Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4): Part II. Parametric and Structural Uncertainty Estimations

WEI LIU,* BOYIN HUANG,+ PETER W. THORNE,# VIVA F. BANZON,+ HUAI-MIN ZHANG,+
ERIC FREEMAN,@ JAY LAWRIMORE,+ THOMAS C. PETERSON,+ THOMAS M. SMITH,& AND
SCOTT D. WOODRUFF**

** Cooperative Institute for Climate and Satellites, North Carolina State University, Raleigh, and NOAA/National Climatic Data Center, Asheville, North Carolina*
+ NOAA/National Climatic Data Center, Asheville, North Carolina
# Nansen Environmental and Remote Sensing Center, Bergen, Norway
@ NOAA/National Climatic Data Center, Asheville, North Carolina, and STG, Inc., Reston, Virginia
& NOAA/STAR/SCSB, and CICS/ESSIC, University of Maryland, College Park, Maryland
*** NOAA/National Climatic Data Center, Asheville, North Carolina, and Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado*

(Manuscript received 30 December 2013, in final form 4 November 2014)

ABSTRACT

Described herein is the parametric and structural uncertainty quantification for the monthly Extended Reconstructed Sea Surface Temperature (ERSST) version 4 (v4). A Monte Carlo ensemble approach was adopted to characterize parametric uncertainty, because initial experiments indicate the existence of significant nonlinear interactions. Globally, the resulting ensemble exhibits a wider uncertainty range before 1900, as well as an uncertainty maximum around World War II. Changes at smaller spatial scales in many regions, or for important features such as Niño-3.4 variability, are found to be dominated by particular parameter choices.

Substantial differences in parametric uncertainty estimates are found between ERSST.v4 and the independently derived Hadley Centre SST version 3 (HadSST3) product. The largest uncertainties are over the mid and high latitudes in ERSST.v4 but in the tropics in HadSST3. Overall, in comparison with HadSST3, ERSST.v4 has larger parametric uncertainties at smaller spatial and shorter time scales and smaller parametric uncertainties at longer time scales, which likely reflects the different sources of uncertainty quantified in the respective parametric analyses. ERSST.v4 exhibits a stronger globally averaged warming trend than HadSST3 during the period of 1910–2012, but with a smaller parametric uncertainty. These global-mean trend estimates and their uncertainties marginally overlap.

Several additional SST datasets are used to infer the structural uncertainty inherent in SST estimates. For the global mean, the structural uncertainty, estimated as the spread between available SST products, is more often than not larger than the parametric uncertainty in ERSST.v4. Neither parametric nor structural uncertainties call into question that on the global-mean level and centennial time scale, SSTs have warmed notably.

1. Introduction

Sea surface temperature (SST) is a fundamental variable in climate studies and climate monitoring (e.g., Hartmann et al. 2013; Blunden and Arndt 2013). Consequently, a succession of historical observed SST analyses have been produced by different groups, including Smith et al. (1996, 2008), Smith and Reynolds (2003, 2004), Kaplan et al. (1998), Rayner et al. (2003, 2006), Kennedy et al. (2011a,b), Ishii et al. (2005), and Hirahara et al. (2014). Now an updated version of the monthly Extended Reconstructed Sea Surface Temperature (ERSST), version 4 (ERSST.v4), is available, as presented in a companion paper (Huang et al. 2015, hereinafter Part I). This is a global monthly dataset on a spatial 2° × 2° grid, covering January 1875 onward.

In ERSST.v4, significant improvements were achieved relative to the previous ERSST versions (Smith and Reynolds 2003, 2004; Smith et al. 2008) by tuning against

Corresponding author address: W. Liu, Scripps Institution of Oceanography, CASPO, UCSD, 9500 Gilman Drive, La Jolla, CA 92093.
E-mail: wel109@ucsd.edu

DOI: 10.1175/JCLI-D-14-00007.1

TABLE 1. Parameter settings in ERSST.v4 operational and ensemble runs. A total of 9 of the changed parameters included in ERSST.v4 (Part I) have been varied, and for each parameter, 2–4 options are possible (operational product settings and 1–3 alternates). These 9 parameters can be categorized into two groups: observation-related and system-dependent parameters. Parameters 1 and 3 belong to the former, while the others belong to the latter. The operational run in ERSST.v4 is conducted by using the first selection of each of the parameters shown in the table. Meanwhile, 100-ensemble runs are carried out via a Monte Carlo ensemble approach in which a random sampling is repeated until achieving 100 unique sets of parameter combinations, based on a probability weighting on each parameter option, in the form of percentage (given in parentheses—in each case, the ensemble will, on average, sample the ERSST.v4 setting more than the alternates). Note, here bias adjustments prior to 1886 are set as the annual cycle in 1886, since the NMAT data in HadNMAT2 and UKMO NMAT are not deemed reliable before 1886 by the dataset creators (Kent et al. 2013).

	Parameter	ERSST.v4 operational	Option 1	Option 2	Option 3
1	SST STD for QC	From OISST.v2 1982–2011 (70%)	From COADS 1950–79 (30%)	—	—
2	SSTA calculation in QC	On an in situ basis (70%)	On a gridbox basis (30%)	—	—
3	NMAT for bias correction*	HadNMAT2 (70%)	UKMO NMAT (30%)	—	—
4	Bias correction smoothing	$f = 0.10$ (40%)	$f = 0.05$ (20%)	$f = 0.20$ (20%)	Linear as v3b (20%)
5	Ship–buoy adjustment	0.12 (50%)	0.10 (25%)	0.14 (25%)	—
6	LF anomaly filling	Nearby anomaly filling (70%)	Zero-anomaly filling (30%)	—	—
7	EOT training period	1982–2011 (50%)	1982–2005 (25%)	1988–2011 (25%)	—
8	EOT weighting	$W = N/(N + \xi^2)\cos\varphi$ (70%)	$W = \cos\varphi$ (30%)	—	—
9	EOT critical value	0.10 (50%)	0.08 (25%)	0.12 (25%)	—

* Adjustment is linear before 1886 using 1886 adjustment.

several plausible options in the selection of several parameters to alight on the final operational algorithm configuration. The parametric uncertainty outlined here is assessed from varying those parameters of the ERSST algorithm modified in Part I and serves to complement the operational version and increase its utility to end users. As noted in Kennedy (2014), there exist many remaining uncertainties and gaps in our SST knowledge, which it is important that, as a global community, we critically assess. The central theme of this paper is to analyze the parametric uncertainty in ERSST.v4 and outline some of its potential applications. Comparisons are made to an equivalent product called the Met Office (UKMO) Hadley Centre SST version 3 dataset (HadSST3; described in section 4), particularly the ensemble that captures its parametric uncertainty in a number of distinct parameters. Finally, we also compare to other long-term SST products to inform an estimate of the current structural uncertainty in SST records.

The remainder of this paper is organized as follows. First, in section 2, we briefly summarize those aspects of the v4 algorithm that are important for this discussion of the parametric uncertainty in the product (see Part I for a full description). In section 3, we examine the sensitivity of individual parameters and then test the nonlinear effect in the combination of multiple parameters. Owing to a demonstrated strong nonlinearity between parameters, we then carry out an ensemble analysis to quantify the parametric uncertainty in ERSST.v4. Subsequently, we make a comparison of parametric uncertainties between ERSST.v4 and the HadSST3 in section 4. Section 5 analyzes structural uncertainty in SST estimates and how this compares to the parametric uncertainties derived in section 3. A discussion is provided in section 6, and section 7 concludes.

2. ERSST.v4 algorithm and parameters

A complete methodological description and justification for the parameter choices made in the operational version of ERSST.v4, along with a comparison to several other datasets, are given in Part I. In this section, we briefly highlight those parameter choices described in Part I, except for the SST and ice data, and we also outline the alternatives that were considered during its development. These include the ERSST.v3b choices and also, in some cases, additional plausible parameter choices. These alternative choices form the basis for the parametric uncertainty estimation derived and analyzed herein. The choices are also tabulated in Table 1, which provides a breakdown for the frequency that each choice is assigned in the ensemble. The following description is in the chronological order in which the operational algorithm undertakes the analysis. Whether these sequential processing choices interact is addressed in section 3a.

The input data used in ERSST.v4 are selected from release 2.5 (R2.5) of the International Comprehensive Ocean–Atmosphere Data Set (ICOADS; Woodruff et al. 2011) through 2007, and then from 2008 forward from Global Telecommunication System (GTS) receipts gathered by the NOAA's National Centers for Environmental Prediction (NCEP). We use R2.5 to construct the parametric uncertainty estimates herein because of demonstrably improved data completeness, enhanced duplicate removal, and other procedures over the previous R2.4 release, which formed the basis for ERSST.v3b. It was felt that use of R2.4 did not constitute a reasonable parametric choice, partly because as input data it is external to the algorithm itself, and partly because R2.5 was demonstrably better and hence that this should not inform the parametric uncertainty estimation herein.

Next, based on a monthly SST climatology on a spatial 2° × 2° grid for 1971–2000 (Xue et al. 2003; note here this reference climatology is prior defined, which may be different from the 1971–2000 SST climatology of the final reconstructed data), individual observations are screened via a quality control (QC) procedure (Smith and Reynolds 2003, 2004). In QC, the SST anomaly (SSTA) is calculated on an in situ basis (before computation of the gridbox average; parameter 2 in Table 1) and extreme values [greater than 4 times the standard deviation (STD)] are excluded based on the monthly STD of Optimum Interpolation SST, version 2 (OISST.v2; Reynolds et al. 2002) for the 1982–2011 base period (parameter 1 in Table 1). For ERSST.v3b these same operations were achieved through consideration on a gridbox basis with comparison to STD estimates from Comprehensive Ocean–Atmosphere Data Set (COADS) records from 1950 to 1979. The two STD climatologies are included as possible parameter choices herein.

Following QC, the algorithm applies bias adjustments. First, ship SST measurements are adjusted through reference to the most recent nighttime marine air temperatures (NMAT) from the Hadley Centre (HadNMAT2; Kent et al. 2013; parameter 3 in Table 1; note that HadNMAT2 also uses ICOADS R2.5 for construction) using a modified scheme from that of Smith and Reynolds (2002). In particular, a locally weighted scatterplot smoothing (LOWESS) filter (Cleveland 1981) with a smoothing parameter $f = 0.1$ is applied to eliminate variations on time scales shorter than decades in the calculation of annual coefficients (parameter 4 in Table 1). ERSST.v3b used COADS NMAT (an earlier dataset with less fully developed corrections and QC checks on NMAT) and applied a simple smoothing scheme to bias corrections (linear regression on annual coefficients pre-1942 and a zero annual coefficient post-1942). This created a correction with a sharp step (see Fig. 6 in Part I). In ERSST.v4 development, another recent NMAT dataset (UKMO NMAT; Parker et al. 1995) and alternative LOWESS filter values (0.05 and 0.2) were tested. Here both UKMO and HadNMAT2 NMAT datasets and four smoothing options (three LOWESS filter values and linear plus step) are considered for the error estimation.

Next, ship–buoy bias adjustment was undertaken by adding 0.12°C to all drifting and moored buoy SSTs (parameter 5 in Table 1). This value was realized through considering all nearby pairs of data points of ship and buoy measurements. No similar correction was undertaken for ERSST.v3b, but the spread of pairwise estimates enables us to explore uncertainty to this choice. Here we use the best estimate ±0.5 STD (0.02°C) as alternative values. We note that this is somewhat smaller than the range of published estimates of the effect, which have undertaken a range of approaches to determining its value and have tended to range between 0.12° and 0.18°C with varying uncertainty estimates (see discussion in Part I). We prefer to base our parametric uncertainty estimates upon the analysis that informed ERSST.v4 development for consistency.

Then the merged ship and buoy SSTAs are analyzed separately for the low- (>15 yr) and high-frequency (<15 yr) components. The low-frequency (LF) component is constructed by averaging and filtering data over a spatial–temporal region (Smith et al. 2008). In areas or periods with sparse data, the LF anomaly is set to a nearby value (parameter 6 in Table 1). In ERSST.v3b, this infilling was instead a zero-anomaly (climatological average) infilling that implicitly assumed no underlying change and would tend to damp anomalies away from the climatology period were there a transient change in the system. To determine sensitivity to this assumption, both parameter options are considered here.

The high-frequency (HF) analysis uses a set of anomaly-increment modes, or spatial patterns, computed using empirical orthogonal teleconnections (EOTs; Van den Dool et al. 2000). Based on a training period of 1982–2011 (parameter 7 in Table 1), a maximum of 130 EOTs were used (Smith and Reynolds 2003), with screening to eliminate any modes not adequately sampled (<10% of the variance of the mode). In data-rich periods, over 120 EOTs are generally selected in ERSST.v4. In data-sparse periods, this drops to as low as 100 and, prior to 1875, even lower still. For the selected modes, a weight for each mode is found by fitting the set of modes to the superobservations (defined as the average of all input data over a given grid box for a given month) and including an additional EOT weighting (Reynolds et al. 2002) such that $W = N/(N + \xi^2)\cos\varphi$, where N is the sum of the record numbers of ship (N_s) and buoy (N_b) in superobservations, ξ is average error value in observations, and φ denotes latitude (parameter 8 in Table 1). The HF component is computed from the weighted sum of the returned EOT modes.

All three of these EOT steps had multiple choices considered in the development of ERSST.v4, and these choices are explored in the uncertainty estimates here. In addition to 1982–2011, EOT training periods of 1982–2005 (as in ERSST.v3b) and 1988–2011 (same number of years as in v3b, but for a more recent period) are considered. The EOT weighting in ERSST.v3b was solely based upon the $\cos\varphi$ area weighting, and that is considered here as a plausible alternative. Finally, the screening criteria (Crit) were varied around the designated operational value of 0.1 (parameter 9 in Table 1). The lower Crit values than in ERSST.v3b (0.2) are set such as to retain plausibly realistic El Niño/La Niña features in early Niño-3.4 time series (see discussion in Part I). Part I considered alternatives of 0.05 and 0.2 for this parameter, whereas here the alternatives explored are more restricted, being 0.08 and 0.12 (Table 1). As documented by Part I, a value of Crit of 0.05 yields unacceptably high noise levels, and 0.2 unacceptably diminishes variability in key regions, such as Niño 3.4. It is therefore felt that neither is an entirely reasonable choice in creating a parametric uncertainty estimate that should span plausible solutions (section 3).

Finally, the reconstructed SSTA by LF and HF components are merged with sea ice information from the UKMO Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST) (1870–2010; Rayner et al. 2003) and NCEP (2011–12; Grumbine 1996). Note, here sea ice extents and concentrations are uncertain, particularly prior to the satellite period. A prior-defined 1971–2000 SST climatology is used to generate the ERSST.v4 product. None of these steps included reasonable alternatives, so these steps do not contribute to the parametric uncertainty estimates herein.

3. The ERSST.v4 parametric uncertainty estimation

Before delving into the details of the parametric uncertainty estimation, some generic discussion is warranted surrounding more theoretical aspects of such uncertainty estimation and the likely implications for our estimates. First and foremost, such estimates are an emerging field and to date several distinct approaches have arisen (e.g., Kennedy et al. 2011b; Morice et al. 2012; Thorne et al. 2011a; Mears et al. 2011; Williams et al. 2012). These reflect both the importance of such estimates, which allow users to assess the sensitivity of their analyses to observational uncertainties in a more informed manner, and also the real challenges in making such estimates.

Central to the challenges is ensuring that the resulting ensemble is neither overly optimistic nor overly pessimistic and contains the true measured values. In the absence of a robust way to determine this, there exist a myriad of defensible approaches to determining what the plausible ensemble should consist of. In essence, this comes down to determination of four central facets: (i) which parameters to vary; (ii) which ranges to allow the parameters to vary within; (iii) whether the parameters are varied in combination; and (iv) whether to provide a priori weights on different parameter value choices to preferentially choose certain values over others.

While we cannot claim to know the correct way to do this, our approach in creating the parametric uncertainty estimates for ERSST.v4, outlined in the remainder of this section, has been as follows. We begin by assessing the effects of single parameter perturbations (section 3a). Next, we have tested whether it is necessary to account for nonlinearities through single- and paired-parameter perturbation experiments (section 3b). Having determined the necessity to consider such nonlinearities, we then create a 100-member ensemble using a Monte Carlo procedure (section 3c).

Returning to the four essential choices in creating such an ensemble alluded to previously: (i) we vary solely those parameters changed in ERSST.v4 and outlined in section 2; (ii) we take the particular fixed values assigned in Table 1 and justified in section 2 of Part I; (iii) we vary them in combination; and (iv) we provide preferential weight to the operational version configuration of each parameter such that, on average, the operational choice for each parameter is visited more often in the ensemble than any possible alternate is. Implicit in this approach are the following assumptions: that all important parameters were considered in moving from v3b to v4; that the alighted-on v4 settings were more optimal in reality than the possible alternates; and that our alternatives span the full range of plausible choices. If such assumptions are incorrect, then the estimates will be overestimates or (more likely) underestimates.

a. Single-parameter perturbations

To explore the sensitivity of each parameter option, we conducted 14 single-parameter perturbation (SPP) runs (cf., Fig. 1 legend). In each SPP run, only one parameter choice is perturbed, while the other parameters remain the same as in the ERSST.v4 operational run. The SST difference between individual SPP runs and the ERSST.v4 operational run shows the sensitivity of the analysis to a particular parameter, or more specifically, a particular parameter option. Because several parameters have two or more alternates (Table 1), the number of SPP runs is slightly greater than the number of parameters varied.

Figure 1a displays the differences in global and broad latitudinal band annual-mean SSTs between all 14 SPPs and the operational ERSST.v4. In the low frequency, the uncertainty in global scale is most affected by changes in those parameters that relate to bias adjustments and bias

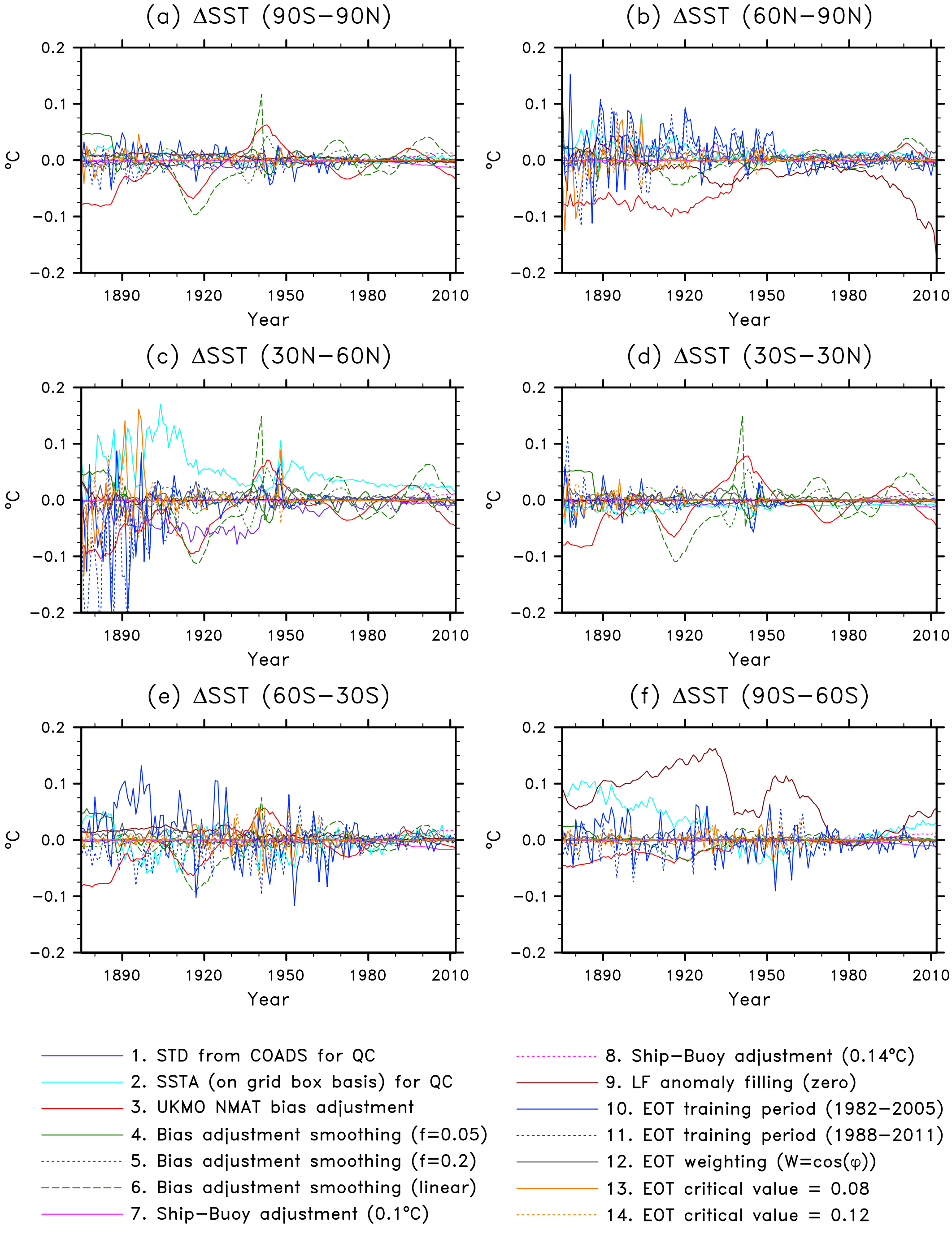


FIG. 1. Differences of averaged annual-mean SST between individual SPP runs and the ERSST.v4 operational run for (a) globe, (b) 60°–90°N, (c) 30°–60°N, (d) 30°S–30°N, (e) 60°S–30°N, and (f) 90°–60°S. In each plot, results from different SPP runs are denoted by distinct colors and line styles, as denoted in the in-line key.

adjustment smoothing choices. Prior to 1886, parameter effects are greatest when using UKMO NMAT bias adjustments instead of HadNMAT2. For the high frequency, the global SST series behavior is highly sensitive to EOT training period and EOT critical value in the early period when data are sparse. Changes in the ship–buoy adjustment and EOT weighting have little influence on global SST.

For the various latitudinal belts (Fig. 1, remaining panels), SST sensitivity can be quite different to global-mean sensitivity. In the low frequency: (i) a change of SST STD for QC has a significant effect on SSTs in

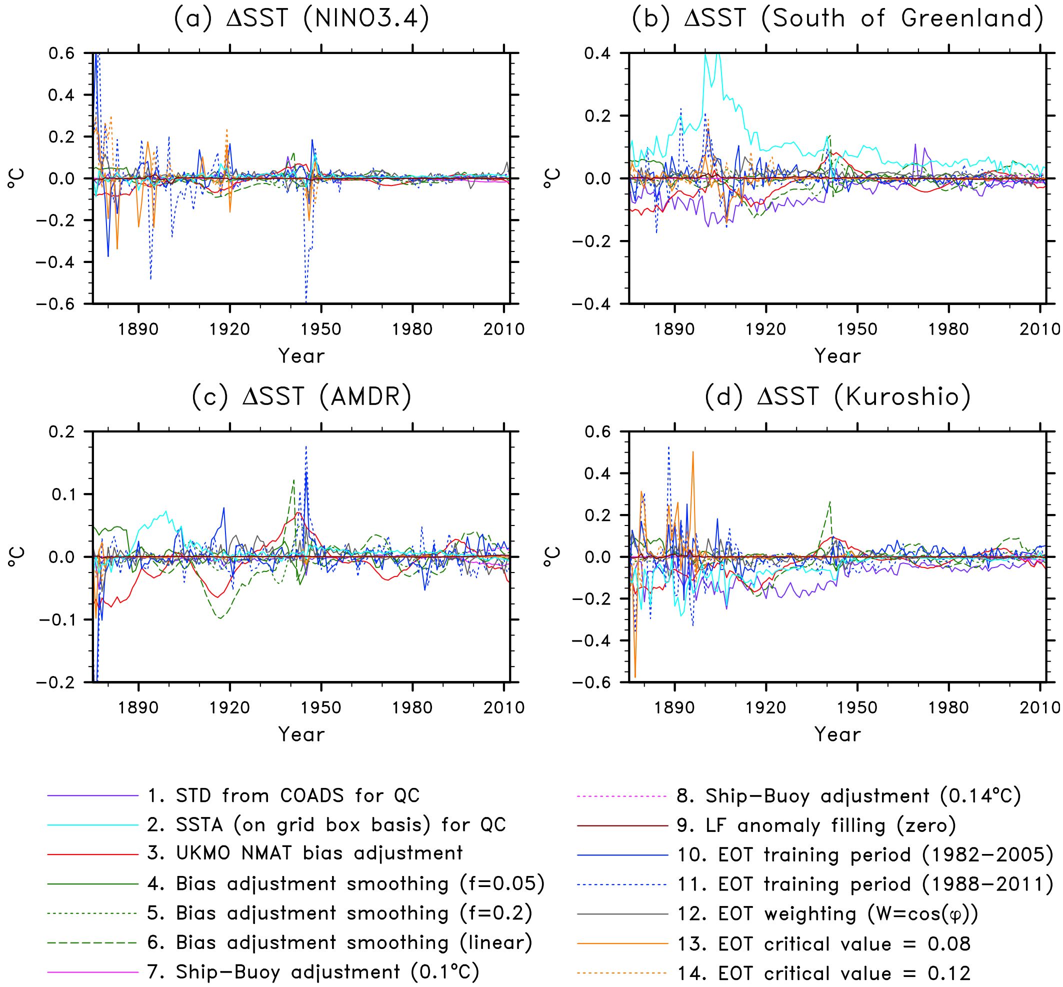


FIG. 2. As in Fig. 1, but for four key regions: (a) the Niño-3.4 region (6°S–6°N, 120°–170°W), (b) the region south of Greenland (40°–60°N, 24°–56°W), (c) the AMDR (10°–20°N, 30°–60°W), and (d) the Kuroshio region south of Japan (24°–34°N, 130°–146°E).

30°–60°N prior to 1942, whereas a change of the SSTA calculation for QC has effects mostly in 30°–60°N and 90°–60°S; (ii) A change to bias adjustment smoothing with a linear scheme, as in ERSST.v3b, produces a negative SST departure between 1910 and 1930 and a positive spike at 1942 within 60°S–60°N; and (iii) the LF anomaly filling becomes dominant in the polar regions and is the dominant parameter choice in the region 90°–60°S through the whole period and in 60°–90°N after 2000. In the high frequency, SST is more sensitive to the EOT training period than to the EOT critical value over 90°–30°S.

Additionally, we examine the parameter sensitivity over four regions of particular interest (Fig. 2): the Niño-3.4 area (6°S–6°N, 120°–170°W), the area south of Greenland (40°–60°N, 24°–56°W), the Atlantic main development region for hurricanes (AMDR; 10°–20°N, 30°–60°W) and the Kuroshio region south of Japan (24°–34°N, 130°–146°E). In the Niño-3.4 area, SST is sensitive to EOT training periods and EOT critical values. Alterations to these two parameters will substantially modulate the Niño-3.4 SST on the interseasonal scale and then influence the recording of ENSO events when sampling is sparse (see also the discussion of EOT effects in this region in Part I). In the area south of Greenland, the SSTA is subject to effects from a variety of parameter options, although parameters of EOT training period, EOT critical value, and the SSTA calculation for QC seem to be more prominent. None of these parameter variations substantially alter the presence of a prolonged local temperature minimum in this region in the late twentieth century.

Over the AMDR, SST is most affected by EOT training period and EOT critical value, especially prior to 1886. In the Kuroshio region, the situation becomes more complex because the SST seems sensitive to all the parameters except EOT weighting and ship–buoy

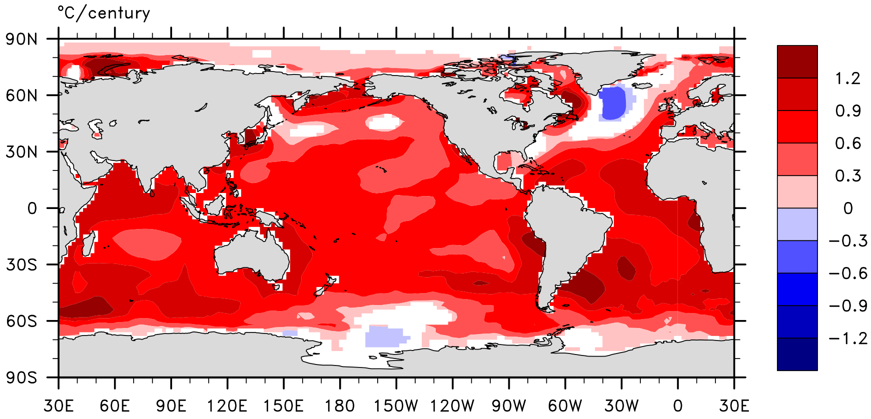


FIG. 3. Global distribution of the SST warming trend (1910–2012) of the ERSST.v4 operational run. The trend is calculated from monthly data and only illustrated when it exceeds a 95% significance based on a two-tailed Student's *t* test.

adjustment. It is worth noting that the magnitude of SST differences for geographical subregions (latitudinal belts or four key regions) is often larger than in the global mean, since local parameter effects will frequently cancel when averaged over the globe.

It follows that there exists an influence of parameter changes on the SST warming pattern and that parametric uncertainty is an important component of the uncertainty to consider when using and analyzing ERSST.v4. Figure 3 displays the global pattern of SST trend during 1910–2012 from the ERSST.v4 operational run for comparison purposes. A general warming dominates over the globe, except for a slight but significant cooling occurring to the north of the Ross Sea and to the south of Greenland, two regions for deep water formation. Such warming minimum/cooling has been attributed to deep mixing and convection there, with implications for deep ocean heat uptake (e.g., Gregory 2000; Huang et al. 2003). Figure 4 presents differences from this field arising from the 14 SPP runs (note different color bar axis).

For data quality control steps, employing an SST STD from COADS (Fig. 4a) will serve to greatly amplify the warming in the Kuroshio and Gulf Stream regions and moderately enhance warming in the Gulf of Alaska and around the southern tip of Africa, while reducing the warming in the Bering Sea and to the south of Greenland. With an SSTA calculated on a gridbox basis (Fig. 4b), the warming is much reduced in the Sea of Okhotsk, the Gulf Stream area, and off Wilkes Land of Antarctica in the Southern Ocean, but is greatly enhanced over the Kuroshio area, the Greenland Sea, and the South Pacific.

For bias correction steps, a switch to UKMO NMAT causes a stronger warming to the east of Greenland, over the Greenland Sea and the Barents Sea (Fig. 4c). Also, the warming is somewhat strengthened in the Southern Ocean and several marginal seas around Japan and the Bering Strait. For the bias adjustment smoothing, a change to smoothing parameter $f = 0.05$ or $f = 0.2$ (Figs. 4d,e) does not significantly modify the warming trend over the globe; however, a switch to the linear scheme (Fig. 4f) will increase the warming globally, especially in the northwestern Pacific, the eastern Greenland Sea, and the western Barents Sea. A ship–buoy adjustment with 0.1°C (0.14°C) will uniformly but very slightly reduce (enhance) the warming trend over the global scale (Figs. 4g,h).

For filtering and infilling step parameter choices (Fig. 4i), an adoption of zero LF anomaly filling is most prominent in reducing the warming in the polar regions (i.e., in the Southern Ocean and in the Arctic). In another aspect, the switch of EOT training period will change the base functions in EOTs so that the SSTAs are represented differently by different sets of EOTs. Figures 4j–l show that changes in the EOT training period and EOT weighting function seem to add noise to the trends. The choice of EOT critical value generally has a small effect on warming over the globe (Figs. 4m,n).

Figure 5 illustrates the dependence of the global-mean SST trend on parameters from the SPP runs. From 1910 to 2012, the period of reasonable global data coverage, the global-mean SST has a warming trend of 0.704°C $\mathrm{century}^{-1}$ in the operational run (Table 2). Of the 14 parameter options, 8 would increase the linear warming trend, as shown in Fig. 4, with a greatest increase of 0.072°C $\mathrm{century}^{-1}$, by adopting the bias adjustment smoothing with a linear scheme (as employed in ERSST.v3b). The remaining 6 parameter options would decrease the trend, with a maximal reduction by 0.014°C by a switch to a zero LF anomaly filling method (again, as used in ERSST.v3b).

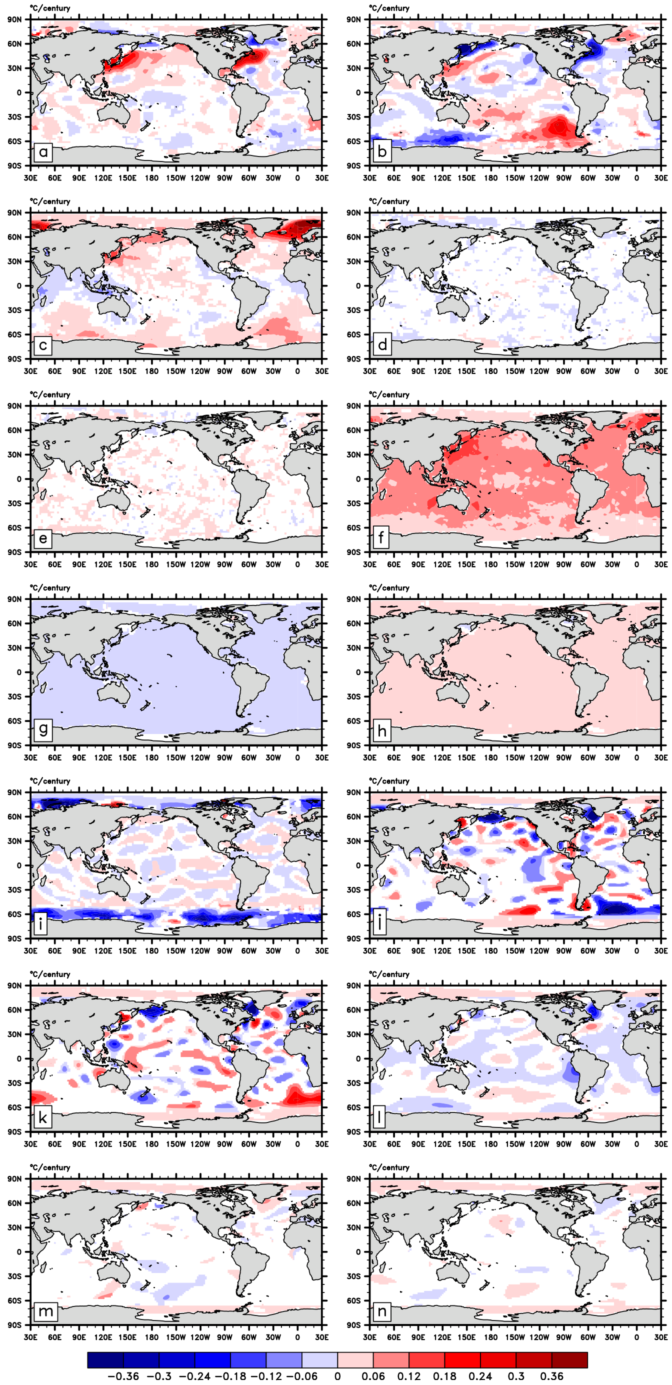


FIG. 4. Maps of differences of the SST warming trend (1910–2012) between individual SPP runs and the ERSST.v4 operational run. (a)–(n) Results from the following SPP runs: (a) STD from COADS for QC, (b) SSTA on gridbox basis for QC, (c) UKMO NMAT bias adjustment, (d) bias adjustment smoothing ($f = 0.05$), (e) bias adjustment smoothing ($f = 0.2$), (f) bias adjustment smoothing (linear), (g) ship–buoy adjustment (0.1°C), (h) ship–buoy adjustment (0.14°C), (i) LF anomaly filling (zero), (j) EOT training period (1982–2005), (k) EOT training period (1988–2011), (l) EOT weighting ($W = \cos\varphi$), (m) EOT critical value = 0.08, and (n) EOT critical value = 0.12. In all plots, results are calculated from monthly data, shown in color, and only illustrated when they exceed a 95% significance based on a two-tailed Student's t test. Note that the color bar key is significantly compressed relative to Fig. 3.

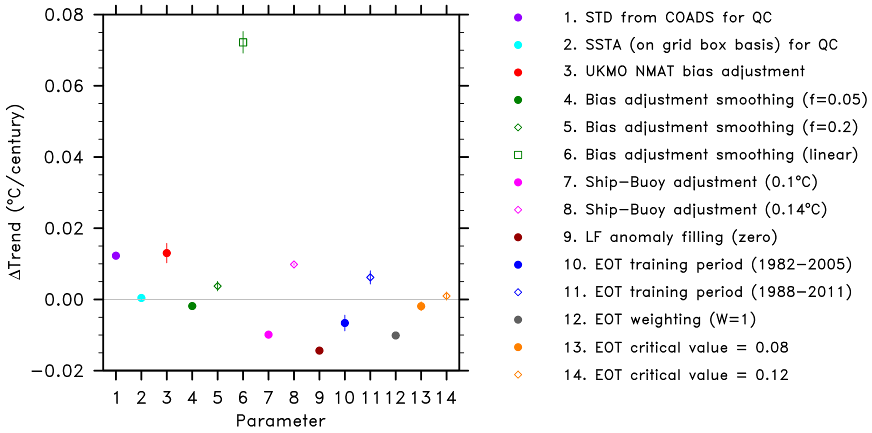


FIG. 5. Differences in global-mean linear trend estimates (with error bar) for 1910–2012 between individual SPP runs and the ERSST.v4 operational run. SPP runs are denoted in the in-line key. The trends are calculated from monthly data.

b. *Testing for nonlinearity*

From the analysis in the preceding subsection, ERSST.v4 exhibits distinct sensitivities to each of the 9 parameters and 14 alternative settings over different regions and different time scales. The simplest way to obtain a parametric uncertainty estimate would be to sum all the component parameter effects from SPP runs under the assumption that there exist only linear combination effects among parameters. However, if nonlinearity exists in parameter combinations, it would require a Monte Carlo ensemble approach (Binder and Heerman 1992) to fully sample the parametric uncertainty. Therefore, it is necessary to test for nonlinearity among multiple parameter perturbations.

To examine the nonlinearity among parameter combinations, we conduct all 85 potential double-parameters perturbation (DPP) runs, in which only 2 are perturbed from the operational settings. Then we compare the sum of comparable SPP and DPP results to test for nonlinearity effects.

First, we name the run with the *i*th parameter perturbed as run SPP_i, then the differential SST δ_i between run SPP_i and the operational run for an individual monthly 2° area situated at point x in year y and month m is defined as

$$\delta_i(x,m,y) = \mathrm{SST}_i(x,m,y) - \mathrm{SST}(x,m,y). \qquad (1)$$

Similarly, the differential SST from another SPP run (SPP_j, $i \neq j$) with the *j*th parameter perturbed is

$$\delta_j(x,m,y) = \mathrm{SST}_j(x,m,y) - \mathrm{SST}(x,m,y). \qquad (2)$$

TABLE 2. The global-mean SST linear trend estimates in ERSST.v4 operational and SPP runs. All the trends are calculated from monthly data. Trends in SPP runs are shown in descending order of trend magnitude. The differential warming trend between each individual SPP and the operational run is shown in the last column.

Parameter setting and changes	SST trend (°C century^{-1})	ΔSST trend (°C century^{-1})
ERSST.v4 operational run	0.704	—
6. Bias adjustment smoothing (linear)	0.776	0.072
3. UKMO NMAT bias adjustment	0.717	0.013
1. STD from COADS for QC	0.716	0.012
8. Ship–buoy adjustment (0.14°C)	0.714	0.010
11. EOT training period (1988–2011)	0.710	0.006
5. Bias adjustment smoothing (f = 0.2)	0.708	0.004
14. EOT critical value = 0.12	0.705	0.001
2. SSTA (on grid basis) for QC	0.704	0.000
4. Bias adjustment smoothing (f = 0.05)	0.702	−0.002
13. EOT critical value = 0.08	0.702	−0.002
10. EOT training period (1982–2005)	0.697	−0.007
7. Ship–buoy adjustment (0.1°C)	0.694	−0.010
12. EOT weighting ($W = \cos\varphi$)	0.694	−0.010
9. LF anomaly filling (zero)	0.690	−0.014

TABLE 3. The ratio r_{ij} (see text for more details) in DPP runs. In ERSST.v4, besides the operational setting, 9 parameters (P1–P9) and 14 parameter options (index 1–14, as shown in Table 2) are considered. The nonlinearity between any two parameters is evaluated by r_{ij} with a threshold of 0.1 [i.e., $r_{ij} \leq 0.1$ implies linearity (values in italics), while $r_{ij} > 0.1$ implies nonlinearity]. Note the following: (i) linearity is evaluated between ship–buoy adjustment (P5, option 7–8) and EOT weighting (P8, option 12), albeit a $r_{ij} = 0.12$ that is slightly higher than but not significantly different from the 0.1 threshold, especially considering an obvious gap between this combination and all nonlinear combinations. (ii) The ratio r_{ij} corresponds to a symmetric matrix, so the top right part shows the value of r_{ij}, and the bottom left part denotes assigned linearity (L) or nonlinearity (N) in parameter combinations.

	P1	P2	P3	P4			P5		P6	P7		P8	P9	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		0.63	0.54	0.52	0.51	0.49	0.18	0.18	0.41	0.44	0.47	0.43	0.50	0.50
2	N		0.51	0.47	0.45	0.51	0.16	0.16	0.43	0.53	0.54	0.42	0.59	0.60
3	N	N		0.78	0.65	0.94	0.22	0.22	0.42	0.31	0.29	0.28	0.76	0.82
4	N	N	N				0.27	0.26	0.37	0.25	0.26	0.25	0.34	0.37
5	N	N	N				0.31	0.32	0.27	0.23	0.24	0.23	0.31	0.37
6	N	N	N				0.24	0.25	0.33	0.29	0.29	0.26	0.40	0.39
7	N	N	N	N	N	N			*0.12*	*0.06*	*0.06*	*0.12*	*0.02*	*0.03*
8	N	N	N	N	N	N			*0.12*	*0.07*	*0.06*	*0.12*	*0.02*	*0.02*
9	N	N	N	N	N	N	L	L		0.24	0.25	0.19	0.29	0.30
10	N	N	N	N	N	N	L	L	N			0.45	0.80	0.79
11	N	N	N	N	N	N	L	L	N			0.45	0.81	0.77
12	N	N	N	N	N	N	L	L	N	N	N		0.42	0.42
13	N	N	N	N	N	N	L	L	N	N	N	N		
14	N	N	N	N	N	N	L	L	N	N	N	N		

Changing both in a DPP can be denoted as

$$\delta_{ij}(x,m,y) = \mathrm{SST}_{ij}(x,m,y) - \mathrm{SST}(x,m,y). \quad (3)$$

Then the SST difference between a DPP run and its corresponding SPP runs is

$$\varepsilon_{ij}(x,m,y) = \delta_{ij} - (\delta_i + \delta_j). \quad (4)$$

If the influence of changing parameters is linear, then

$$\varepsilon_{ij}(x,m,y) = 0. \quad (5)$$

From Eq. (5), a pure linear combination requires a zero ε_{ij} at any location and at any time. However, in practice, the pure linear case is rare, and many cases are quasi-linear, in which case ε_{ij} is very small but not exactly zero. Thus, we provide a definition of small ε_{ij} for a case to be considered approximately linear. To do that, we start by integrating ε_{ij}^2, δ_{ij}^2, and $(\delta_i + \delta_j)^2$ over space (global) and over time (1875–2012):

$$\sigma_{\varepsilon ij}^2 = \iiint \varepsilon_{ij}^2(x,m,y)\,dx\,dm\,dy; \quad (6)$$

$$\sigma_{ij}^2 = \iiint \delta_{ij}^2(x,m,y)\,dx\,dm\,dy; \quad (7)$$

$$\sigma_{i+j}^2 = \iiint [\delta_i(x,m,y) + \delta_j(x,m,y)]^2\,dx\,dm\,dy; \quad (8)$$

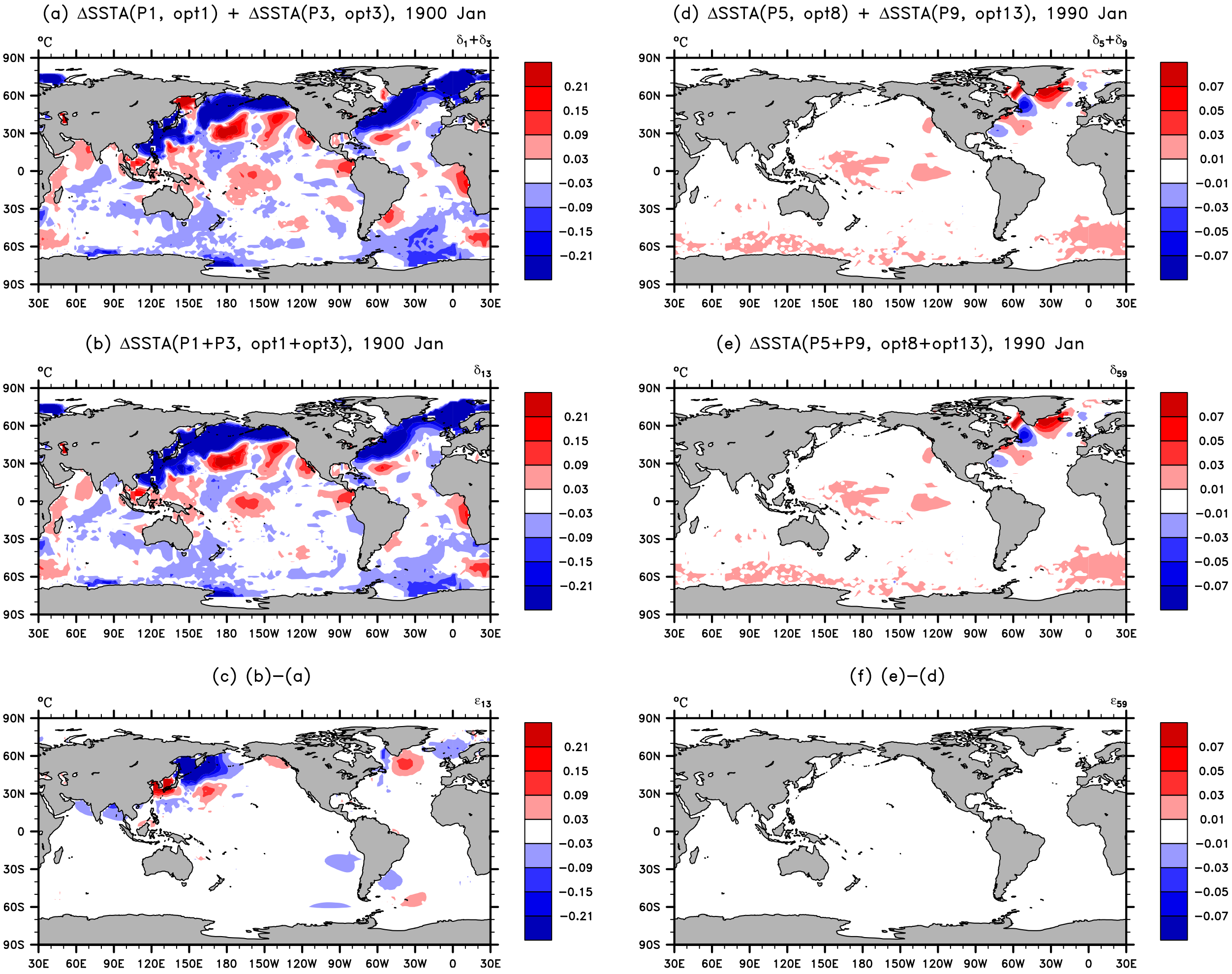
and comparing $\sigma_{\varepsilon ij}^2$ with σ_{ij}^2 or σ_{i+j}^2 by a ratio

$$r_{ij}^2 = \frac{\sigma_{\varepsilon ij}^2}{\min(\sigma_{i+j}^2,\ \sigma_{ij}^2)}, \quad (9)$$

where $\min(\sigma_{i+j}^2,\ \sigma_{ij}^2)$ means the minimum between σ_{ij}^2 and σ_{i+j}^2. Using r_{ij} (here r_{ij} denotes the positive root of r_{ij}^2 so as to represent the ratio of magnitudes), we define that when $r_{ij} \leq 0.1$, there is an approximately linear combination of ith and jth parameters. Thus, if the typical perturbation caused by changing parameters is an order of magnitude larger than the nonlinearity difference, then the combination is said to be approximately linear.

Table 3 summarizes the nonlinearity test results in combinations between any two parameters in ERSST.v4. Most combinations are nonlinear. The largest ratio r_{ij} is equal to 0.94 in a combined perturbation between UKMO NMAT bias adjustment and bias adjustment smoothing with a linear scheme. The only linear combination is between parameters in the ship–buoy adjustment category and parameters in the HF category: that is, ship–buoy adjustment (0.1° or 0.14°C) with EOT training period (1982–2005 or 1988–2011), ship–buoy adjustment (0.1° or 0.14°C) with EOT weighting ($W = \cos\varphi$), and ship–buoy adjustment (0.1° or 0.14°C) with EOT critical value equal to 0.08 or 0.12.

Figure 6 shows an example of the nonlinear/linear parameter combination in which $\delta_i + \delta_j$, δ_{ij}, and ε_{ij} are



(a) ΔSSTA(P1, opt1) + ΔSSTA(P3, opt3), 1900 Jan

(d) ΔSSTA(P5, opt8) + ΔSSTA(P9, opt13), 1990 Jan

(b) ΔSSTA(P1+P3, opt1+opt3), 1900 Jan

(e) ΔSSTA(P5+P9, opt8+opt13), 1990 Jan

(c) (b)−(a)

(f) (e)−(d)

FIG. 6. Examples of the spatial nonlinearity and linearity of parameter combinations: (a)–(c) A nonlinear example of STD calculated from COADS and UKMO NMAT basis for bias adjustments; (d)–(f) the linear example of ship–buoy of 0.14°C and EOT critical sampling at 0.08; (top) the sum of the SPP perturbations; (middle) the equivalent DPP estimate; and (bottom) the difference between these two cases and for linear combinations, which is shown to be almost zero everywhere. The index of parameter (P) follows the index in Table 1, and the index of parameter option (opt) follows the index in Table 2.

mapped over the globe at an arbitrary time point (since ε_{ij} must be around zero at any location and at any time for a linear combination). The nonlinear case is the combined perturbations between parameter options of the SST STD for QC calculated from COADS and the use of UKMO NMAT bias adjustment. As shown in Figs. 6a and 6b, $\delta_i + \delta_j$ and δ_{ij} for January 1900 have rather different patterns in the northwestern Pacific, especially close to the Sea of Okhotsk, so that ε_{ij} is large and comparable with $\delta_i + \delta_j$ and δ_{ij} (Fig. 6c), which indicates significant nonlinear effects by simultaneously altering the QC method and NMAT data for bias adjustment. On the other hand, with combined perturbations between parameter options of ship–buoy adjustment (0.14°C) and EOT critical sampling as 0.08, $\delta_i + \delta_j$ and δ_{ij} in January 1990 share a highly similar pattern over the globe (Figs. 6d,e) such that the difference between the two approaches is at least one order of magnitude smaller than either delta field and negligible, suggesting a linear combination between these two parameter options.

Given the propensity for nonlinear interactions in DPP experiments, it is reasonable to assume that higher-order multiple perturbed parameter (MPP) combinations will be almost ubiquitously nonlinear in nature. Furthermore, because the effects are nonlinear, the SPP perturbations will not be a realistic basis from which to infer likely multiparameter combination effects. In many cases, the effects may cancel; in others, they may be multiplied. This can be thought of as akin to dropping a marble at the top of a very steep mountain, whereby both the small stones and the large rocks may deviate its path and affect its final resting place, and it will not always be the largest rocks that are the greatest determining factor. The only way in which to adequately assess the uncertainty is through sampling, in some meaningful sense, the very large perturbed parameter solution space.

In summary, nonlinearity is extensive in parameter combinations of ERSST.v4, necessitating a Monte Carlo ensemble approach to fully sample the parametric uncertainty. This is unsurprising given the sequential nature of the processing as outlined in section 2. Indeed, based upon a tacitly stated assumption of such nonlinearity existing in dataset construction techniques more generally, many emergent parametric uncertainty estimates for both in situ (e.g., Kennedy et al. 2011b; Morice et al. 2012; Thorne et al. 2011a) and satellite (Mears et al. 2011) data products have used Monte Carlo estimation techniques to quantify their parametric uncertainties. To our knowledge, this is the first time that the need or otherwise for such a step has been formally quantified and proven, at least in an observational climate record reconstruction context for a given algorithmic approach. However, it does not necessarily follow that a Monte Carlo technique is required for all such estimates.

c. *Ensemble design and analysis*

As justified in section 3b, a Monte Carlo ensemble approach is employed to estimate the parametric uncertainty in ERSST.v4. First, we assign each parameter option a weighting that indicates how much chance a given parameter value has to be selected in any given member of the ensemble. From Table 1, the operational parameter option is assigned a larger weighting than the others to make sure that it is the primary parameter value choice. This is to avoid the chances of sampling too many parameter combinations distal from the operational choices outlined and justified in Part I under the assumption that these operational parameter settings are the most reasonable/optimal settings. Parameter combinations are derived by a random sampling that uses a uniform range distribution in the selection of each parameter to obtain 100 unique sets of the 9 parameter combinations. Finally, based on the resulting 100-parameter combination settings, we conduct 100-ensemble runs to create 100 additional ERSST.v4 realizations, most of which have 3–5 parameter perturbations from the operational run.

All 100-ensemble runs share the same reference climatological SST as in the operational run. So to facilitate comparison of the SST evolution on various latitudes, we calculate SSTAs from the ensemble and operational runs relative to the reference climatological SST during 1971–2000 and utilize these SSTA series to study the parametric uncertainty.

Figure 7 illustrates the global SSTA in the ensemble and operational runs. Over the whole period, the ensemble mean of SSTA is generally close to the value of the operational run (Part I). Ensemble ranges are large prior to 1900, with two spikes around World Wars I and II (WWI and WWII), and a rapid decrease after 1942. In general, the parametric uncertainty has narrowed over the globe in the past century. Over latitudinal bands, distinct characteristics are found in parametric uncertainties. The uncertainties in the tropics (30°S–30°N) are relatively low in comparison with those in the mid and high latitudes (north of 30°N or south of 30°S). The uncertainty before 1920 mainly arises in the extratropical Northern Hemisphere (30°–90°N), while after 1920, there is a more uniform distribution over the globe due to an increase of the quantified uncertainty in the Southern Hemisphere (Fig. 9a, next section).

The parametric uncertainty in the Southern Ocean (90°–60°S) exhibits a distinct feature. As shown in Fig. 7g, the ensemble SSTAs in the Southern Ocean visibly bifurcate into two groups before 1935 and within 1955–70, so the ensemble-mean SSTA deviates from the operational value, and the SST uncertainty decreases instead of increasing around WWII. A similar case also happens in the northern North Atlantic and the Arctic (60°–90°N) after 2000. The reason is that the parameter choice of an LF anomaly filling technique overwhelms the other parameters over the polar regions during these periods, and the two choices have a large systematic offset from one another (Figs. 1b,f).

We also examine the parametric uncertainty in those four key regions considered in Fig. 2. In the Niño-3.4 area, ensemble deviations are discernable only prior to 1950 and mainly project onto the higher frequencies (Figs. 8a and 10a), suggesting a prominent parametric uncertainty over the interseasonal time scale and a potential effect in estimation of the magnitude of-historical ENSO events, although the timing of events appears to be robust to these parametric uncertainty estimates (Fig. 8a). Parametric uncertainty appears to have a larger effect on some ENSO events than on others, with particular uncertainties in the events in the late 1940s, around 1920, and prior to 1900 (Fig. 10a). In the area south of Greenland, uncertainty peaks during 1900–10, with a rapid reduction after WWII (Figs. 8b and 10a). Over

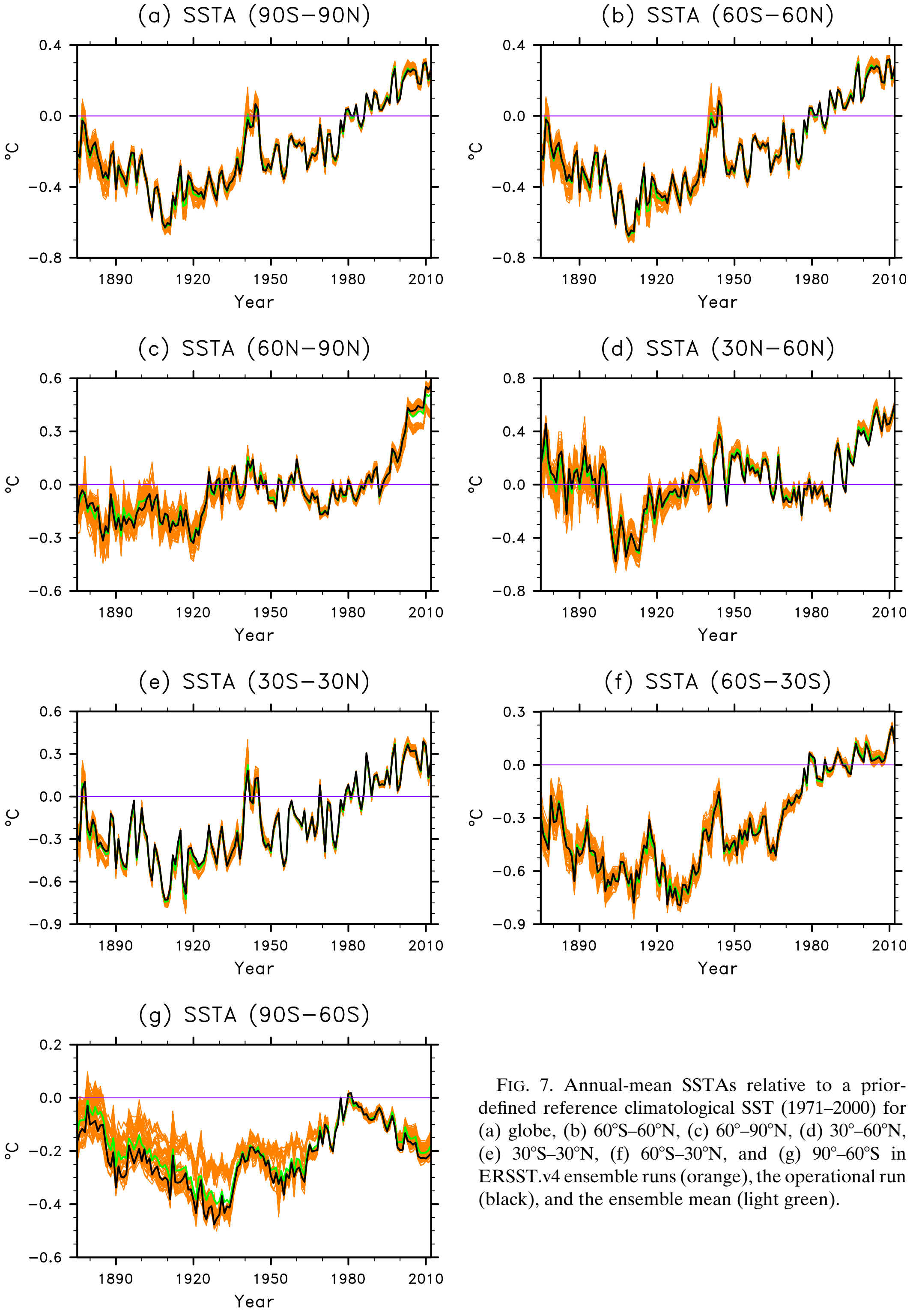


FIG. 7. Annual-mean SSTAs relative to a prior-defined reference climatological SST (1971–2000) for (a) globe, (b) 60°S–60°N, (c) 60°–90°N, (d) 30°–60°N, (e) 30°S–30°N, (f) 60°S–30°N, and (g) 90°–60°S in ERSST.v4 ensemble runs (orange), the operational run (black), and the ensemble mean (light green).

the AMDR, parametric uncertainty is large from the early 1860s to the late 1870s, with a spike in the 1940s (Figs. 8c and 10a). In the Kuroshio region, the uncertainty in SST is greater than in the other three areas, especially during the 1880s and 1890s (Fig. 10a) so that the ensemble-mean SSTA obviously deviates from the operational SSTA in the early stage of the time series (Fig. 8d).

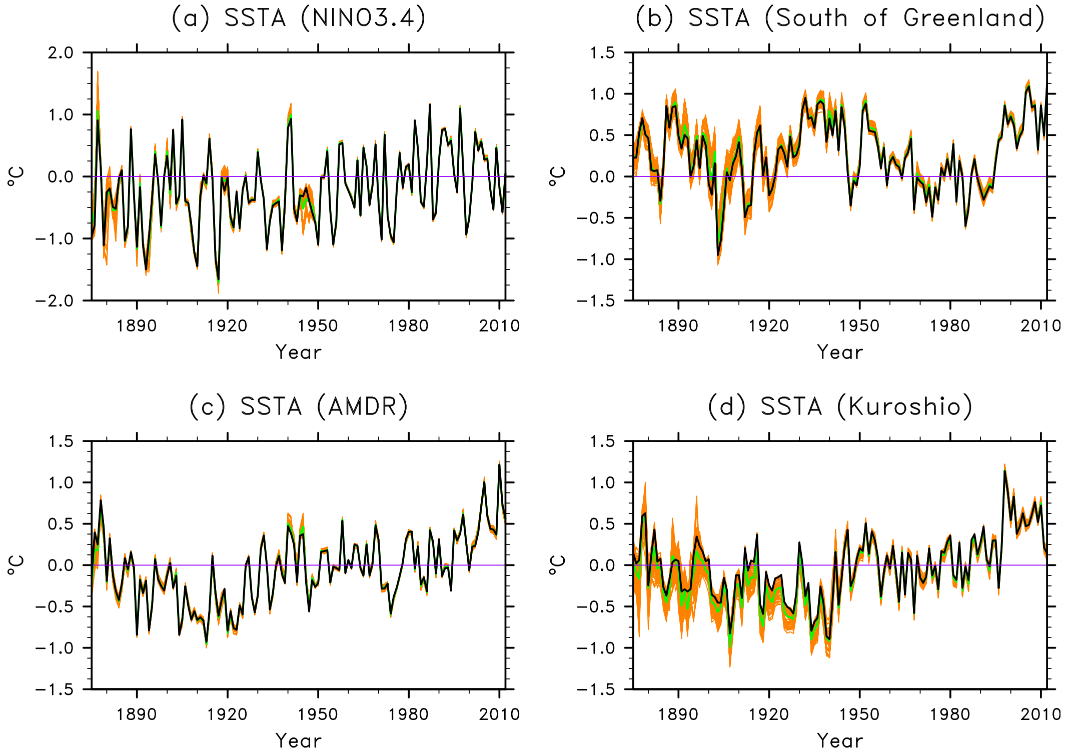


FIG. 8. As in Fig. 6, but for the four key regions (see Fig. 2).

4. Comparisons of ERSST.v4 and HadSST3 ensemble runs

To further evaluate the parametric uncertainty estimates in ERSST.v4, we next compare the ERSST.v4 ensemble runs with those from HadSST3. The HadSST3 parametric uncertainty ensemble is only for the SST bias adjustment components with remaining aspects derived as additional terms outside the ensemble generation process. Similar to ERSST.v4, a suite of 100 ensemble realizations is implemented in HadSST3 with independent parameters and schemes (Kennedy et al. 2011b). These 100 realizations are generated by varying parameters within plausible ranges. For each realization, a new value is taken for each parameter and drawn from a flat prior (this is distinct from our ensemble, which preferentially picks the operational setting on average). Parameters varied in HadSST3 include the data deck–dependent bias adjustments, the bucket adjustments for wooden and canvas buckets, bias from ships using engine room intake (ERI) thermometers and hull-contact sensors, ERI recorded as bucket, canvas to rubber, etc. [see Kennedy et al. (2011b) for more details].

Unlike in ERSST.v4, realizations in HadSST3 are expressed as deviations from the 1960–90 SST climatology of the final reconstructed data, and for each realization of the anomalies, there is a corresponding climatological average for the period 1961–90. As a result, we cannot adopt the method in section 3(c) by simply focusing the analysis on SSTA. Instead, parametric uncertainty in HadSST3 results from both SSTA and the climatological SST.

The other major difference is that HadSST3 is not an infilled product, nor does it undertake any form of smoothing. As a result, there are plenty of missing grid values in HadSST3, especially in the polar regions (90°–60°S and 60°–90°N) and during the early period. Further, many of the ERSST.v4 parameters varied related to the smoothing and infilling steps (section 2, Table 1) for which, obviously, no corollary exists in the HadSST3 ensemble. Equally, several of the HadSST3 parameters that were varied have no corollary in the ERSST.v4 method and, hence, uncertainty estimates.

Although both products are 100-member ensembles, these ensembles are very distinct in their construction and their characteristics. Nevertheless it is still valuable to compare them to understand better our uncertainty in SSTs and the state of parametric uncertainty estimation in SST fields and time series.

In this context, we compare here the resulting parametric uncertainties in ERSST.v4 and HadSST3 from a global- and regional-mean perspective. In the comparison, we referred to "the ensemble median" of the 100 HadSST3 realizations as the "operational run" of

HadSST3 for simplicity. Within each dataset, we first calculate a global or regional-mean SST from individual ensemble runs, and based on that, we then calculate the ensemble deviations. For HadSST3, considering the characteristics of its ensemble runs, we denote its SSTA as a and its reference climatological SST in 1961–90 as c and format the ensemble variances (square of ensemble deviations) of SST as

$$\sigma^2_{(a+c)} = \frac{1}{N_E}\sum_{i=1}^{N_E}[(a_i + c_i) - (\overline{a} + \overline{c})]^2, \qquad (10)$$

where a_i and c_i are SSTA and the reference climatological SST from the ith ensemble run; $\overline{a}$ and $\overline{c}$ are the ensemble mean of SSTA and the reference climatological SST; and N_E is the ensemble number and equal to 100. Equation (10) can be further written as

$$\sigma^2_{(a+c)} = \sigma^2_a + \sigma^2_c - \sigma^2_r, \qquad (11)$$

where

$$\sigma^2_a = \frac{1}{N_E}\sum_{i=1}^{N_E}(a_i - \overline{a})^2, \qquad (12)$$

$$\sigma^2_c = \frac{1}{N_E}\sum_{i=1}^{N_E}(c_i - \overline{c})^2, \quad \text{and} \qquad (13)$$

$$\sigma^2_r = -\frac{1}{N_E}\sum_{i=1}^{N_E}2(a_i - \overline{a})(c_i - \overline{c}). \qquad (14)$$

The terms σ^2_a, σ^2_c, and σ^2_r represent the parametric uncertainties from SSTA, the reference climatological SST, and a residual term, respectively. For ERSST.v4, all ensemble runs share the same reference climatology (i.e., $c_i - \overline{c} = 0$; $\sigma^2_{(a+c)} = \sigma^2_a$) so that parametric uncertainties can be studied from the analysis of SSTA only.

Figure 9 displays the ensemble deviations in ERSST.v4 and HadSST3 over the globe and for latitudinal bands. To facilitate as direct as possible a comparison between the two datasets, we collocate ERSST.v4 on the grid of HadSST3 such that the grid sampling is the same between ERSST.v4 and HadSST3. Since there are substantial missing data over the polar region (outside 60°S–60°N), we limit the comparison to the region 60°S–60°N. From the figure, several differences exist between the parametric uncertainty estimates of these two datasets:

1) Unlike ERSST.v4, parametric uncertainty arises roughly equally from the SSTA and the reference climatological SST for HadSST3. Averaged over the whole period, $\sigma^2_{(a+c)} = 2.154 \times 10^{-3}\,\mathrm{K}^2$, $\sigma^2_a = 1.065 \times 10^{-3}\,\mathrm{K}^2$, and $\sigma^2_c = 1.206 \times 10^{-3}\,\mathrm{K}^2$ (i.e., σ^2_r only accounts for 5.4% of $\sigma^2_{(a+c)}$ and thus is ignorable).
2) ERSST.v4 has a much larger parametric uncertainty at any given time step than HadSST3, both globally and for each latitudinal band. This is understandable, as HadSST3 uncertainty is only from the SST bias adjustment aspects.
3) Parametric uncertainty in ERSST.v4 is maximal in the mid and high latitudes (outside 30°S–30°N), whereas HadSST3 uncertainty maximizes in the tropical regions.
4) In most regions, ERSST.v4 has an enhanced parametric uncertainty around WWII, while HadSST3 has a reduced parametric uncertainty during the same period.

In addition, we compare parametric uncertainties in ERSST.v4 and HadSST3 within four key regions (Fig. 10). As in the analyses for the global and latitudinal bands, the parametric uncertainty of SST is much smaller in HadSST3 than in ERSST.v4 over all four areas. It is interesting to note that, over the Niño-3.4 area, the parametric uncertainty in HadSST3 does not exhibit significant interseasonal time-scale variations, as in ERSST.v4.

Finally, we consider the contribution of the parametric uncertainty to the uncertainty of long-term trends. For ERSST.v4, we compute the trend of the global-mean SST during 1910–2012 (the period of reasonably globally complete coverage) from the ensemble and operational runs in both datasets to construct a box-and-whisker plot (Fig. 11). The warming in the ERSST.v4 operational run is 0.704°C century^{-1}, which is slightly lower than the ensemble median of 0.711°C century^{-1}. Over the 100-ensemble runs, the maximum and minimum warming is 0.796°C century^{-1} and 0.673°C century^{-1}, while the warming trend of the 14 SPP runs is between the maximum and minimum.

For comparison with HadSST3 we limit our calculation within 60°S–60°N and calculate the trend of this regional-mean SST from HadSST3 and ERSST.v4 collocated with the former. The comparison results show that ERSST.v4 has a larger global-mean warming trend than HadSST3 over this period. The warming in the ERSST.v4 operational run is 0.720°C century^{-1}, which is slightly lower than the ensemble median of 0.734°C century^{-1}, whereas the warming in the HadSST3 operational run is 0.641°C century^{-1}, which is slightly higher than the ensemble median of 0.636°C century^{-1}. Unlike the large disparity in parametric uncertainty estimates on the monthly-to-interannual time scales (ERSST.v4 estimates much larger than for HadSST3), the estimates are more

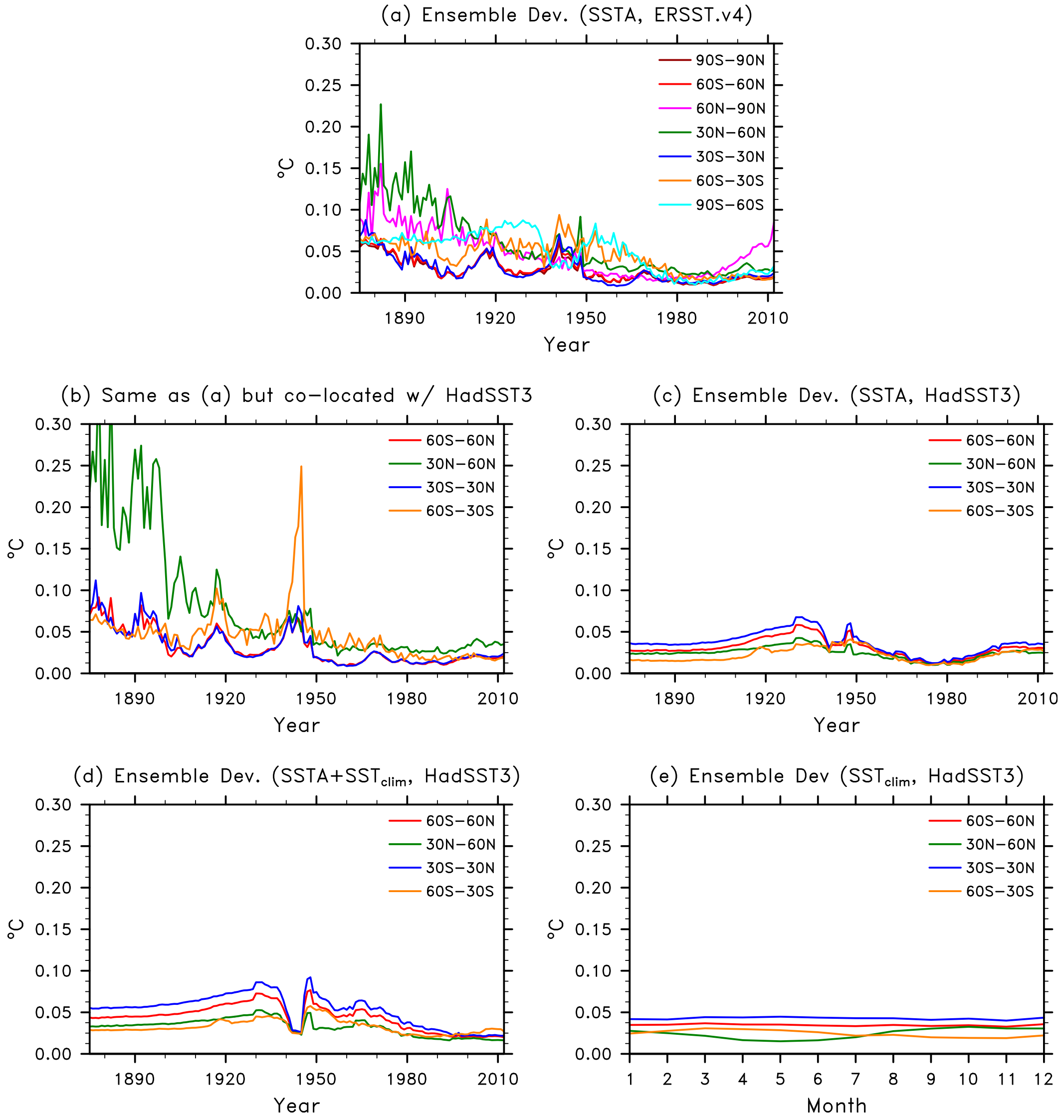


FIG. 9. Ensemble deviations σ_a, σ_c, and $\sigma_{(a+c)}$ (see text for definition) in ERSST.v4 and HadSST3 from a global- and regional-mean perspective. (a) Annual-mean σ_a, also $\sigma_{(a+c)}$ in ERSST.v4; (b) annual-mean σ_a in HadSST3; (c) annual-mean $\sigma_{(a+c)}$ in HadSST3; and (d) seasonal σ_c in HadSST3. In the collocated ERSST.v4, the grid sampling is the same between ERSST.v4 and HadSST3 such that the uncertainties are comparable between two datasets. Since there are substantial missing data over the polar region (outside 60°S–60°N), the comparison between two datasets is then confined within the region 60°S–60°N.

comparable for these long-term large-scale diagnostics, with HadSST3 now providing the broader range. The interquartile range in ERSST.v4 gives a warming from 0.718° to 0.747°C century^{-1} (0.029°C century^{-1} dispersion) while the range for HadSST3 gives a warming from 0.618° to 0.661°C century^{-1} (0.043°C century^{-1} dispersion). ERSST.v4 exhibits a degree of skewness in the ensemble estimates for this global trend diagnostic that is not apparent in HadSST3. This skewness is such that larger values for this trend diagnostic in ERSST.v4 are marginally more likely than smaller values.

To further test the possible extreme solutions of global warming that could be achieved in ERSST.v4 and whether we could better reconcile the two data products, we conduct two additional runs by deliberately perturbing those five SPPs with options that can generate more/less warming than the operation run (Fig. 5 and Table 2). Results show that the estimated warming extreme is 0.736° and 0.669°C century^{-1}. The former is well within the range of ensemble estimation, while the latter is slightly below the range of ensemble estimation. This reaffirms the presence of significant

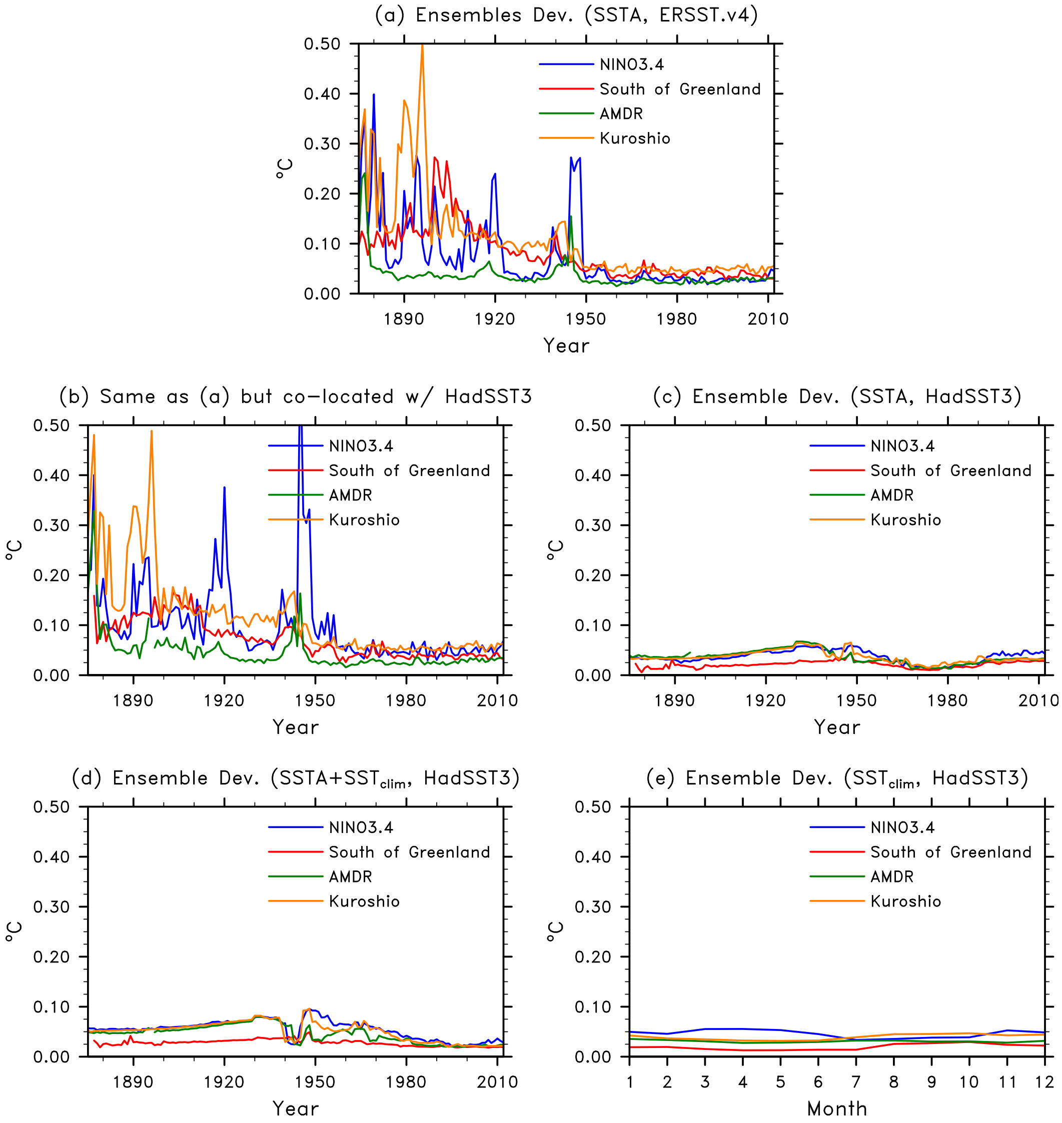


FIG. 10. As in Fig. 9, but for the four key regions (see Fig. 2).

nonlinearity (section 3b) but also provides some degree of confidence that the 100-member ensemble is likely a reasonable indicator of the true range of possible solutions (given, of course, the choice of parameters and possible settings).

The trends in both ensembles are significantly nonzero in the sense that the recognized parametric uncertainties can rule out the presence of a zero trend in either product. Although the two ranges marginally overlap, this does not necessarily mean they are consistent (Lanzante 2005).

This analysis of the two ensemble estimates has served to highlight how they are clearly distinct from one another. ERSST.v4 considers sources of uncertainties that project far more strongly onto higher-frequency components yielding broader uncertainties on high-frequency and local scales. In contrast, all of the parametric uncertainty components in HadSST3 relate to uncertainties in the bias adjustments that project mainly onto long-term changes on broader spatial scales. It is clear that neither estimate is ‘‘complete,’’ in the sense that the sources of uncertainty considered are not holistic in either. Therefore, these estimates need to be assessed further in regards to which is more appropriate for a given application and how they should be interpreted. As noted in Kennedy et al. (2011b) and Kennedy (2014), it is necessary to increase the number of independent estimates and parametric uncertainty estimates to more holistically understand historical SSTs and their uncertainties.

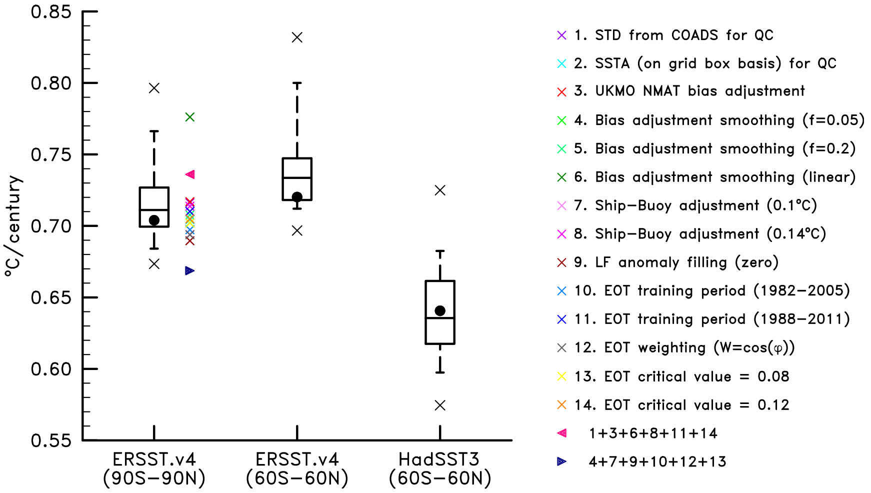


FIG. 11. Box-and-whisker plots of the trend of mean SST (1910–2012) in ensemble and operational runs from ERSST.v4 over the globe, the collocated ERSST.v4 with HadSST3 within 60°S–60°N, and HadSST3 within 60°S–60°N. The box shows the median, the lower quartile, and the upper quartile for the ensemble members. The black crosses indicate the lower and upper extreme of the ensemble members. The trend of operational run is denoted by a black dot. For ERSST.v4, trends of the global-mean SST from SPP runs (cross) and two additional runs (triangle, see text for the details of these two runs) are shown on the right on the box. All the results are calculated from monthly data.

5. Comparison to other SST estimates and inferences

There exist several additional SST data products for which an operational "best guess" product exists but which do not produce a parametric uncertainty estimate. Nevertheless these multiple estimates, under the assumption of reasonableness, allow an assessment of: 1) whether the parametric uncertainty is likely holistic; and 2) whether the various recognized dataset construction uncertainties impact first-order conclusions about the variations in SST. The various issues around structural uncertainty assessments are discussed in substantially greater depth in the recent review of uncertainties in in situ SST by Kennedy (2014) than is possible here. For the purposes of the present analysis, it is necessary to note that the various estimates used in this section arise from different versions of the raw data holdings (ICOADS and its numerous precursors) and undertake distinct methodological choices to QC, adjustment, and averaging.

In Fig. 12 the parametric uncertainty estimate from section 3 is compared to the structural uncertainty apparent from available estimates. Here, the structural uncertainty is estimated as the deviation of the difference between the six available estimates under the aforementioned assumption of reasonableness. As is evident in Fig. 12a, the structural uncertainty is generally larger than the ERSST.v4 parametric uncertainty estimates. The clear implication here and from section 5 is that the parametric uncertainty estimate from a single dataset is insufficient to fully sample the structural uncertainty inherent in the data (Thorne et al. 2011b). That other datasets fall outside the parametric uncertainty estimates of a single dataset is not unique to SST. For example, Mears et al. (2011) found that for their Microwave Sounding Unit dataset parametric uncertainty estimates, other independently derived datasets fell outside their estimates over 50% of the time.

Neither the parametric uncertainty estimates in ERSST.v4 (Fig. 12b) nor the structural uncertainty apparent from the range of datasets (Fig. 12c) call into fundamental question the finding that centennial timescale SSTs at the global-mean level have increased substantially. It is worth noting that the three most recently derived estimates [HadSST3, the Centennial Observation-Based Estimates of SST version 2 (COBE-SST2), and ERSST.v4], which use the latest or latest but one version of ICOADS and each apply bias adjustments since 1941, show somewhat distinct characteristics from the remainder (descriptions of COBE-SST2 can be found in Hirahara et al. 2014). As noted by Kennedy (2014), there is an acute need for new analyses or revisiting old analyses but using the most up-to-date data sources and knowledge and better quantifying uncertainties inherent to each product. This analysis and the companion piece (Part I) help address this need.

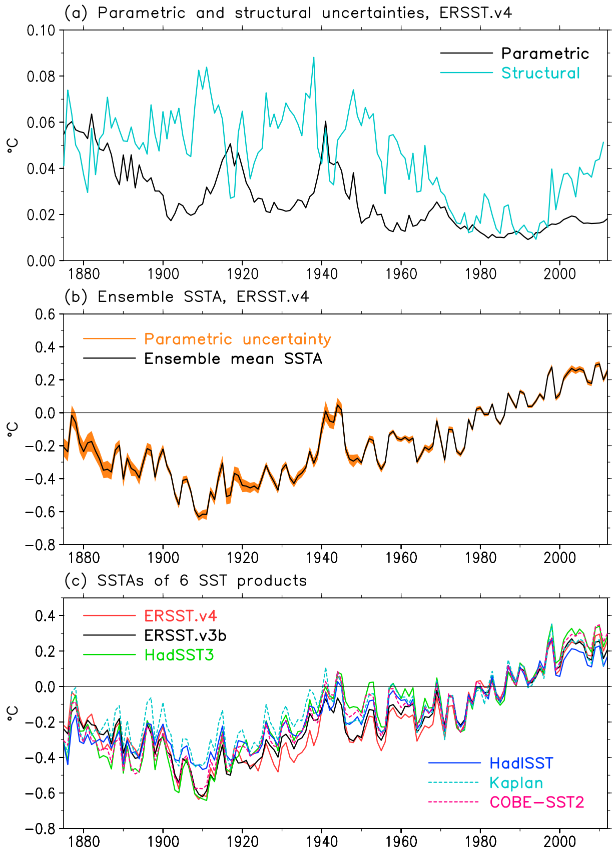


FIG. 12. (a) Parametric and structural uncertainties of ERSST.v4; (b) parametric uncertainty span (shading) and the ensemble-mean SSTA of ERSST.v4; and (c) SSTAs of six SST products, including ERSST.v4. All the results are globally and annually averaged. Structural uncertainty is defined as the deviation among six SSTAs under the assumption that each estimate is a random draw from the very large number of plausible approaches to SST dataset construction.

6. Discussion

A parametric uncertainty estimate for ERSST.v4 is provided in this study by performing sensitivity experiments. In the low-frequency (decadal- and century-scale), single-parameter experiments indicate that SST is sensitive to steps associated with quality control, bias adjustment, smoothing, and, in some regions, high-frequency spatial field characterization using EOTs. The EOT uncertainty dominates the high-frequency components everywhere, projects most strongly onto interseasonal-to-interannual variability, and is particularly acute in regions of high variability, such as the tropical Pacific Niño-3.4 region in data-sparse epochs.

Significant nonlinear effects were found to exist in most combinations of multiple parameters, thereby requiring a Monte Carlo ensemble approach to fully sample the parametric uncertainty. The ERSST.v4 ensemble analysis shows that this uncertainty is largest prior to 1900 over the globe and all regions associated with the sparsity of data and uncertainty in the data's adjustment, with a further spike occurring around WWII, except in the Southern Ocean. Distinct from the other regions, SST in the Southern Ocean (also SST in 60°–90°N after 2000) is subject to a large uncertainty throughout the entire record because of the low-frequency infilling method choices, given that there are never sufficient observations in this region. Additionally, parametric uncertainty is investigated in four key regions of likely interest to end users of the dataset. In the Niño-3.4 area, there is substantial uncertainty in early time series behavior and in some specific events in the early twentieth century that arises mainly from EOT-associated parameter choices. This uncertainty relates to the magnitude rather than timing of ENSO events.

Through a comparison of the ensembles, parametric uncertainties are found to differ significantly between ERSST.v4 and HadSST3. In contrast to ERSST.v4, parametric uncertainty in HadSST3 is much smaller on monthly-to-interannual time scales over the globe. Also, latitudinal patterns are reversed between the two datasets, with parametric uncertainties being maximal over the mid and high latitudes (outside 30°S–30°N) in ERSST.v4 but largely confined to the tropics (30°S–30°N) in HadSST3.

The global-mean long-term trend computed from ERSST.v4 is most sensitive to two parameters: the bias adjustment smoothing and the low-frequency anomaly filling. From 1910 to 2012, the warming trend increases by 0.072°C century^{-1}, when the bias adjustment smoothing with linear scheme is used but decreases by 0.014°C century^{-1} with zero low-frequency anomaly filling (both used in ERSST.v3b). Unlike for the high-frequency series behavior, the parametric uncertainty estimates in HadSST3 and ERSST.v4 are broadly comparable in magnitude for global-mean long-term trends.

Structural uncertainty has been assessed through a comparison to multiple available estimates. These estimates are of varying heritage and complexity and are derived from different versions of historical marine databases. As noted by Kennedy (2014), these issues complicate a clean analysis of structural uncertainty. This is clearly an evolving field where new analyses will help better inform this aspect of the uncertainty, and we would join Kennedy (2014) in advocating for such new analyses and reanalyses.

There are some recognized uncertainties that we have not covered in the present analysis. These mainly revolve around the issues that relate to the use of a finite set of EOTs to reconstruct spatial fields. At least three intertwined issues pertain here. Even with spatially complete perfect data as input, the EOT method would yield some degree of smoothing and information loss, as 130 EOTs will only capture some percentage of the full spatial information present. When input data are incomplete, this information loss becomes greater, and when they are imperfect, it becomes greater still. These sources of uncertainty have not been explicitly addressed here, but initial, ongoing analyses suggest that the effects will tend to be larger at smaller scales.

Finally, for users to utilize the ERSST.v4 parametric uncertainty estimates, the ensemble runs will be supplied alongside the ERSST.v4 product. Because the 100-member ensemble is identically formatted to the operational product, it should be simple for users to assess the uncertainty in a meaningful manner as it pertains to their region, time scale, and diagnostic of interest. However, unlike the operational dataset version described in Part I the ensemble will not be updated every month. We would strongly encourage users to make use of these ensembles to understand the impact of recognized, quantified uncertainties on their own analyses and applications of interest.

7. Conclusions

In conclusion, we have quantified herein the parametric uncertainty in the ERSST.v4 product, assessed the impact upon a range of space and time scales, and intercompared these estimates with the preexisting estimates from HadSST3 arising solely from SST bias adjustment parameters. The uncertainties in ERRST.v4 are largest in data-sparse periods and regions and have distinct impacts at different space and time scales. For long-term global-mean trends, the parametric uncertainties are an order of magnitude smaller than the estimated trend, as they are for HadSST3. Furthermore, the structural uncertainties were somewhat larger than our parametric uncertainty estimates, and, to the extent they can be ascertained from the handful of available centennial-time-scale SST products, were also substantially smaller than the long-term trend. Therefore, unless the various available means of assessing dataset construction uncertainties are underestimated by a large factor, it can be concluded that globally averaged SSTs have increased since the early twentieth century, with some uncertainty inherent in the exact magnitude. The best estimate for the magnitude of the global-mean SST warming since 1910, according to our ERSST.v4 product, is around 0.7°C, with asymmetry in the parametric uncertainty such that greater values are somewhat more likely than smaller values.

Acknowledgments. In the early phases of this work, PWT was supported by NOAA/NCDC whilst an employee of the Cooperative Institute for Climate and Satellites. We thank John Kennedy for providing some of the data sources for constructing Fig. 12 and useful discussions around the value of the various products used therein. Comments from two anonymous NCDC internal reviewers helped to improve the manuscript. We thank three anonymous reviewers who provided substantive reviews that considerably improved the manuscript.

REFERENCES

Binder, K., and D. W. Heerman, 1992: *Monte Carlo Simulation in Statistical Physics: An Introduction.* Springer-Verlag, 129 pp.

Blunden, J., and D. S. Arndt, 2013: State of the climate in 2012. *Bull. Amer. Meteor. Soc.,* **94,** S1–S258, doi:10.1175/2013BAMSStateoftheClimate.1.

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

Gregory, J. M., 2000: Vertical heat transports in the ocean and their effect on time-dependent climate change. *Climate Dyn.,* **16,** 501–515, doi:10.1007/s003820000059.

Grumbine, R. W., 1996: Automated passive microwave sea ice concentration analysis at NCEP. NOAA Tech. Note 120, 13 pp. [Available online at http://polar.ncep.noaa.gov/mmab/papers/tn120/ssmi120.pdf.]

Hartmann, D. L., and Coauthors, 2013: Observations: Atmosphere and surface. *Climate Change 2013: The Physical Science Basis,* T. F. Stocker et al., Eds., Cambridge University Press, 159–254. [Available online at http://www.climatechange2013.org/images/report/WG1AR5_Chapter02_FINAL.pdf.]

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., P. H. Stone, A. P. Sokolov, and I. V. Kamenkovich, 2003: The deep-ocean heat uptake in transient climate change. *J. Climate,* **16,** 1352–1363, doi:10.1175/1520-0442-16.9.1352.

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

Ishii, M., A. Shouji, S. Sugimoto, and T. Matsumoto, 2005: Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using ICOADS and the Kobe Collection. *Int. J. Climatol.,* **25,** 865–879, doi:10.1002/joc.1169.

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.

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 errors. *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., N. A. Rayner, D. I. Berry, M. Saunby, B. I. Moat, J. J. Kennedy, and D. E. Parker, 2013: Global analysis of night marine air temperature and its uncertainty since 1880: The HadNMAT2 data set. *J. Geophys. Res. Atmos.,* **118,** 1281–1298, doi:10.1002/jgrd.50152.

Lanzante, J. R., 2005: A cautionary note on the use of error bars. *J. Climate,* **18,** 3699–3703, doi:10.1175/JCLI3499.1.

Mears, C. A., F. J. Wentz, P. Thorne, and D. Bernie, 2011: Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo estimation technique. *J. Geophys. Res.,* **116,** D08112, doi:10.1029/2010JD014954.

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.

Parker, D. E., C. K. Folland, and M. Jackson, 1995: Marine surface temperature: Observed variations and data requirements. *Climatic Change,* **31,** 559–600, doi:10.1007/BF01095162.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.,* **108,** 4407, doi:10.1029/2002JD002670.

——, P. Brohan, D. E. Parker, C. K. Folland, J. J. Kennedy, M. Vanicek, T. J. Ansell, and S. F. B. Tett, 2006: Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset. *J. Climate,* **19,** 446–469, doi:10.1175/JCLI3637.1.

Reynolds, R. W., 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.

Smith, T. M., and R. W. Reynolds, 2002: Bias corrections for historical sea surface temperatures based on marine air temperatures. *J. Climate,* **15,** 73–87, doi:10.1175/1520-0442(2002)015<0073:BCFHSS>2.0.CO;2.

——, and ——, 2003: Extended reconstruction of global sea surface temperatures 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.

——, ——, R. E. Livezey, and D. C. Stokes, 1996: Reconstruction of historical sea surface temperature 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.

Thorne, P. W., and Coauthors, 2011a: A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes. *J. Geophys. Res.,* **116,** D12116, doi:10.1029/2010JD015487.

——, J. R. Lanzante, T. C. Peterson, D. J. Seidel, and K. P. Shine, 2011b: Tropospheric temperature trends: History of an ongoing controversy. *Wiley Interdiscip. Rev.: Climate Change,* **2,** 66–88, doi:10.1002/wcc.80.

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

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

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.

Xue, Y., T. M. Smith, and R. W. Reynolds, 2003: Interdecadal changes of 30-yr SST normals during 1871–2000. *J. Climate,* **16,** 1601–1612, doi:10.1175/1520-0442-16.10.1601.