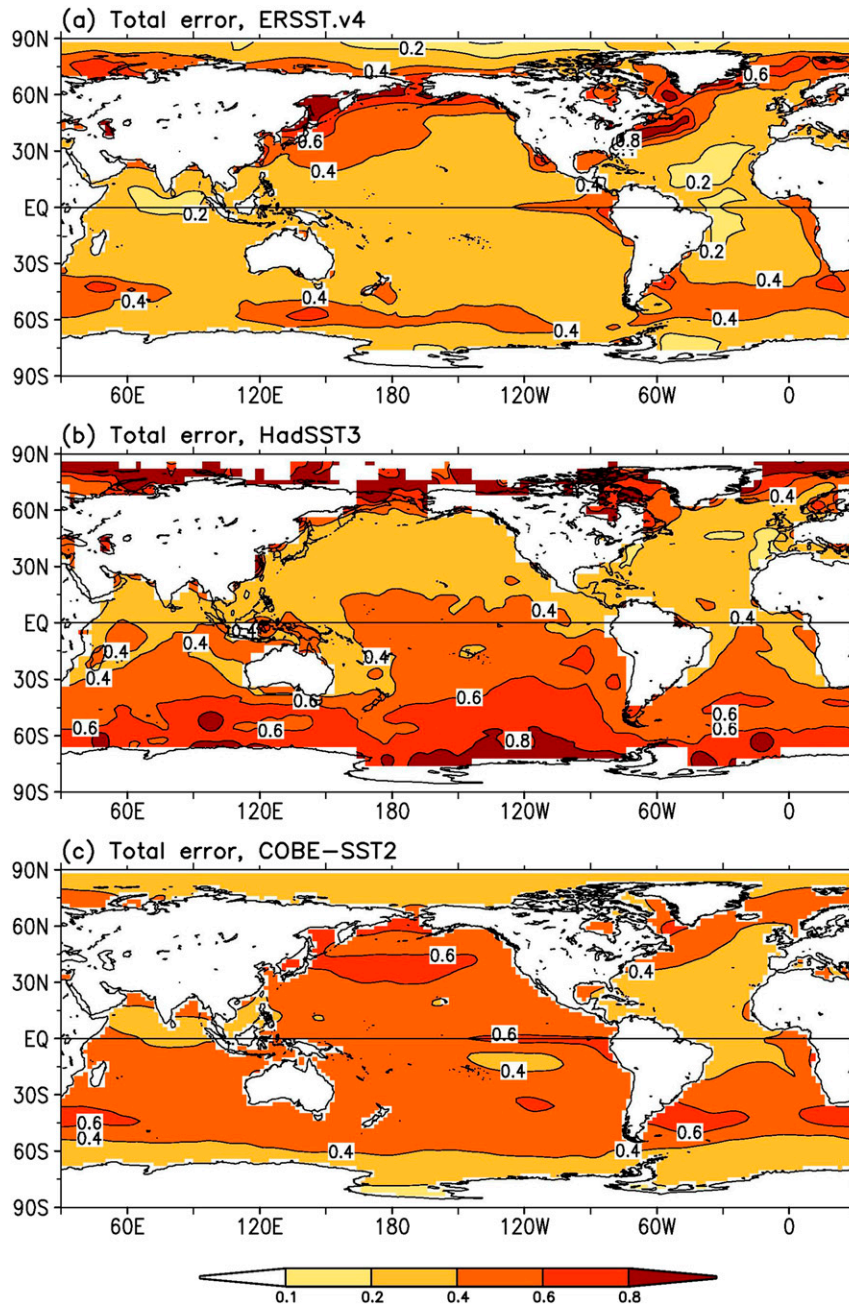
FIG. 8. Averaged (1871–2005) total uncertainty (1σ) of (a) ERSST.v4, (b) HadSST3, and (c) COBE-SST2. Contours are 0.1°, 0.2°, 0.4°, 0.6°, and 0.8°C. The averaging period of 1871–2005 is to be consistent with the common period of the reconstruction uncertainty in Figs. 1 and 3a.

0.7°C in the later 1860s, and gradually decreases to approximately 0.3°C after the 1960s. The U_t value peaks in the early 1890s and then subsequently during the two World Wars. The U_t is mostly associated with the parametric uncertainty before the 1900s and during the two World Wars (red line in Fig. 3b), but it is mostly associated with the reconstruction (dashed orange line

in Fig. 2b) uncertainty after the 1960s when observations become more plentiful and the uncertainty related to bucket corrections is no longer as important.

In comparison to the globally averaged U_t for local SSTs (Fig. 9), the U_t for global-mean SST in ERSST.v4 is much smaller (solid black line; Fig. 10; note the magnitude difference in the y axis between Figs. 9 and 10). The
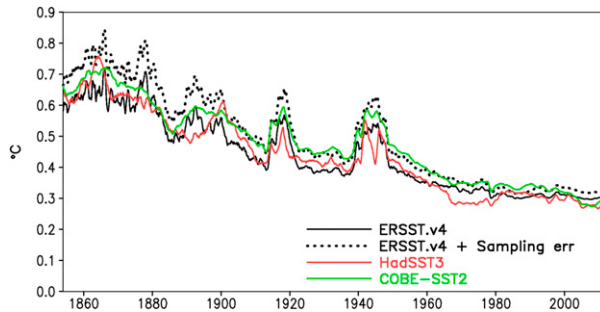
FIG. 9. Globally averaged total uncertainties (1σ) of ERSST.v4 (black line), ERSST.v4 + sampling error (black dotted line), HadSST3 (red line), and COBE-SST2 (green line). A 12-month running mean is applied.

FIG. 10. Total uncertainties (1σ) of globally averaged SST in ERSST.v4 (black line), ERSST.v4 + sampling error (black dotted line), HadSST3 (red solid line), HadSST3 − Correlation (red dashed line), and HadSST3 + Correlation (red dotted line), overlapped with the sampling uncertainty in COBE-SST2-OI (green solid line) and COBE-SST2-MTA (green dotted line). A 12-month running average is applied except for COBE-SST2-OI and COBE-SST2-MTA, which are adapted from Hirahara et al. (2014).

smaller U_t is expected since many of the uncertainties in individual grid boxes will tend to cancel when SSTs are averaged over the global domain, and the averaged SST becomes more reliable based on the larger number of observations. Despite the large difference in magnitude, the temporal variation of the U_t for globally averaged SST is very consistent with that of the U_t for local SSTs (e.g., both are relatively higher in the early 1860s and during the two World Wars). However, the U_t for globally averaged SST arises almost entirely from the parametric uncertainty (red line in Fig. 3c), and the contribution from the reconstruction uncertainty (dashed orange line in Fig. 2c) is negligible (see earlier analysis of U_r for further discussion on this point).

b. Intercomparisons to independently derived uncertainty estimates

The intercomparison to independently derived estimates of SST uncertainty may provide evidence as to whether the quantified SST uncertainty can explain the apparent disagreement between SST datasets. Comparisons, however, are complicated because all the groups consider distinct subsets of the possible sources of uncertainty. Even where the same sources are considered they are invariably quantified in distinct ways.

Different uncertainty models have been used for quantifying SST uncertainties by different groups over time. For example, Davis (1976) and Shen et al. (1998) proposed to assess the uncertainties associated with spatial SST modes. In HadSST3, uncertainties were quantified based on uniqueness of SST call signs, statistics of observations within a specific grid box, and their correlations with surrounding grid boxes (Kennedy et al. 2011a); and the uncertainties include sampling uncertainty with and without correlation, bias, and coverage uncertainties (Table 3). In earlier versions of ERSST, uncertainties were assessed based on statistics of low- and high-frequency characteristics of SSTs from
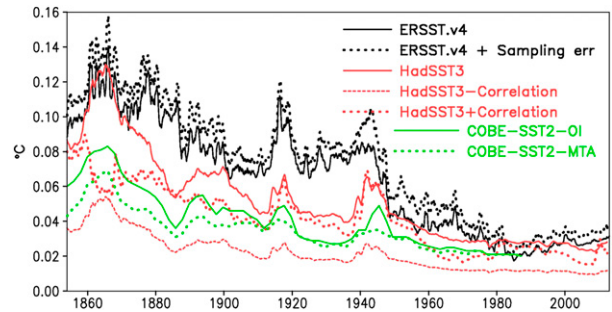
both model simulation and SST analysis (Smith and Reynolds 2003, 2004, 2005), and the uncertainties include sampling and bias uncertainty. In the initial ERSST.v4 analysis a subset of the parametric uncertainty considered herein was quantified (Liu et al. 2015). Similar to the uncertainty assessment in previous versions of ERSST (Smith and Reynolds 2003), the uncertainties in COBE-SST2 were evaluated by subsampling modern observations to the sampling of data-sparse periods (Hirahara et al. 2014); the uncertainties include sampling uncertainties using optimal interpolation and multiple time scale analysis (MTA; Table 3).

Herein we compare solely the most recent products' uncertainty estimates under the assumption that most users will consider the newest version of the various available products. The SST uncertainties in HadSST3 between 1850 and 2013 (Kennedy et al. 2011a; data are available at http://www.metoffice.gov.uk/hadobs/hadsst3/data/download.html) and in COBE-SST2 between 1850 and 1990 (Hirahara et al. 2014; data are available at https://amaterasu.ees.hokudai.ac.jp/~ism/pub/cobe-sst2) are compared to those in ERSST.v4. In comparison to HadSST3 (Fig. 8b), the averaged (1871–2005) U_t in ERSST.v4 (Fig. 8a) is 0.2°–0.4°C larger in the northern North Pacific and northern North Atlantic, and 0.2°–0.4°C smaller in the Southern Hemisphere oceans and in the Arctic. In contrast, the averaged U_t is approximately 0.2°C smaller in ERSST.v4 than in COBE-SST2 in most of the tropical–subtropical oceans, but approximately 0.2°C larger in the Arctic. Despite these distinct regional expressions of the quantified uncertainties, the globally averaged U_t for local SSTs is very consistent among ERSST.v4, HadSST3, and COBE-SST2 (Fig. 9).

The U_t of globally averaged SST, however, is 0.02°–0.06°C higher in ERSST.v4 than in HadSST3 before the
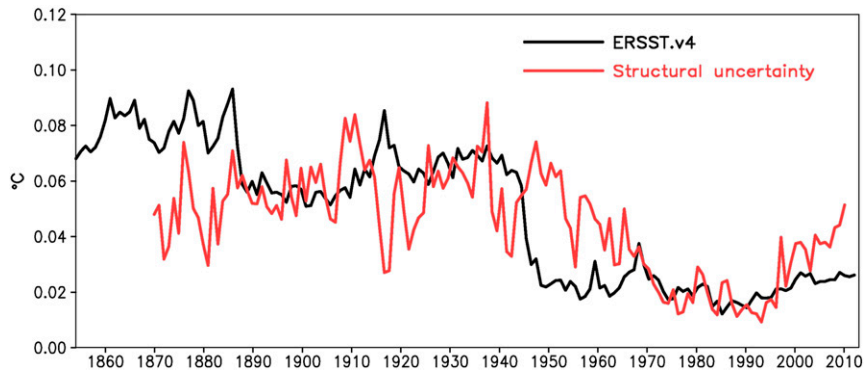
FIG. 11. Total quantified uncertainty in annually and globally averaged SST in ERSST.v4 (black line) and structural uncertainty (red line) from six SST products. These six products are ERSST.v4, ERSST.v3b, HadSST3, HadISST, Kaplan SST, and COBE-SST2.

1940s (Fig. 10) and then becomes relatively consistent with the remaining products after the 1940s. This raises the question as to why these distinctions may occur. First, the global-mean uncertainty in ERSST.v4 is critically dependent on the magnitude of parameter perturbations in the U_p ensemble (section 4b). If these are too broad then the resulting uncertainty estimates may be too large, and this would be expected to be expressed primarily in the data-sparse early period. Second, the methodologies (Table 3) used to estimate the uncertainties are substantively different among the products and are themselves uncertain. The U_t is estimated using 32 ensemble members of reconstruction uncertainty and 1000 ensemble members of parametric uncertainty in ERSST.v4, whereas it is estimated using 100 ensemble members in HadSST3 (Kennedy et al. 2011b). The sampling uncertainty is estimated using 5 years of observations during the data abundant period of

2006–10 in COBE-SST2 (Hirahara et al. 2014). Further complication arises because there exist a range of approaches for uncertainty estimation. For example, in COBE-SST2 the sampling uncertainty using optimum interpolation (COBE-SST2-OI) is nearly 0.02°C larger than using multiple time-scale analysis (COBE-SST2-MTA; Fig. 10) before the 1890s and in the later 1910s and early 1940s. Similarly, the U_t is 0.02°–0.04°C larger with interbox correlation (HadSST3 + Correlation) than without interbox correlation (HadSST3 − Correlation) before the 1960s. The lack of knowledge of the true SST evolution precludes a definitive assessment of the adequacy of the three sets of uncertainty estimates, although arguably none is likely to be absolutely holistic.

The U_t in ERSST.v4 is roughly consistent with the "structural" uncertainty (Kennedy 2014) that is defined as the SST STD among different products. Figure 11 shows the U_t for annually and globally averaged SST in
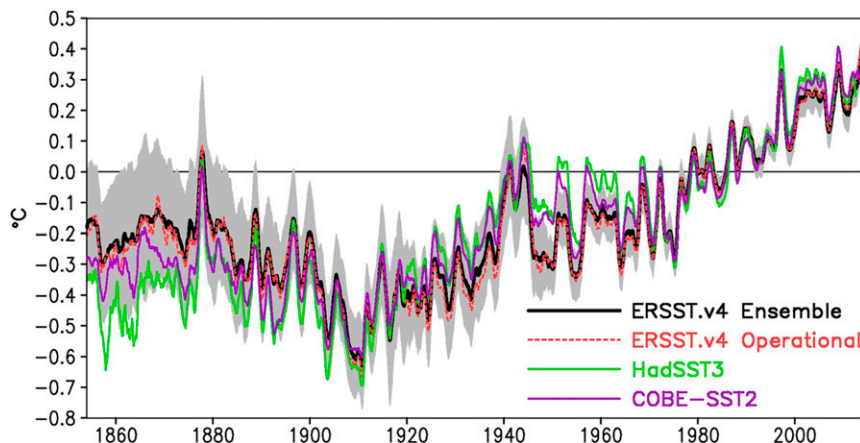

FIG. 12. Global averaged SSTAs of ERSST.v4 ensemble average (black line), ERSST.v4 operational (dashed red line), HadSST3 (green line), and COBE-SST2 (purple line). The shaded region represents the 95% confidence interval due to total quantified uncertainty in ERSST.v4.

ERSST.v4 and structural uncertainty defined as the spread among ERSST.v4, ERSST.v3b, HadSST3, HadISST, Kaplan SST, and COBE-SST2. The U_t in ERSST.v4 is mostly consistent with the structural uncertainty except before the 1880s, between the later 1910s and 1920s, and between the 1940s and 1960s. The averaged (1870–2010) U_t in ERSST.v4 and the structural uncertainty are approximately 0.046° and 0.045°C, respectively. The U_t and structural uncertainty increase slightly after the 1990s, which may be associated with (a) the number of buoy observations increasing rapidly after the 1980s (noting that some of the estimates considered in the structural uncertainty term apply ship-buoy adjustments whereas others do not) and (b) the coverage of in situ (ship + buoy) observations decreasing slightly after that time.

As implied by the structural uncertainty, the SSTAs are different among independently produced datasets. Figure 12 shows that HadSST3 is near or beyond the 95% confidence interval of ERSST.v4 U_t quantified herein before the 1910s, between the 1920s and 1930s, and between the later 1940s and 1960s. However, the COBE-SST2 is mostly within the 95% confidence interval except for between the later 1940s and 1960s. As demonstrated by Huang et al. (2015a), the SSTA difference between ERSST.v4 and HadSST3 is largely associated with the difference of their respective SST bias adjustments. This suggests that the range of the bias uncertainties within the ERSST.v4 parametric uncertainty system cannot account for the bias adjustment differences between ERSST.v4 and HadSST3. In other words the ERSST.v4 parametric ensemble cannot adequately emulate the HadSST3 bias adjustment approach through perturbation of within-algorithm uncertain parameters. The range of the SST bias uncertainty in ERSST.v4 is principally predetermined by two versions of NMATs and the choices of adjustment smoothers, although exhaustive efforts were made to identify all other parameters that could possibly be varied. Overall, parameters are perturbed by 10%–50% values used in operational production, even by 100% for some key parameters such as EOT critical values. Independent approaches are clearly required to fully explore uncertainties in climate data records.

6. Summary, discussion, and conclusions

a. Principal findings

The SST uncertainty in ERSST.v4 has been assessed using a variety of test datasets and consideration of uncertainty arising from an expanded selection of intrinsically uncertain internal parameter values' settings.

Comparisons indicate that the reconstruction uncertainty, which is the unavoidable information loss at local scales during reconstruction, changes only slightly when different reasonable spatially complete test data are applied. The reconstruction uncertainty using the test data from MOISST is very similar to that using the test data from GFDL-ESM2G and HadGEM2-AO GCMs, and 0.1°C larger than that using the test data from HadISST. The parametric uncertainty based upon the expanded set of parameters varied is approximately 100% larger than was estimated by Liu et al. (2015). The larger parametric uncertainty results from a combination of (a) broader ranges of some parameters considered in Liu et al. (2015), (b) more internal parameters being varied, and (c) the entirely random selection of parameter options. The reconstruction uncertainty estimated by applying the test data from MOISST and the parametric uncertainty based on 1000 ensemble analyses is used to estimate the total SST uncertainty of ERSST.v4, recognizing that this cannot be construed as an absolute estimate given the statistically ill-posed nature of the underlying problem.

The total uncertainty is closely associated with the availability of historical SST observations. It is larger in the high-latitude oceans and before the 1950s, because observations are sparse in these regions in the earlier period of historical observations. It is also large when observations are sparse due to the World Wars I and II. In contrast, the total uncertainty is small in the lower-latitude oceans and in the modern period after the 1970s when sampling has been good. However, the total uncertainty does somewhat increase again in the most recent period owing to uncertainties in the ship to buoy transition and due to slightly reduced coverage of observations. The globally averaged uncertainties are close to the median uncertainties, whereas the 90th percentile uncertainties are almost 2 times the median uncertainties, reflecting skew in the geographical distribution of uncertainties with a long tail of high uncertainty in certain regions. There are several areas with relatively large uncertainties, confined within a small region located in the Arctic, northern North Pacific, northwestern North Atlantic, and part of the South Ocean south of 40°S where observations are sparse.

At the grid box scale (2° × 2°), the total uncertainty (0.3°–0.7°C) of local SST is roughly equally associated with both reconstruction and parametric uncertainties. At the global scale, the total uncertainty (0.03°–0.14°C) of globally averaged SST is much smaller than the globally averaged total uncertainty of local SSTs. The reasons for the smaller total uncertainty of globally averaged SST are that many of the uncertainties of grid box–scale measurements partially or completely cancel when global averaging is performed. The total uncertainty

of globally averaged SST arises mainly from the parametric uncertainty, while the contribution from the reconstruction uncertainty is very small.

Tests show that the parametric uncertainty, long-term SST trend dispersion ($0.41°$–$0.78°C\,century^{-1}$), and ensemble averaged SST trend ($0.67°C\,century^{-1}$) and its error at 95% confidence level ($0.15°C\,century^{-1}$) are not very sensitive to the number of ensemble members when the number is larger than 500. All positive ($>0.41°C\,century^{-1}$) SST trends of the 1000 ensembles suggest that the warming trend of the historical SST observations represented by the ERSST.v4 analysis system is very robust to the recognized and quantified uncertainties. The estimates quantified herein would need to be an underestimate of the true uncertainty by a factor of at least 4 to call into question the conclusion that globally SSTs have risen since the early twentieth century. Such expanded uncertainties would also not be able to preclude that SSTs have warmed at a far greater rate than current estimates suggest. Additional tests show that the dispersion of the long-term SST trend is mostly associated with the parameters in QC procedures and ship SST bias adjustment schemes. Possible improvements in undertaking these steps will hence naturally be a focus of the ERSST algorithm's future development. The range of SST trend is larger in the recent "hiatus" period than over 1951–2012, indicating a larger uncertainty of the SST trend in the hiatus period. We note that trends over short periods are inherently more uncertain because the period is short, and both random and shorter-term systematic effects will not cancel as they would for longer-term trend periods. The quantified total uncertainty for various periods of 15-yr length is more uncertain than for multidecadal or centennial time scales. For different 15-yr segments, different factors are important reflecting the changes in sampling and observing techniques. In general, periods of stable coverage and technique exhibit lower uncertainty of 15-yr time scale trends. The hiatus period suffers from both a reduction in sampling and the effects of moving from 90% ship measures to 90% buoy measures. Clearly maintaining a consistent monitoring capability moving forwards would be beneficial for climate assessment.

Finally, the uncertainty estimates from ERSST.v4 have been compared with those arising from HadSST3 and COBE-SST2. The comparisons indicate that the magnitude and temporal variation of total uncertainty for both local and globally averaged SSTs are broadly consistent among ERSST.v4, HadSST3, and COBE-SST2. However, differences are found in the spatial distribution of quantified uncertainties. The uncertainty is small ($0.1°$–$0.4°C$) in the Arctic in both ERSST.v4 and COBE-SST2, while it is larger ($0.6°$–$0.8°C$) in HadSST3. The uncertainty is large ($0.4°$–$0.8°C$) in the northern North Pacific and northwestern Atlantic in both ERSST.v4 and COBE-SST2, while it is smaller ($0.4°C$) in HadSST3. The uncertainty is small ($0.4°$–$0.6°C$) in the Southern Ocean in both ERSST.v4 and COBE-SST2, while it is larger ($0.6°$–$0.8°C$) in HadSST3.

The reasons for aforementioned uncertainty differences may result from (a) the selection of parameter values in ERSST.v4, (b) the methodologies applied in the estimation of SST uncertainties in HadSST3 and COBE-SST2, and (c) the distinct treatments of random and sampling errors, which will form a focus of our future development. Further studies are needed to clarify what causes the differences of the estimated uncertainties among different SST products, which may help understand the underlying physical and/or statistical reasons resulting in the differences in SST uncertainties so that the future estimation of SST uncertainty could be improved.

b. Caveats

The random uncertainty term is considered to be uncorrelated in the present study, while this will not be true for a specific ship track as indicated in Kennedy et al. (2011a). Improved ship-track data in future ICOADS releases may permit a more nuanced approach to the consideration of this term that allows for the inclusion of the correlated aspects in the ERSST algorithmic framework. At this time owing to the gross incompleteness of the track data this is not possible to incorporate. The inclusion of a correlated random term based upon tracks may logically yield regional false SST biases at monthly scales in the input data, and hence have an effect on the EOT selection, weighting, and ordering in particular, and as a result serve to increase the uncertainty in reconstructed small-scale to regional SSTs at the monthly scale. The possible effect on global-mean estimates and their trends is not entirely clear although the impacts are likely to cancel in space and time and be largest in data-sparse regions sampled by few independent platforms.

The sampling uncertainty U_s, which is due to incomplete sampling over the grid, was treated very differently in different products (Hirahara et al. 2014; Kennedy et al. 2011a; Smith et al. 2008). Using DOISST data (Table 2), the sampling uncertainty is tested by a pair of analyses: one spatially complete (fully sampled) and the other incomplete (subsampled) to match observed sampling (Smith and Reynolds 2003, 2004, 2005; Hirahara et al. 2014): $U_s^2(x, y, t) = [A_f(x, y, t) - A_s(x, y, t)]^2$, where A_f and A_s represent fully sampled and subsampled analyses, respectively. Our tests show, however, that the U_s may not be independent from U_p owing to the large number of sampling related parameters varied substantially
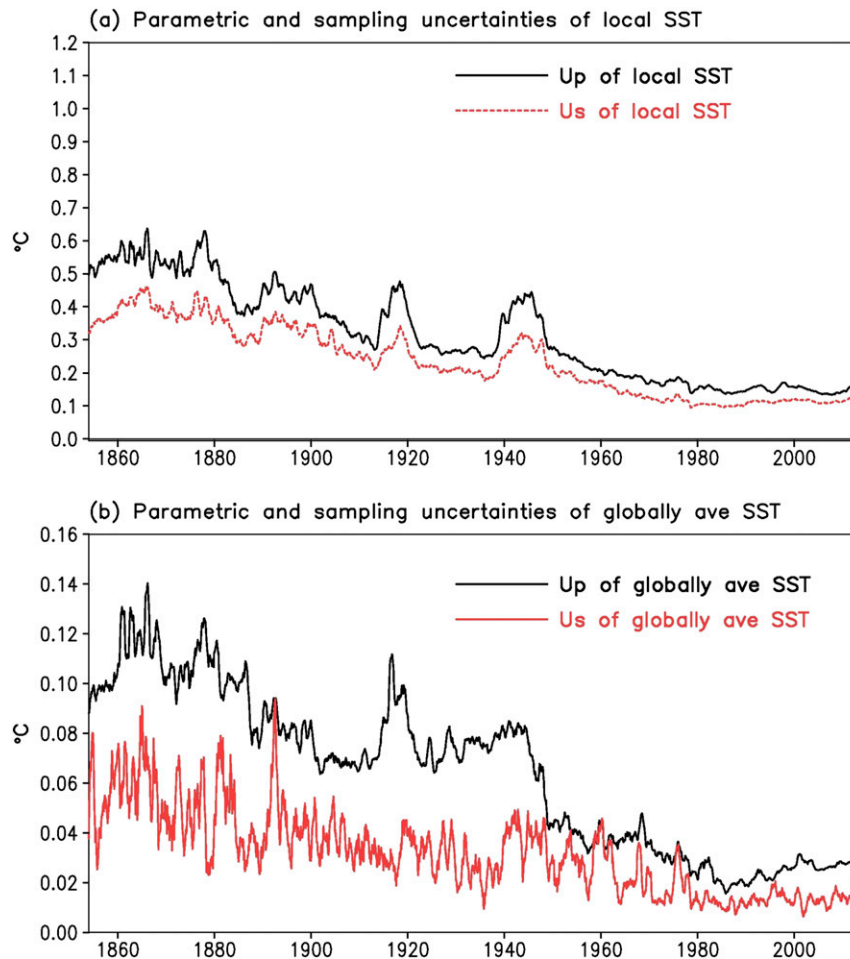
FIG. 13. (a) Global averaged parametric (U_p) and sampling (U_s) uncertainties of local SST. Their correlation coefficient is 0.99. (b) The U_p and U_s of globally averaged SST. Their correlation coefficient is 0.80. The U_p is about 35% larger than U_s in (a) and about 110% larger in (b).

in the 1000-member ensemble. Their correlation is 0.99 between the globally averaged uncertainties for local SST (Fig. 13a), and 0.80 between the uncertainties for globally averaged SST (Fig. 13b). The magnitude of U_p is approximately 35% and 110% larger than that of U_s for local and globally averaged SSTs, respectively. In addition, the total uncertainty in ERSST.v4 (solid black lines in Figs. 9 and 10) is comparable to or larger than those in HadSST3 and COBE-SST2. If the sampling uncertainty were included (dotted black lines in Figs. 9 and 10), the total uncertainty would be higher in comparison with HadSST3 and COBE-SST2, particularly for the globally averaged SST. This may reflect a true underestimation of the actual uncertainty in these preceding products.

The impact of sampling on the parametric uncertainty estimation was further assessed by additional experimental analyses. Using spatially complete model simulation as "observations" (zero uncertainty arising from sampling by definition), the globally averaged parametric uncertainty of local SST reduces to a near constant of 0.1°C, while the parametric uncertainty of globally averaged SST is near zero. The same conclusions are reached when spatially complete DOISST analysis is used as the virtual observations. Hence the vast majority of the variant behavior in U_p arises from sampling effects and their interactions with additional methodological steps.

Given that the correlation between the sampling and parametric uncertainties is high and the parametric uncertainty is near zero when sampling is perfect, we argue that, in the framework of our uncertainty estimation, much or all of the U_s term should be considered as constituting a component of the parametric uncertainty. Therefore, we do not officially include the sampling error in the estimation in the total uncertainty in Eq. (4),

but the total uncertainty including the sampling uncertainty is provided in Figs. 9, 10, and 13 for readers' reference. In future work, tests with different algorithms will be designed to further verify whether the sampling uncertainty should be included in the parametric uncertainty using fully sampled model output and/or other analyzed SST dataset as ERSST continues to be developed.

c. Concluding remarks

In conclusion, this paper has documented an expanded uncertainty model used in ERSST.v4, quantified and analyzed each source, and compared the resulting estimate to those from two other state-of-the-art SST datasets. Uncertainties are primarily controlled by the density and coverage of observations such that total uncertainty decreases over time with peaks at the time of the two World Wars and then increases slightly again since the late 1990s. The ERSST.v4 uncertainty estimates are broadly comparable in the global mean to other estimates. The uncertainty in centennial time scale trends is 4 times smaller than the estimated SST trend. Therefore the conclusion that the global ocean surface has warmed since 1900 remains extremely robust to recognized and quantified sources of uncertainty.

APPENDIX

Options of ERSST.v4 Internal Parameters

Options 1–8 of 24 internal parameters in ERSST.v4 are provided and listed in Table 1. One of these options is implemented in the operational ERSST.v4 production, and the other alternative options are used for the parametric uncertainty estimates. The details of these parameter options are as follows (presented in the order that they appear in the algorithm processing chain):

1) *First-guess (FG) used for quality control (QC)*: The deviation of observations from the FG is assessed to ensure the outlier observations are not included in the analysis. The FG from previous ERSST v3b is used in v4. Since the v3b SSTs are bias adjusted while raw observations are not bias adjusted, the unadjusted SST from v3b is used to assess the contribution of FG to the uncertainty of SST analysis.

2) *Standard deviation (STD) used for quality control (QC)*: Observed SSTs may be discarded within the QC procedure in selecting raw observations, if they deviate from the FG by more than 4.5 times the SST STD. Two sets of SST STDs are used. One was calculated from COADS observations from 1950 to 1979 and implemented in v3b; the other is from monthly OISST from 1982–2011 and implemented in v4. The STD is generally smaller in OISST than in COADS, which suggests that fewer SST observations may be used when STD from OISST is applied (Huang et al. 2015a). The factor of 4.5 is termed the STD multiplier and may also vary as described in parameter 5.

3) *Minimum (Min) SST STD*: To maintain a good QC procedure, a minimum STD (1.0°C) for parameter 2 is set in the ERSST.v4, and its alternative options are 0.5° and 1.5°C.

4) *Maximum (Max) SST STD*: In contrast to parameter 3, a maximum STD (4.5°C) for parameter 2 is set in the ERSST.v4, and its alternative options are 3.5° and 5.5°C.

5) *SST STD multiplier*: The multiplier to parameter 2 is set to 4.5 in ERSST.v4, and its alternative options are 3.5 and 5.5. A larger value of minimum and maximum STD and STD multiplier will enable the ERSST.v4 to include more extreme input SST observations in subsequent processing steps.

6) *Random error of observations*: In the uncertainty estimation the random error is added to a single ship or buoy observation as described in section 3c. The mean of the random error is set to 0, and the STD of the random error is set to 1.3° and 0.5°C for ship and buoy observations, respectively (Reynolds et al. 2002; Kent and Challenor 2006).

7) and 8) *Ship and buoy SST error STD*: Random error STDs of ship and buoy observations are different, which are approximately 1.3° and 0.5°C (Reynolds et al. 2002), respectively. These empirically derived STDs are somewhat uncertain when they are taken into account in weighting EOTs [refer to Eq. (3) in Huang et al. (2015a)]. Therefore their values are perturbed by ±0.1°C accordingly as their alternative options.

9) *Ship-buoy SST adjustment*: Studies (Reynolds et al. 2002, 2007; Kent et al. 2010; Huang et al. 2015a) showed that observations from ships and buoys exhibit a systematic difference. The averaged ship-buoy difference is approximately 0.12°C with an STD of 0.04°C. The ship-buoy SST adjustment is therefore set to 0.12°C in ERSST.v4, and its alternatives are set

to 0.08° and 0.16°C. This is broader than the ranges explored by Liu et al. (2015) for this parameter.

10) *Buoy SST weighting*: An earlier study (Reynolds and Smith 1994) indicated that the 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 alternative weightings are set to 5.8 and 7.8. Parameter 10 may not be completely independent from parameters 6–8, and therefore the uncertainty from these parameters may slightly be underestimated.

11) *SSTA calculation*: In ERSST.v3b, bin averaged SSTs were calculated first on a regular $2° \times 2°$ grid, and then SSTAs were calculated as the differences between SST and its climatology (1971–2000). In ERSST.v4, SSTAs at in situ locations are calculated between SSTs and SST climatology at these locations, and then SSTAs are bin-averaged to a $2° \times 2°$ grid. The order of operations can have an impact. These two options of SSTA methods are used for the parametric uncertainty estimation.

12) *NMAT for SST bias adjustment.* In both ERSST v3b and v4, NMAT values are used to calculate the ship SST bias (Huang et al. 2015a). In v3b, an earlier version of UKMO NMAT was used, while HadNMAT2 is used in v4. In both v3b and v4, SST biases are fitted to a global climatological model of SST NMAT difference. However, tests showed the SST biases may change if they are fitted to regional climatological models, say 90°–30°S, 30°S–30°N, and 30°–90°N. Therefore, bias uncertainty is taken into account by including options of using different NMATs and their modes.

13) *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 bias in ERSST.v4 [see details in Huang et al. (2015a)]. Alternative filters are considered in the parametric uncertainty estimation when coefficient f is set to 0.05 and 0.20. In pursuing a full bias uncertainty, additional options of linear fitting and annually averaged filtering are also included.

14) *Minimum number of months for annual average*: In retrieving the 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.v4. The alternative numbers are set to 1 and 3 months.

15) *Minimum ratio of superobservations*: In retrieving the LF anomaly, a $26° \times 26°$ spatial running mean filter is applied. In the regions without observations where the value of superobservations is labeled as missing, the missing value is replaced by the averaged value within a $26° \times 26°$ subdomain, if the ratio of superobservations coverage within the subdomain is greater than 0.03 (five valid superobservations vs a maximum of 169 grids). In the estimation of parametric uncertainty, the alternative ratios are set to 0.02 and 0.04. Superobservations are defined as the bin-averaged SST observations over the $2° \times 2°$ grid boxes.

16) *Maximum observation number*: In applying the $26° \times 26°$ spatial filter in parameter 15, an averaged superobservation is calculated by weighting each $2° \times 2°$ grid box area and observation numbers within the grid box. To protect from the averaged superobservations being overwhelmed by a single densely observed grid box, a maximum observation of 10 is set in ERSST.v4. Its impact on the parametric uncertainty is considered by alternative numbers of 5 and 15.

17) *EOT training period and domain restriction*: In ERSST v3b and v4, HF SSTAs are decomposed with EOTs to filter out small-scale noise. The EOTs were calculated using monthly OISST derived from weekly OISST v2 from 1982 and 2005 in v3b, but from 1982 to 2011 in v4. As shown by Huang et al. (2015a), the selection of EOT training periods leads to sensitivity in the SSTA reconstruction, particularly in the tropical oceans. Therefore, several groups of EOTs are derived: (a) EOTs from three alternative training periods (1982–2005; 1988–2011; 1982–2011), (b) EOTs nondamped in the high latitudes south of 60°S and near 60°N, (c) EOTs from even-year data (1982, 1984, ..., 2012) and odd-year data (1983, 1984, ..., 2013), and (d) EOTs with damping scales of 5000, 4000, and 3000 km in longitude, and 4000, 3000, and 2000 km in latitude to explore the effects of domain truncation.

18) *EOT weighting*: In fitting the HF SSTAs, an EOT mode was weighted by grid box area in ERSSTv3b. Additional weighting of observation number and its associated error is considered in ERSST.v4 (Huang et al. 2015a). Therefore, these two weighting options are used in the parametric uncertainty estimation.

19) *EOT critical value*: Not all 130 EOT modes are actually used to reconstruct HF SSTAs. An EOT is selected if the EOT critical value (Huang et al. 2015a) is higher than a certain criterion. The EOT critical value assesses whether that particular EOT mode is supported or is potentially an artifact. Huang et al. (2015a) showed that the critical value is sensitive in determining the resulting SSTA

reconstruction. The critical value was set to 0.2 in v3b and is set to 0.1 in v4. Therefore, three alternative options of 0.05, 0.1, and 0.2 are set for the parametric uncertainty estimation.

20) *Ice concentration factor*: Ice concentration from HadISST (1870–2010) is used in ERSST.v4, which is approximately 10% higher than the previous UKMO ice concentration in the Northern Hemisphere. The difference between these two versions of the ice concentration data may imply a measure of uncertainty in observing the ice concentration. Therefore, the ice concentration is alternated by multiplying a factor of 0.9, 1.0, and 1.1.

21) and 22) *Min/max ice for SST adjustment*: In ERSST.v4, the combined SST from low- and high- frequency components is adjusted in the ice-covered area when the ice concentration falls between a min and max of 0.6 and 0.9, respectively (Smith and Reynolds 2004). These minimum and maximum values are perturbed by ±0.1 as their alternative options.

23) *LF filter period*: In ERSST, SSTAs are decomposed into LF and HF components. The LF component is retrieved by applying a median 15-yr filter to annually averaged SSTAs. The LF periods are perturbed among 11, 15, and 19 years to include the potential contribution to the SST uncertainty.

24) *HF filter period*: In ERSST, the HF component SSTA is filtered using a 3-month running filter to account for missing superobservations. An alternative option without the filter is added to quantify its impact on the SST uncertainty.

REFERENCES

Bojinski, S., and Coauthors, 2014: The concept of essential climate variables in support of climate research, applications, and policy. *Bull. Amer. Meteor. Soc.*, **95**, 1431–1443, doi:10.1175/BAMS-D-13-00047.1.

Cleveland, W. S., 1981: LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *Amer. Stat.*, **35**, 54–65, doi:10.2307/2683591.

Collins, W. J., and Coauthors, 2008: Evaluation of the HadGEM2 model. Hadley Centre Tech. Note 74, Met Office, 47 pp. [Available online at www.metoffice.gov.uk/media/pdf/8/7/HCTN_74.pdf.]

Cowtan, K., and Coauthors, 2015: Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophys. Res. Lett.*, **42**, 6526–6534, doi:10.1002/2015GL064888.

Davis, R. E., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **6**, 249–266, doi:10.1175/1520-0485(1976)006<0249:POSSTA>2.0.CO;2.

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, doi: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, doi:10.1002/qj.49712152206.

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

Grumbine, R. W., 2014: Automated passive microwave sea ice concentration analysis at NCEP. NOAA MMAB Tech. Note 321, 39 pp.

Hartmann, D. L., and Coauthors, 2014: Observations: Atmosphere and surface. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 159–254.

Hirahara, S., M. Ishii, and Y. Fukuda, 2014: Centennial-scale sea surface temperature analysis and its uncertainty. *J. Climate*, **27**, 57–75, doi:10.1175/JCLI-D-12-00837.1.

Huang, B., M. L'Heureux, J. Lawrimore, C. Liu, V. Banzon, Z.-Z. Hu, and A. Kumar, 2013: Why did large differences arise in the sea surface temperature datasets across the tropical Pacific during 2012? *J. Atmos. Oceanic Technol.*, **30**, 2944–2953, doi:10.1175/JTECH-D-13-00034.1.

——, and Coauthors, 2015a: Extended Reconstructed Sea Surface Temperature version 4 (ERSST.v4). Part I: Upgrades and intercomparisons. *J. Climate*, **28**, 911–930, doi:10.1175/JCLI-D-14-00006.1.

——, W. Wang, C. Liu, V. Banzon, H.-M. Zhang, and J. Lawrimore, 2015b: Bias adjustment of AVHRR SST and its impacts on two SST analyses. *J. Atmos. Oceanic Technol.*, **32**, 372–387, doi:10.1175/JTECH-D-14-00121.1.

Kaplan, A., M. A. Cane, Y. Kushnir, A. C. Clement, M. B. Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res.*, **103**, 18 567–18 589, doi:10.1029/97JC01736.

Karl, T. R., and Coauthors, 2015: Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, **348**, 1469–1472, doi: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, doi: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, doi: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, doi:10.1029/2010JD015220.

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, doi:10.1175/JTECH1844.1.

——, J. J. Kennedy, D. I. Berry, and R. O. Smith, 2010: Effects of instrumentation changes on sea surface temperature measured in situ. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 718–728, doi:10.1002/wcc.55.

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

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, doi:10.1029/2011JD017187.

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 temperatures since the late nineteenth century. *J. Geophys. Res.*, **108** (D14), 4407, doi:10.1029/2002JD002670.

Reynolds, R. W., 2009: What's new in version 2. [Available online at http://www.ncdc.noaa.gov/sst/papers/oisst_daily_v02r00_version2-features.pdf.]

——, and T. M. Smith, 1994: Improved global sea surface temperature analyses using optimum interpolation. *J. Climate*, **7**, 929–948, doi: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 for climate. *J. Climate*, **15**, 1609–1625, doi:10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2.

——, T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, 2007: Daily high-resolution blended analyses for sea surface temperature. *J. Climate*, **20**, 5473–5496, doi:10.1175/2007JCLI1824.1.

Shen, S. S. P., T. M. Smith, C. F. Ropelewski, and R. E. Livezey, 1998: An optimal regional averaging method with error estimates and a test using tropical Pacific SST data. *J. Climate*, **11**, 2340–2350, doi:10.1175/1520-0442(1998)011<2340:AORAMW>2.0.CO;2.

——, A. N. Basist, G. Li, C. Williams, and T. R. Karl, 2004: Prediction of sea surface temperature from the Global Historical Climatology Network data. *Environmetrics*, **15**, 233–249, doi:10.1002/env.638.

——, H. Yin, and T. M. Smith, 2007: An estimate of the error variance of the gridded GHCN monthly surface air temperature data. *J. Climate*, **20**, 2321–2331, doi:10.1175/JCLI4121.1.

——, N. Tafolla, T. M. Smith, and P. A. Arkin, 2014: Multivariate regression reconstruction and its sampling error for the quasi-global annual precipitation from 1900 to 2011. *J. Atmos. Sci.*, **71**, 3250–3268, doi:10.1175/JAS-D-13-0301.1.

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, doi:10.1175/1520-0442-16.10.1495.

——, and ——, 2004: Improved extended reconstruction of SST (1854–1997). *J. Climate*, **17**, 2466–2477, doi: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, doi:10.1175/JCLI3362.1.

——, ——, R. E. Livezey, and D. C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate*, **9**, 1403–1420, doi:10.1175/1520-0442(1996)009<1403:ROHSST>2.0.CO;2.

——, ——, 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, doi:10.1175/2007JCLI2100.1.

van den Dool, H. M., S. Saha, and A. Johansson, 2000: Empirical orthogonal teleconnections. *J. Climate*, **13**, 1421–1435, doi:10.1175/1520-0442(2000)013<1421:EOT>2.0.CO;2.

Woodruff, S. D., and Coauthors, 2011: ICOADS release 2.5: Extensions and enhancements to the surface marine meteorological archive. *Int. J. Climatol.*, **31**, 951–967, doi:10.1002/joc.2103.