Evaluating SST Analyses with Independent Ocean Profile Observations

Boyin Huang, William Angel, and Tim Boyer

NOAA/National Centers for Environmental Information, Asheville, North Carolina

Lijing Cheng

International Center for Climate and Environment Sciences, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

Gennady Chepurin

Department of Atmospheric and Oceanic Science, University of Maryland, College Park, College Park, Maryland, and NOAA/National Centers for Environmental Information, Asheville, North Carolina

Eric Freeman and Chunying Liu

ERT, Inc., Laurel, Maryland, and NOAA/National Centers for Environmental Information, Asheville, North Carolina

Huai-Min Zhang

NOAA/National Centers for Environmental Information, Asheville, North Carolina

ABSTRACT

The difficulty in effectively evaluating sea surface temperature (SST) analyses is finding independent observations, since most available observations have been used in the SST analyses. In this study, the ocean profile measurements [from reverse thermometer, CTD, mechanical bathythermograph (MBT), and XBT] above 5-m depth over 1950–2016 from the World Ocean Database (WOD) are used (data labeled pSSTW). The biases of MBT and XBT are corrected based on currently available algorithms. The bias-corrected pSSTW over 1950–2016 and satellite-based SST from the European Space Agency (ESA) Climate Change Initiative (CCI) over 1992–2010 are used to evaluate commonly available SST analyses. These SST analyses are the Extended Reconstructed SST (ERSST), versions 5, 4, and 3b, the Met Office Hadley Centre Sea Ice and SST dataset (HadISST), and the Japan Meteorological Administration (JMA) Centennial In Situ Observation-Based Estimates of SST version 2.9.2 (COBE-SST2). Our comparisons show that the SST from COBE-SST2 is the closest to pSSTW and CCI in most of the Pacific, Atlantic, and Southern Oceans, which may result from its unique bias correction to ship observations. The SST from ERSST version 5 is more consistent with pSSTW than its previous versions over 1950–2016, and is more consistent with CCI than its previous versions over 1992–2010. The better performance of ERSST version 5 over its previous versions is attributed to its improved bias correction applied to ship observations with a baseline of buoy observations, and is seen in most of the Pacific and Atlantic.

1. Introduction

Sea surface temperature (SST) is a key parameter in the ocean–atmosphere system, and is widely used in atmospheric simulations and weather forecasting. Studies showed that there is a threshold of SST in triggering tropical convection, cyclone genesis, and hurricane development (e.g., Fu et al. 1990; Dutton et al. 2000; McTaggart-Cowan et al. 2006; Johnson and Xie 2010; Dare and McBride 2011). These studies indicated that the accuracy of SST analyses may critically affect climate simulation and weather forecasting. The in situ observations from ships and buoys are typically used in

---

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

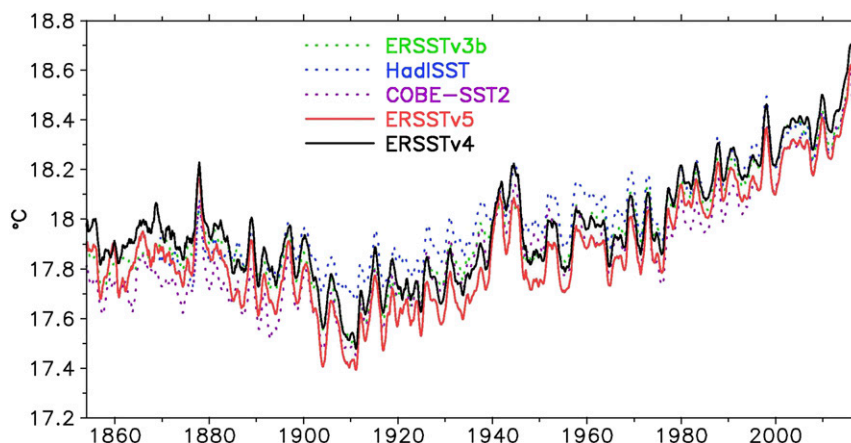*Corresponding author*: Boyin Huang, boyin.huang@noaa.gov

FIG. 1. Globally averaged SST from ERSST.v5, ERSST.v4, ERSST.v3b, HadISST, and COBE-SST2. The difference between ERSST versions 5 and 4 is approximately −0.1°C. A 12-month (6 month at starting and ending points) running average has been applied for plotting.

SST analyses (e.g., Kennedy et al. 2011a,b; Hirahara et al. 2014; Huang et al. 2015, 2017). Most recently, the observations from Argo floats are also added to an SST analysis (Huang et al. 2017). Satellite observations derived from the Advanced Very High Resolution Radiometer (AVHRR) and other instruments are included in some blended SST analyses in the satellite era (Rayner et al. 2003; Reynolds et al. 2002, 2007).

The accuracy of SST analyses is mostly dependent on how the biases of observations from different instruments are corrected (Kent et al. 2017). The study of Huang et al. (2017) suggested that the globally averaged SST in the Extended Reconstructed SST, version 5 (ERSST.v5), is about 0.1°C systematically lower than the previous version, ERSST.v4 (Fig. 1, solid black and red lines). The lower SST in ERSST.v5 results from the biases of ship observations adjusted to more accurate or homogeneous buoy observations. However, the SSTs in ERSST.v5 and earlier versions have not fully been evaluated by independent observations, particularly before the 1990s. In this study, near-surface (0–5-m depth) temperatures derived from independent ocean profile measurements from 1950 to 2016 are used to evaluate the commonly available SST analyses including ERSST.v5 and ERSST.v4. These profile measurements are from reversing thermometers (RT; attached to Nansen–Niskin hydrographic bottles), conductivity–temperature–depth (CTD), mechanical bathythermographs (MBT), and expendable bathythermographs (XBT). RT and CTD data (RT/CTD hereinafter) are considered highly accurate (currently to 0.002°C) and with no known systematic bias. CTD data are routinely calibrated to concurrent reversing thermometer measurements.

The paper is organized as follows: SST datasets from available SST analyses and evaluation datasets from ocean profile measurements and satellite-based observations are described in section 2. The evaluation datasets are compared with in situ observations from ships, buoys, and Argo floats to ensure their quality in section 3. The SST analyses are evaluated against those ocean profile measurements and satellite-based measurements in section 4. The study is summarized and discussed in section 5.

2. Data

a. SST analyses

The following SST analyses will be assessed in section 4: ERSST versions 5, 4, and 3b (Huang et al. 2017; Huang et al. 2015; Smith et al. 2008), the Met Office Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST; Rayner et al. 2003), and the Japan Meteorological Agency (JMA) Centennial In Situ Observation-Based Estimates of SST, version 2.9.2 (COBE-SST2; Hirahara et al. 2014). The globally averaged SSTs from these analyses (Fig. 1) are generally consistent, but differences are seen in some periods resulting from unique bias-correction schemes (Huang et al. 2015).

The ERSST.v5 is based upon ERSST.v4, with improved bias correction to ship observations using more homogeneous buoy observations as a baseline. ERSST.v5 has a better representation of spatial and temporal variations in high-latitude oceans and El Niño and La Niña activities in the tropical Pacific (Huang et al. 2017). ERSST.v4 is based upon ERSST.v3b by correcting the biases of ship observations after the 1940s, which is neglected in ERSST.v3b, and therefore ERSST.v4 has a better representation of global SST warming trends in the most recent decades (Karl et al. 2015; Hausfather et al.

2017). The nighttime marine air temperature (NMAT; Kent et al. 2013) is used in correcting the biases of ship observations in all ERSST versions. The bias-corrected in situ SSTs are reconstructed using localized empirical orthogonal teleconnection (EOT) modes on a monthly and global $2° \times 2°$ grid from 1854 to the present. ERSST.v5, ERSST.v4, and ERSST.v3b use in situ ship and buoy observations; Argo observations above 5-m depth are added in ERSST.v5. The reason for including Argo observations in ERSST.v5 is that the number of observations from Argo floats has rapidly expanded over the global oceans since 2000. By including Argo observations, the coverage of in situ data increases by approximately 10% since 2006 (Huang et al. 2017), and therefore the operational SST analysis becomes more reliable at regional scale while the globally averaged SST remains close.

The HadISST includes ship and buoy observations as well as satellite AVHRR observations. The biases of ship observations are corrected before 1941 based upon a bucket model of Folland and Parker (1995), and no bias correction is applied after 1941 (Kent et al. 2013). One of the difficulties in applying the bucket model is that metadata of bucket types are not always available in the historical record, although new methods to improve metadata have recently been developed (Carella et al. 2018). Therefore, the bucket type is randomly selected in case of an unknown type of buckets in HadISST. The resolution of HadISST is monthly on a global $1° \times 1°$ grid from 1870 to 2016. This bucket model was also implemented in the Met Office Hadley Centre SST, version 3 (HadSST3; Kennedy et al. 2011a,b), which includes in situ observations only. The advantage of HadISST over HadSST3 is that HadISST is reconstructed with gaps filled over the global oceans using empirical orthogonal function (EOF) modes. In contrast, HadSST3 is reconstructed on the monthly $5° \times 5°$ grids where in situ observations are available, and no reconstruction/interpolation is made on the grids without in situ observations. Therefore, HadISST is selected and averaged to the same grids ($2° \times 2°$) of ERSST for comparison purposes.

The COBE-SST2 includes in situ ship and buoy observations only. The same bucket model of Folland and Parker (1995) is used to calculate the biases of ship observations before 1941. After 1941, the mean biases for insulated buckets, uninsulated buckets, and engine room intake are first calculated. The ship biases are then derived by weighting the mean biases of insulated buckets, uninsulated buckets, and engine room intake. The weighting is determined by assuming that the anomaly of the unknown type of observations is equivalent to the anomaly of other known types of observations

within a $5° \times 5°$ grid box (Hirahara et al. 2014). The in situ observations are finally reconstructed using EOF modes on a monthly global $1° \times 1°$ grid from 1850 to 2016. The COBE-SST2 is box-averaged to $2° \times 2°$ grids for this comparison study.

b. pSSTW from the World Ocean Database

The SST data from profile measurements by RT/CTD, MBT, and XBT at a depth of 0–5 m over 1920–2016 are from the World Ocean Database (WOD) (Boyer et al. 2016), and the data are labeled as pSSTW hereinafter. Near-surface transients in XBT profiles are accounted for in three ways: reporting, quality, and bias correction. Often the top 2.7 m of an XBT trace are not reported, or reported as missing values in order to avoid the noise of near-surface transients. XBT data are assigned quality flags from data originators. There are two major categories of bias corrections: 1) for depth and 2) for temperature measurements. The details of the bias corrections are described in section 3a.

A pair of pSSTW are downloaded: one without bias correction and the other with bias correction. The pSSTW data undergo a quality-control procedure used in ERSST.v5 that filters out outliers departing four standard deviations or more from the ERSST.v5 analysis (Huang et al. 2017). The quality-controlled SST data are bin-averaged to $2° \times 2°$ grid boxes at monthly time scale from 1920 to 2016, and globally integrated numbers and area coverages of observations are calculated. The area coverage is a ratio of the total gridbox area containing observations over the total ocean area. Calculations show that numbers of observations are solely from RT/CTD over 1920–40, dominantly from MBT and RT/CTD over 1940–70, and mostly from XBT and RT/CTD over 1970–2010 (Fig. 2a). Coverages (as high as 25%) are solely from RT/CTD over 1920–40, mostly from MBT and RT/CTD over 1940–70, and dominantly from XBT over 1970–2010 (Fig. 2b).

The recorded depths of these pSSTW are usually shallower than 0.5 m before the 1980s, and increase to 1–3 m over 1990–2010 (Fig. 2c). These pSSTW data within the depth of 0–5 m are considered as SSTs; this is to be consistent with the fact that the Argo measurements of 0–5-m depth were treated as SSTs (Roemmich et al. 2015; Huang et al. 2017). It should be noted that RT/CTD depths before the 1990s may have been recorded as 0 m when actual depths are deeper than 0.5 m because of the logistics of obtaining a near-surface measurement from the profiling instruments (JPOTS 1991, 25–26). Furthermore, the first recorded depth of XBT is approximately 0.7 m, taken 0.1 s after the probe has hit the water surface. The depths of pSSTW observations differ from those of drifting buoys (0.2–0.3 m), moored buoys
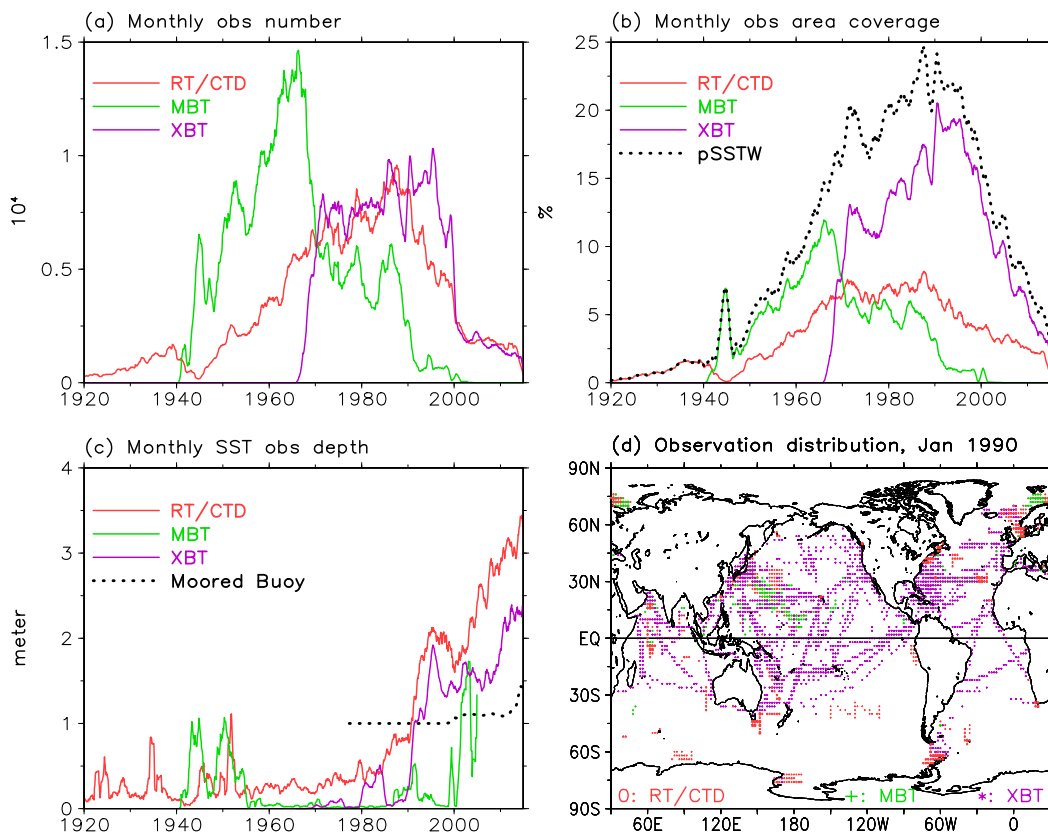
FIG. 2. (a) Number, (b) area coverage, and (c) depth of monthly observations from RT/CTD, MBT, and XBT over the global oceans and (d) observation distribution of RT/CTD, MBT, and XBT in January 1990. The merged pSSTW coverage in (b) is overlaid with a black dotted line. The depth of moored buoy observations (black dotted line) in (c) is from the Tropical Atmosphere Ocean (TAO), Triangle Trans-Ocean Buoy Network (TRITON) in the tropical Indian Ocean, and the Prediction and Research Moored Array in the Tropical Atlantic (PIRATA).

(1–1.5 m; Fig. 2c), and ships (1–10 m) (Reynolds et al. 2002; Lumpkin and Pazos 2007; Barale et al. 2010; Matthews 2013). The difference in observing depths may impact the observed SSTs. For example, the recorded depth of RT/CTD measurements increases from 1 m in the late 1980s to 3 m in the 2010s (Fig. 3, red line). The observed temperatures from RT/CTD are lower than buoy observations at a nominal depth of 0.2 m by 0.05°– 0.2°C over 1990–2010 (Fig. 3, black line).

These pSSTW data from RT/CTD, MBT, and XBT profiles are merged onto 2° × 2° grid boxes by weighting their number of observations. The SSTs derived from other types of profiles are not included because of their low observation numbers, low area coverages, or short time periods. One advantage of this pSSTW collection is its relatively long period of data, from the 1920s to the 2010s, while its area coverage is in a range of 5%–25% from the 1950s to the 2010s. The other advantage is that it is independent from currently available SST analyses described in section 2a, and can therefore be used to assess these SST analyses. To our knowledge, these

pSSTW data are rarely used in evaluating these SST analyses. The profile measurements before 1950 are very sparse (Figs. 2a,b) and therefore not used in validating SST analyses in section 4. Since Argo observations have been included in ERSST.v5 and are no longer independent of ERSST.v5, they are not used as evaluation data.

c. In situ SST from ships, buoys, and Argo floats

Since in situ SSTs are independent from pSSTW, they are first used to assess pSSTW to ensure the quality of pSSTW as an evaluation dataset. The qualified pSSTW is then used to assess SST analyses in section 4. For this purpose, the in situ SSTs are merged from observations from ships, buoys, and Argo floats (hereinafter ShipBuoyArgo). The ShipBuoyArgo is averaged on monthly 2° × 2° grid boxes over the global oceans from 1854 to 2016 (Huang et al. 2017). The ship and buoy SSTs are from the International Comprehensive Ocean–Atmosphere Data Set (ICOADS) release 3.0 (Freeman et al. 2017), and Argo temperatures are from the Argo Global Data Assembly Centre (Argo 2000). The ShipBuoyArgo
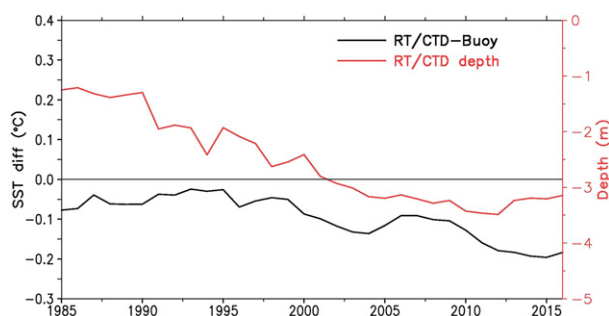
FIG. 3. Annually averaged SST difference (left axis, solid black) between RT/CTD and baseline buoy observations with 5-yr running filter when plotting, and annually averaged RT/CTD observing depth collocated with baseline buoy observations (right axis, solid red). The correlation coefficients between the SST difference and depth are 0.72.

SST is derived as follows (Huang et al. 2017): First, the SSTs from ship and Argo float observations are corrected using buoy observations over 1990–2010 as a baseline; the corrections account for the differences caused by instrument and depth changes. Second, the corrected (buoy baselined) ship and Argo SSTs are merged with buoy SSTs by weighted averaging. The weights for ship, buoy, and Argo SSTs are 1, 6.8, and 6.8 times the number of their observations in every $2° \times 2°$ grid box, respectively. The weights are determined according to the ratio of random-error variances of ship and buoy observations (Reynolds and Smith 1994), and it is assumed that the random-error variance of Argo observations is equivalent to that of buoy observations (Huang et al. 2017). The ShipBuoyArgo SST represents an estimation of available in situ observations in the global oceans and is used to reconstruct temporal and spatial variations of SST over the global oceans in ERSST (e.g., Huang et al. 2015, 2017).

In addition to the merged ShipBuoyArgo SST, the baseline buoy observations are also used in comparing pSSTW and the satellite-based SST from the European Space Agency (ESA) Climate Change Initiative (CCI) because of their high-quality and relatively uniform distribution over the global oceans. The area coverage of buoy observations increases from approximately 5% in the late 1980s to approximately 35% in the late 2000s (Huang et al. 2017).

d. CCI SST

The SSTs from the ESA CCI, level 4 version 1.1 (Merchant et al. 2014), are used first to compare pSSTW (section 3b) and then to evaluate selected SST analyses (section 4). The original resolution of CCI SSTs is daily $0.05° \times 0.05°$, which is processed into monthly $2° \times 2°$ resolution for comparison purpose. The CCI SSTs

represent a near-surface temperature at depths of 0.2 m where most drifting buoys also measure SSTs. The CCI SSTs at depths of 0.2 m are derived by adding an adjustment to the skin temperature. The adjustment is estimated according to ocean skin effect and near-surface thermal stratification (Donlon et al. 1999).

As described in Merchant et al. (2014), the CCI SSTs were derived from both AVHRR and Along-Track Scanning Radiometer (ATSR) observations using a reduced-state-vector optimal estimation algorithm. CCI SSTs are almost independent from in situ observations from ships, buoys, and Argo floats, since no in situ observations are used in estimation of CCI SSTs but ocean reanalysis data are used as a background field in the absence of observations. Earlier comparisons (e.g., Huang et al. 2017) showed that CCI SSTs are very consistent with buoy observations, suggesting that the quality of CCI SSTs is comparable with buoy SSTs. Therefore, CCI SSTs are used to validate SST analyses based on in situ observations in section 4.

Likewise, CCI SSTs are independent from pSSTW. Therefore, CCI SSTs are used to validate the quality of pSSTW in section 3b. The validation of pSSTW using CCI SSTs is to ensure a qualitative assessment of SST analyses in section 4. The advantage of CCI SSTs is their high spatial coverage of near 100% over the global oceans, which can provide an extra validation in areas where no pSSTW observations are available. However, the disadvantage is their short time period—from September 1991 to December 2010.

3. Biases in XBT and MBT observations and intercomparisons

a. Bias correction to XBT and MBT

Biases are found in XBT observations resulting from a depth offset and temperature observations. The study of Levitus et al. (2009, hereinafter L09) provided an adjustment for temperature observations, and the study of Cheng et al. (2014, hereinafter C14) provided adjustments for both the depth offset and temperature observations. The surface correction for most years is larger than the correction at the next nearest depth (10 m) despite the expectation of a positive relationship between the temperature correction and increased depth. The larger surface correction results from near-surface transients as described in section 2b.

The studies of L09 and C14 indicated that the magnitude of the warm bias of XBT observations is approximately 0.1°–0.2°C over 1960–90 and 0.05°C over 2000–10 (Fig. 4, solid black and red lines), which is consistent with other studies (e.g., Gouretski and Koltermann 2007; Reverdin et al. 2009; DiNezio and
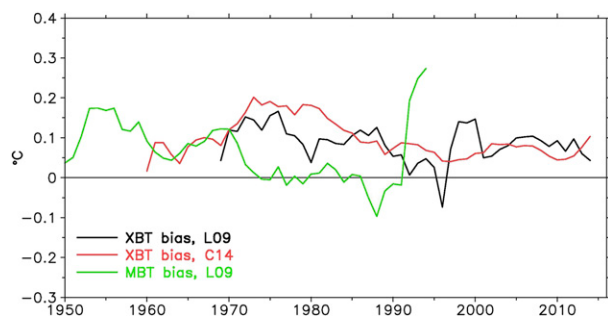
FIG. 4. Estimated annually averaged median biases of near-surface (0–5 m) XBT (solid black) and MBT (solid green) from L09 and XBT from C14 (solid red). The estimated biases of XBT and MBT in L09 are derived relative to RT/CTD and Argo measurements. The estimated bias of XBT in C14 is derived as the difference between XBT datasets without and with the correction of C14 in the WOD.

Goni 2010; Gouretski and Reseghetti 2010; Cowley et al. 2013; Goes et al. 2015; C14; Cheng et al. 2016). A warm bias (approximately 0.1°C) of MBT is identified by L09 over 1950–75 (Fig. 4, solid green line), while the higher biases in the early 1990s are associated with sparse MBT and RT/CTD–Argo comparison pairs. These biases largely result from inaccurate depths of temperature measurements (i.e., depth error) and inaccurate temperature measurements themselves (i.e., temperature error) (Cheng et al. 2016). The XBT depths are calculated according to a fall-rate equation and the time lapse since the instrument hits the ocean surface (Reverdin et al. 2009; Ishii and Kimoto 2009; DiNezio and Goni 2010). The inaccuracy of the fall-rate equation is mainly responsible for the depth error, and the temperature error is due to a combination of errors in temperature thermistor, recording system, and so on.

The difference of XBT bias corrections between L09 and C14 is notable. The XBT biases fluctuate in L09 before 2000, and are relatively smooth in C14, which may result from their methods in calculating the biases. The biases of XBT (or MBT) observations in L09 were calculated with respect to RT/CTD and Argo observations as follows: First, averaged SST anomalies with respect to the same monthly climatology were calculated in RT/CTD and Argo, XBT, and MBT each year within a 5-yr data window over the global oceans on $4° \times 2°$ grids, and then the median differences between RT/CTD and XBT (or MBT) over the global oceans were defined as the biases. In contrast, the XBT biases in C14 were calculated using RT/CTD–Argo and XBT pairs on monthly $1° \times 1°$ grid. By using the pairs, C14 modeled and corrected the depth and temperature errors separately, which vary with time, ocean temperature, and probe types (C14). The data from C14 in Fig. 4 are retrieved in this study by the

annually and globally averaged difference between XBT SSTs without and with the bias correction of C14, and are available from the WOD.

b. Bias-corrected XBT and MBT versus ShipBuoyArgo and satellite SST

The quality of bias-corrected pSSTW is assessed by comparing it with independent ShipBuoyArgo observations. To make a fair comparison, the collocated SST differences between the two datasets over the global oceans are calculated on monthly $2° \times 2°$ grids where both datasets have valid observations. Comparisons show that the differences between bias-corrected pSSTW and ShipBuoyArgo are approximately 0.1°C over 1950–85, decreasing to near zero over 1990–2010, increasing to 0.1°C by 2016 (Fig. 5, solid red line). These differences are similar to an earlier estimation between surface and near-surface observations before 2010 (Gouretski et al. 2012, their Fig. 1c).

The reduced difference between 1985 and 1995 may be associated with a more accurate bias correction applied to ship observations by using buoy observations as a baseline after 1985 in ERSST.v5 (Huang et al. 2017). Therefore, the merged ShipBuoyArgo becomes closer to pSSTW after 1985. The high difference in the 2010s may be associated with the overall reduced number and area coverage of pSSTW after 2010 (Figs. 2a,b), particularly in the central Pacific (not shown in figure).

Our test shows that the high difference between 2010 and 2016 reduces to near zero (not shown in figure) when Argo observations are removed from ShipBuoyArgo and added to pSSTW. The reasons for the near-zero difference are that Argo observations are very close to corrected ship and buoy observations, and that the area coverage of pSSTW increases by 10%–35% while area coverage of ShipBuoyArgo decreases by 10% (Huang et al. 2017). Our test implies that the difference between ShipBuoyArgo and pSSTW would be small if the area coverage of pSSTW before the 1990s was high, since the instruments in Argo floats and CTDs are made with the same specifications.

The pSSTW is also compared against independent CCI SST over 1992–2010 and independent baseline buoy SST over 1985–2016. Comparisons show that the difference between pSSTW and CCI (Fig. 5, purple line) are very close to that between pSSTW and baseline buoy observations (Fig. 5, green line). It should be noted that the disagreement between pSSTW and baseline buoy observations fluctuates near 1995, which may be largely associated with changes in the area coverage of combined pSSTW and baseline buoy observations and their geographic locations. Calculations indicate (not shown in figure) that the area coverage decreases in the central Pacific from 1992 to 1995, and increases in the western
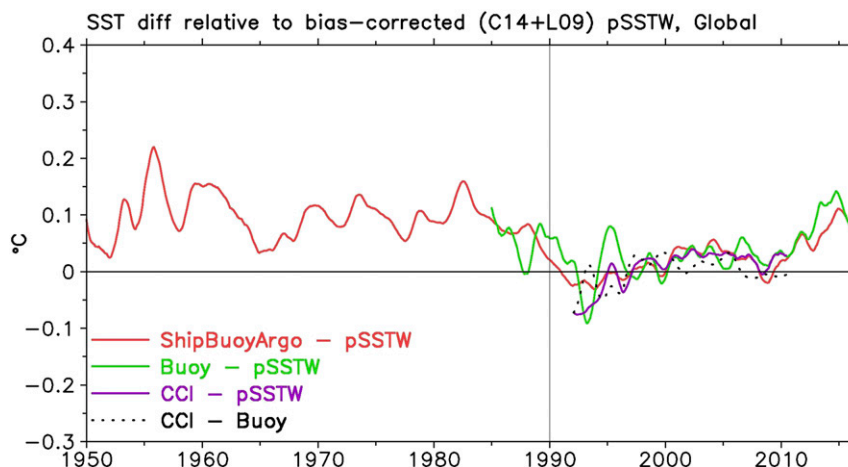
FIG. 5. Globally averaged SST differences of ShipBuoyArgo (solid red), baseline buoy observations (solid green), and CCI (solid purple). The differences are calculated relative to pSSTW with bias correction of C14 for XBT and L09 for MBT. The difference between CCI and baseline buoy observations (dotted black) is overlaid. A 12-month running average has been applied for plotting.

Atlantic near 1995. However, the disagreement between CCI and pSSTW between 1992 and 1995 is clearly not associated with the high area coverage of pSSTW during that period of time (Fig. 2b). Comparisons between CCI and baseline buoy observations indicate that CCI is very consistent with baseline buoy observations between 1995 and 2010, but exhibits a cold bias in the early period of 1992–95 (Fig. 5, black dotted line). This cold bias explains why all five SST analyses are warmer than CCI during this period of time, as shown by Huang et al. (2017, their Fig. 14, and later in Fig. 9a).

These comparisons indicate that the bias-corrected pSSTW is slightly colder than ShipBuoyArgo, particularly over 1950–90. The averaged (1950–90) difference is $0.093° \pm 0.012°C$. As analyzed earlier in section 2b (Fig. 3), the colder pSSTW may be associated with the fact that the XBT and MBT are corrected according to RT/CTD and Argo, and RT/CTD and Argo observations are colder than buoy observations near the ocean surface because the depths of RT/CTD and Argo observations are deeper than those of the buoy observations. For the same reason, the SST derived from Argo near-surface (0–5 m) measurements is approximately 0.03°C colder than buoy measurements, as reported earlier, and the offset was corrected when Argo data were infused into the ERSST.v5 (Huang et al. 2017). The colder pSSTW in comparison with buoy measurements may also be potentially associated with the diurnal effect near the ocean surface (Gille 2009), since the baseline buoy SST near the surface may be heated more than RT/CTD and Argo measurements at deeper depth during the daytime. However, our test using data during nighttime (between times

of 1 h after sunset and 1 h after sunrise) indicates that these differences remain, suggesting that the diurnal effect may be overwhelmed by the difference in observing depths.

If the large difference between pSSTW and ShipBuoyArgo in Fig. 5 is solely associated with deeper observing depths of RT/CTD and Argo, the question is why the difference between pSSTW and ShipBuoyArgo is larger before the 1990s. The RT/CTD observations deviated from any surface observations should be smaller before the 1990s, since RT/CTD observing depths are reported shallower (Fig. 2c). There are potentially three reasons to explain the relatively large difference between pSSTW and ShipBuoyArgo:

1) The actual RT/CTD depths before the 1990s are deeper than those shown in Fig. 2c, since the RT/CTD depths may had been set to zero when it was near the surface (JPOTS 1991, 25–26).

2) The biases of XBT and MBT may not be completely corrected before the 1990s. The biases of XBT and MBT in L09 and C14 are based on the median difference, not the averaged difference between XBT (MBT) and RT/CTD–Argo, over the global domain. Therefore, the bias correction based on the median difference may partially retain the biases when bias-corrected XBT and MBT are merged with RT/CTD and averaged over the global oceans.

3) The ShipBuoyArgo SSTs are warm-biased. The ShipBuoyArgo SSTs before the 1990s are mostly based on ship observations at a depth of 1–10 m. The biases in ship observations are corrected based
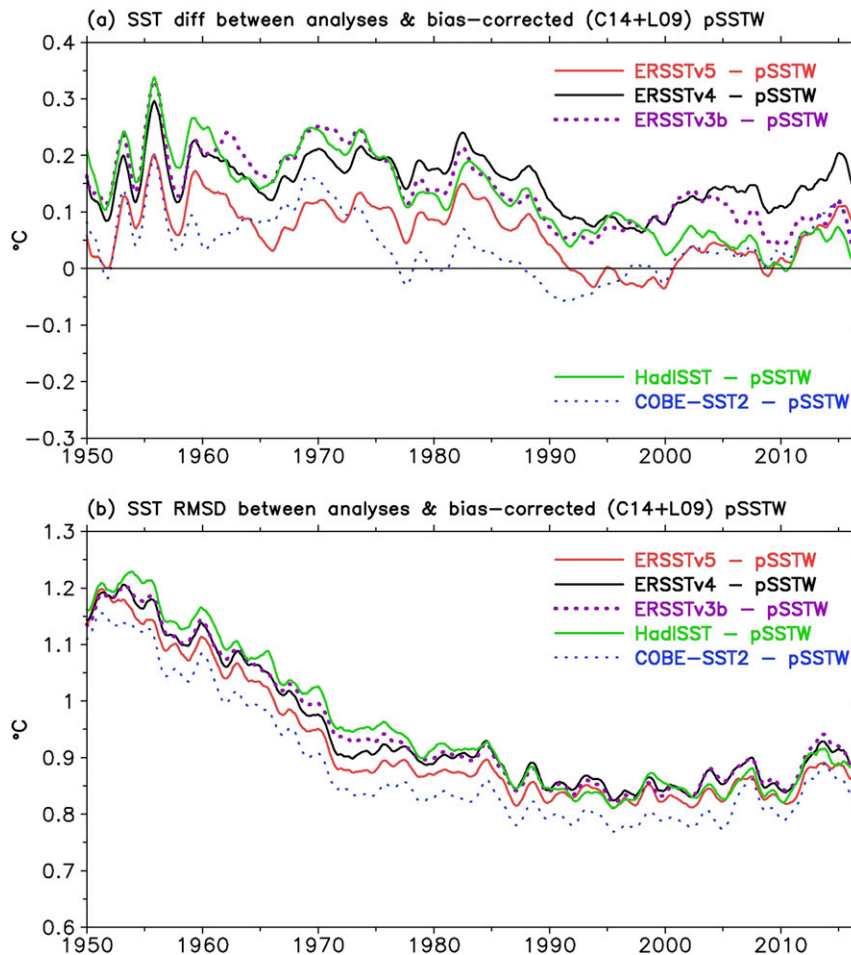
FIG. 6. (a) Globally averaged SST differences between analyses and pSSTW with bias cor-
rection of C14 for XBT and L09 for MBT, and (b) as in (a), but for RMSD. A 12-month running
average has been applied for plotting.

on NMAT and readjusted by buoy SSTs over the
baseline period of 1990–2010 (Huang et al. 2015,
2017), which may have uncertainties (Kent et al.
2017). It is interesting to note that the differences
between bias-uncorrected pSSTW and bias-corrected
ShipBuoyArgo are actually very small before the
1990s, which suggests that both of them could be
biased warm or biased cold at the same time.

4. Evaluation of SST analyses

a. Comparison with bias-corrected pSSTW

By removing the biases of XBT using the C14 method
and of MBT using the L09 method, the SST analyses are
evaluated against bias-corrected pSSTW. The collo-
cated differences between SST analyses and pSSTW are
calculated over the global oceans on $2° \times 2°$ grids where

both analyses and pSSTW have valid data (Fig. 6a).
Figure 6a shows that the globally averaged differences
are overall positive over 1950–2016 with substantial in-
terannual variations, indicating that pSSTW is generally
colder, as shown in Fig. 5. The colder pSSTW may
result from a deeper observing depth, as discussed in
section 3b. The differences are smaller in ERSST.v5
than ERSST.v4, ERSST.v3b, and HadISST, and are the
smallest in COBE-SST2. The averaged differences over
1950–2010 are $0.064° \pm 0.014°$, $0.152° \pm 0.012°$, $0.148° \pm
0.016°$, $0.137° \pm 0.019°$, and $0.002° \pm 0.014°$C (Table 1,
second row) in ERSST.v5, ERSST.v4, ERSST.v3b,
HadISST, and COBE-SST2, respectively. The small
deviation in COBE-SST2 may be attributed to its unique
SST bias correction for unknown-type ship observations
described in section 2a.

To account for random noise in pSSTW and SST an-
alyses, root-mean-square differences (RMSDs) between

TABLE 1. Averaged difference and its uncertainty at 95% confidence level (°C) of SST analyses with respect to bias-corrected pSSTW (1950–2010) and CCI (1992–2010), respectively.

	ERSST.v5	ERSST.v4	ERSST.v3b	HadISST	COBE-SST2
Corrected pSSTW (1950–2010)	0.064 ± 0.014	0.152 ± 0.012	0.148 ± 0.016	0.137 ± 0.019	0.040 ± 0.014
CCI SST (1992–2010)	0.034 ± 0.015	0.127 ± 0.013	0.107 ± 0.015	0.080 ± 0.024	0.042 ± 0.007

SST analyses and bias-corrected pSSTW are calculated. The RMSDs are generally much larger than the arithmetic average because of the inclusion of noise. The RMSDs decrease from 1.2°C in the 1950s to 0.8°C over the 1990s–2010s because of increased coverage of pSSTW over 1950–90 (Fig. 2b). The RMSDs are slightly higher over 2000–14 than 1990–2000 because of a reduced number of pSSTW observations. The averaged (1950–2010) RMSDs are 0.93°, 0.95°, 0.95°, 0.96°, and 0.89°C (Table 2, second row) in ERSST.v5, ERSST.v4, ERSST.v3b, HadISST, and COBE-SST2, respectively. The RMSDs confirm that ERSST.v5 improves over its previous versions 4 and 3b, and HadISST; and COBE-SST2 performs the best among those SST products.

The differences between those SST analyses and pSSTW vary over the global oceans. The averaged differences between 1950 and 2016 show that those SST analyses are overall warmer than pSSTW in most of the Pacific, Atlantic, and Southern Oceans (Fig. 7). The positive differences in these regions are clearly weaker in COBE-SST2 and ERSST.v5 than in ERSST.v4, ERSST.v3b, and HadISST, which is consistent with the globally averaged differences shown in Fig. 6a and Table 1. The SST differences in the Arctic and Indian Ocean differ a lot among those SST analyses. Large positive differences are found in the Arctic region in COBE-SST2 (Fig. 7e) and HadISST (Fig. 7d). Negative differences are found in the Arctic in ERSST.v5 (Fig. 7a), ERSST.v4 (Fig. 7b), and ERSST.v3b (Fig. 7c), and are found in the Indian Ocean in ERSST.v5 (Fig. 7a), HadISST (Fig. 7d), and COBE-SST2 (Fig. 7e).

The magnitude of the differences between 1950 and 2016 is quantified by the RMSD between SST analyses and pSSTW (Fig. 8). The distributions of RMSDs over the global oceans are very similar among those SST analyses. The magnitude of RMSDs is systematically large in the northwestern North Pacific, eastern equatorial Pacific, northwestern North Atlantic, Arctic, and Southern Ocean where in situ observations (ship, buoy,

Argo, and pSSTW) are sparse. In contrast, the RMSDs are relatively small in most of the tropical oceans between 30°S and 30°N.

b. Comparison with CCI SST

As CCI SSTs over 1992–2010 are independent from ShipBuoyArgo SSTs (Merchant et al. 2014), CCI SSTs are used to evaluate the SST analyses from ERSST.v5, ERSST.v4, ERSST.v3b, HadISST, and COBE-SST over the global oceans (Fig. 9a). Figure 9a clearly shows that, on global average, the disagreement from CCI is much less in ERSST.v5 (0.034° ± 0.015°C; Table 1, third row) than in ERSST.v4 (0.127° ± 0.013°C), ERSST.v3b (0.107° ± 0.015°C), and HadISST (0.080° ± 0.024°C), and slightly less than in COBE-SST2 (0.042° ± 0.007°C). The RMSDs are smaller in ERSST.v5 (0.51°C; Table 2, third row) than ERSST.v4 (0.54°C), ERSST.v3b (0.54°C), and HadISST (0.55°C), and slightly higher than COBE-SST2 (0.46°C). Therefore, the performance of ERSST.v5 is in a better agreement with independent CCI SSTs than its previous versions 4 and 3b over 1992–2010. Overall, the differences and RMSDs of SST analyses deviated from CCI SST are smaller than those from pSSTW, which may largely be associated with much higher area coverage of CCI SST (near 100%) than pSSTW (5%–25%).

The spatial distributions of the averaged difference (Fig. 10) and RMSDs (Fig. 11) between SST analyses and CCI are very similar to those between SST analyses and pSSTW. The SST analyses are warmer than CCI in the northern North Pacific, northern North Atlantic, and Southern Ocean. However, the SST analyses are lower than CCI in the Indian Ocean, tropical Pacific, and tropical Atlantic in ERSST.v5 (Fig. 10a), HadISST (Fig. 10d), and COBE-SST2 (Fig. 10e). The magnitude of RMSD is higher in the northern North Pacific, northern North Atlantic, Arctic, and Southern Ocean (Fig. 11), which is very similar among selected SST analyses. The magnitude of averaged differences and RMSDs are clearly lower

TABLE 2. RMSDs of SST analyses with respect to bias-corrected pSSTW (1950–2010) and CCI (1992–2010), respectively (°C).

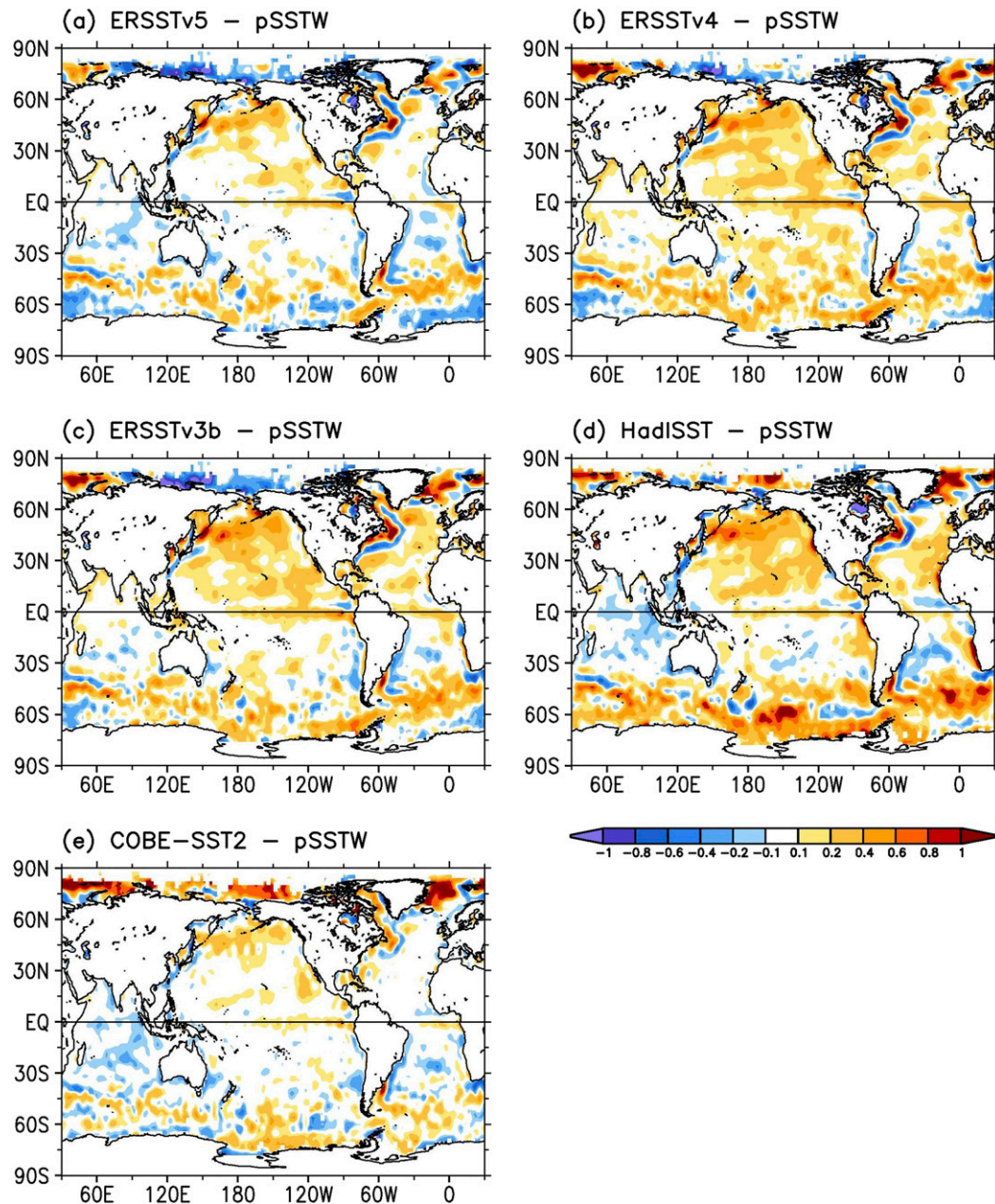|	ERSST.v5	ERSST.v4	ERSST.v3b	HadISST	COBE-SST2
Corrected pSSTW (1950–2010)	0.93	0.95	0.95	0.96	0.89
CCI SST (1992–2010)	0.51	0.54	0.54	0.55	0.46

FIG. 7. Averaged (1950–2016) differences of (a) ERSST.v5, (b) ERSST.v4, (c) ERSST.v3b, (d) HadISST, and (e) COBE-SST2. The differences are calculated relative to pSSTW. Shading intervals are ±0.1°, ±0.2°, ±0.4°, ±0.6°, ±0.8°, and ±1°C.

in COBE-SST2 than in other SST analyses in the Southern Ocean, which results in a lower RMSD of global average in COBE-SST than other SST analyses, as shown in Fig. 9b and Table 2. Overall, the differences between SST analyses and CCI are smaller than those between SST analyses and pSSTW, which is consistent with the globally averaged results shown in Figs. 6b and 9b.

5. Summary and discussion

The pSSTW data from ocean profile measurements of RT/CTD, MBT, and XBT over 1920–2016 are retrieved from the WOD at depth of 0–5 m. The numbers of monthly pSSTW data range from 10^3 to 10^4, and their global area coverages over $2° \times 2°$ grid boxes range from 5% to 25% over 1950–2016. The biases of MBT and
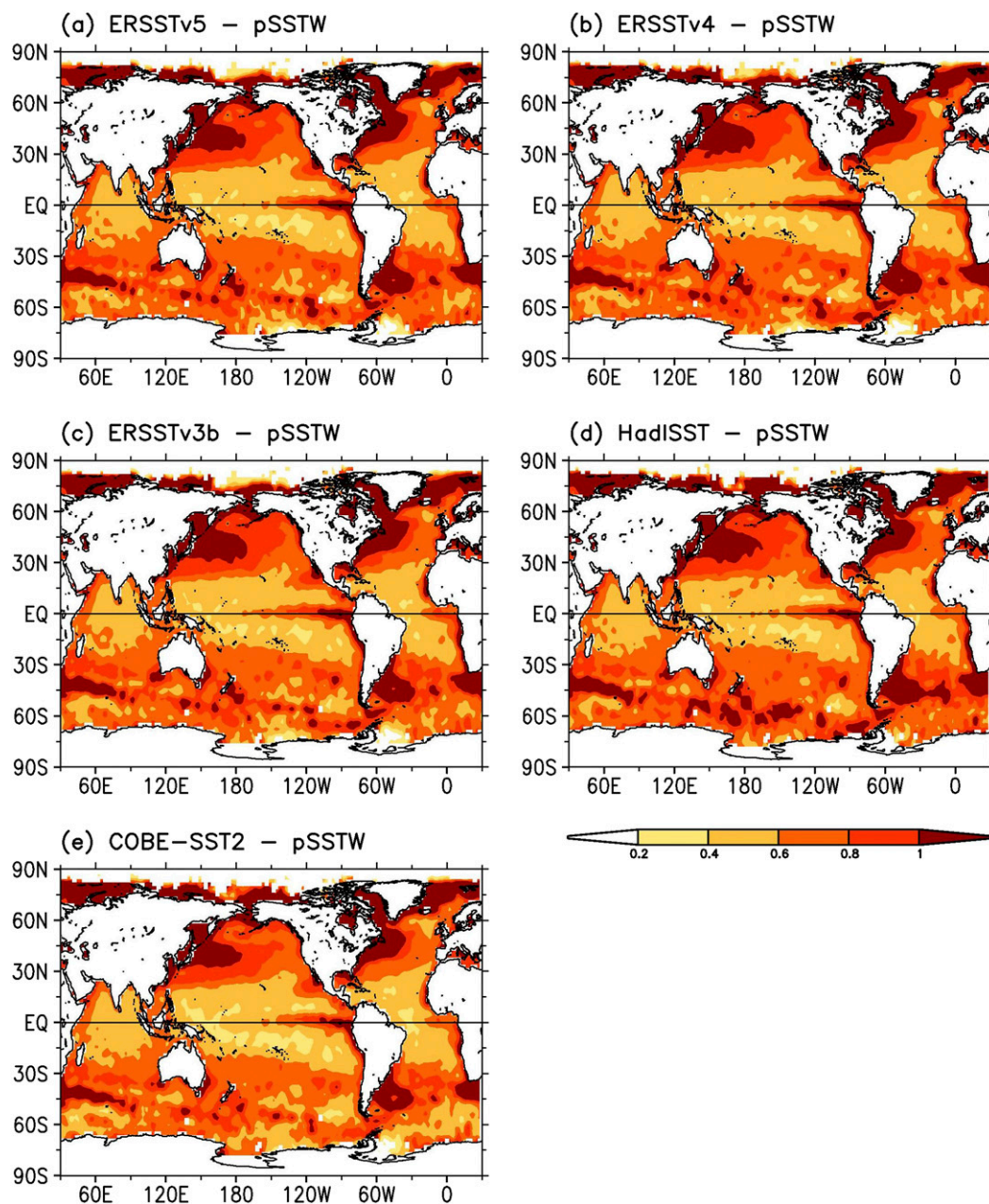
FIG. 8. RMSDs of (a) ERSST.v5, (b) ERSST.v4, (c) ERSST.v3b, (d) HadISST, and (e) COBE-SST2. RMSDs are calculated relative to pSSTW from 1950 to 2016. Shading intervals are 0.2°, 0.4°, 0.6°, 0.8°, and 1°C.

XBT measurements are corrected according to L09 and C14, respectively. The pSSTW data from 1950 to 2016 are used to evaluate commonly available centennial-scale SST analyses of ERSST.v5, ERSST.v4, ERSST.v3b, HadISST, and COBE-SST2. The averaged (1950–2010) difference from pSSTW is 0.06°C in ERSST.v5, which is smaller than those in ERSST.v4, ERSST.v3b, and HadISST (0.14°–0.15°C), but slightly larger than that in COBE-SST2 (0.04°C). Over 1992–2010,

the ESA CCI SSTs are used to evaluate these SST analyses. The averaged difference from CCI is as small as 0.03°C in ERSST.v5, which is much smaller than those in ERSST.v4, ERSST.v3b, and HadISST (0.08°–0.13°C), and slightly smaller than that in COBE-SST2 (0.04°C).

The disagreement between five SST analyses and pSSTW (or CCI) is large in the northern North Pacific, eastern equatorial Pacific, northern North Atlantic,
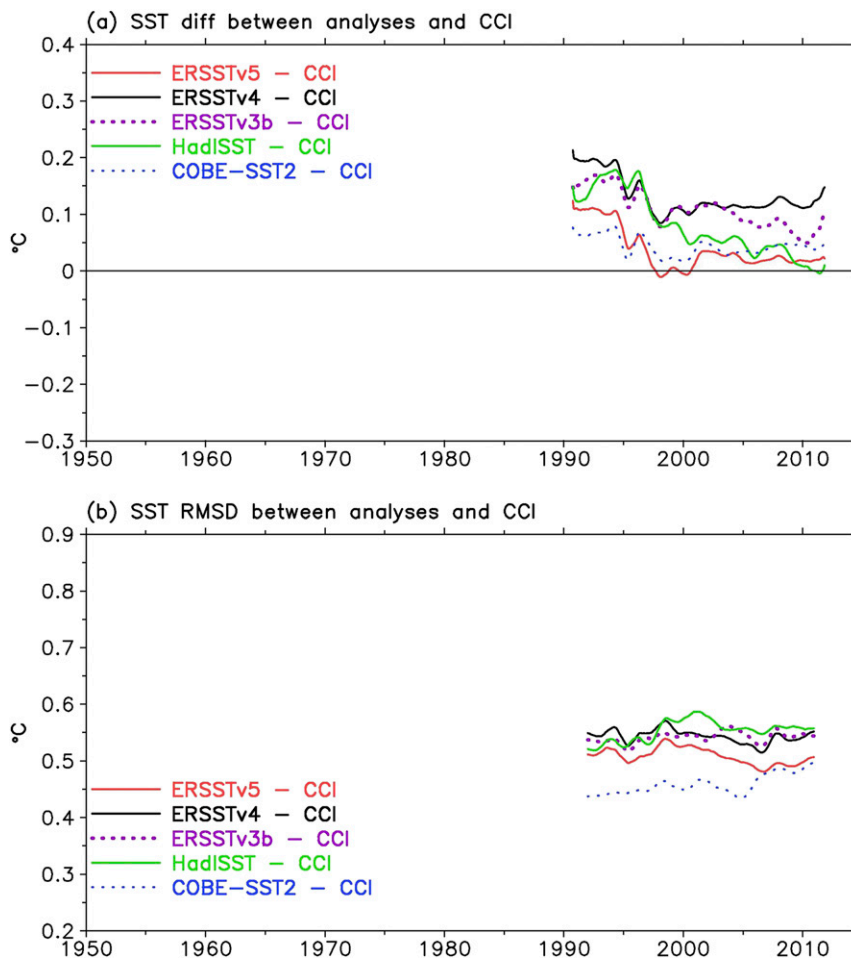
FIG. 9. (a) Globally averaged SST differences between analyses and CCI SST, and (b) as in (a), but for RMSD. A 12-month running average has been applied for plotting.

and Southern Ocean, and smaller in most of the tropical oceans between 30°S and 30°N. These structures are directly associated with the availability of pSSTW in those regions. The disagreement in the Pacific and Atlantic is relatively smaller in COBE-SST2 and ERSST.v5 than in the other three SST analyses when they are compared against pSSTW. In contrast, the disagreement in the Southern Ocean is relatively smaller in COBE-SST2 than in the other four SST analyses when they are compared against CCI.

Previous studies (e.g., Gouretski and Koltermann 2007; L09; C14) showed that pSSTW exhibits biases due to XBT and MBT measurements. These biases are usually determined by the difference between XBT–MBT and RT/CTD–Argo measurements. However, our comparison shows that the bias-corrected pSSTW is slightly colder than ShipBuoyArgo SSTs near the ocean surface over 1950–90. XBT and MBT biases are

corrected to RT/CTD and Argo data, while RT/CTD and Argo data usually do not measure to the surface. ShipBuoyArgo data are corrected to drifting buoys at 0.2–0.3-m depth measurement. Therefore, the colder pSSTW may be associated with the deeper observing depth by pSSTW than that by buoys. This difference should be accounted for when using pSSTW (and Argo) in SST reconstruction and evaluation, which may be assessed by comparisons of pSSTW (Argo) and more accurate/homogenized buoy observations (Huang et al. 2017).

In conclusion, the recent ERSST.v5 represents a better analysis and is closer to independent bias-corrected pSSTW and satellite-based measurements than its previous versions over 1950–2010. The accuracy of ERSST.v5 benefits from using a bias correction of ship SSTs with reference to a baseline of accurate buoy SSTs, as well as the advanced knowledge and techniques in reconstructing the temporal
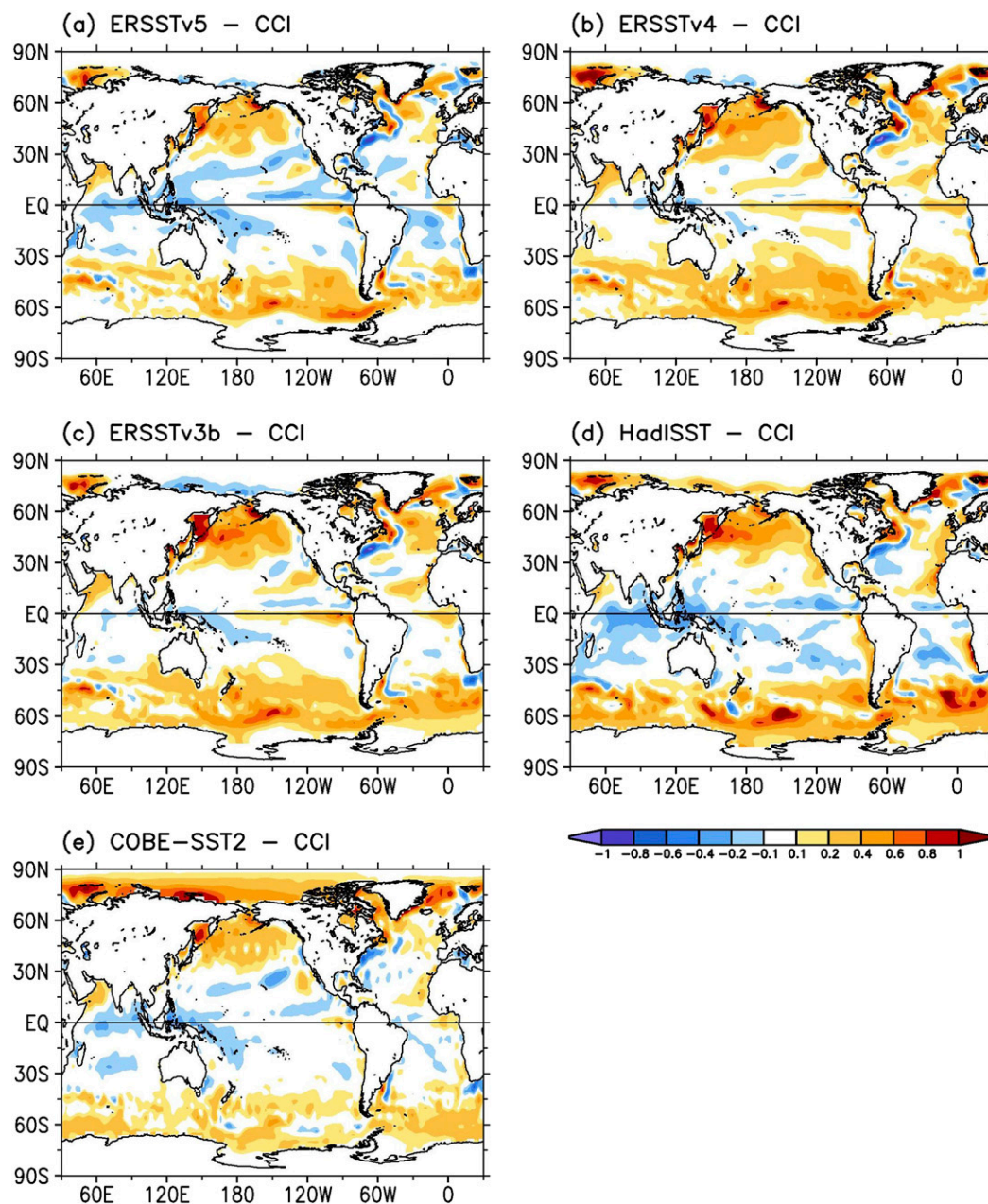
FIG. 10. Averaged (1992–2010) differences of (a) ERSST.v5, (b) ERSST.v4, (c) ERSST.v3b, (d) HadISST, and (e) COBE-SST2. The differences are calculated relative to CCI. Shading intervals are ±0.1°, ±0.2°, ±0.4°, ±0.6°, ±0.8°, and ±1°C.

and spatial variations of SSTs at centennial and global scales (Huang et al. 2017). It is worth noting that, even at a global scale, there remains a difference of 0.0°–0.2°C between SST observations and SST analyses. Such a difference potentially has many sources, as noted, but it suggests a lower limit on uncertainty in SST analyses until such a time when these sources can be attributed.
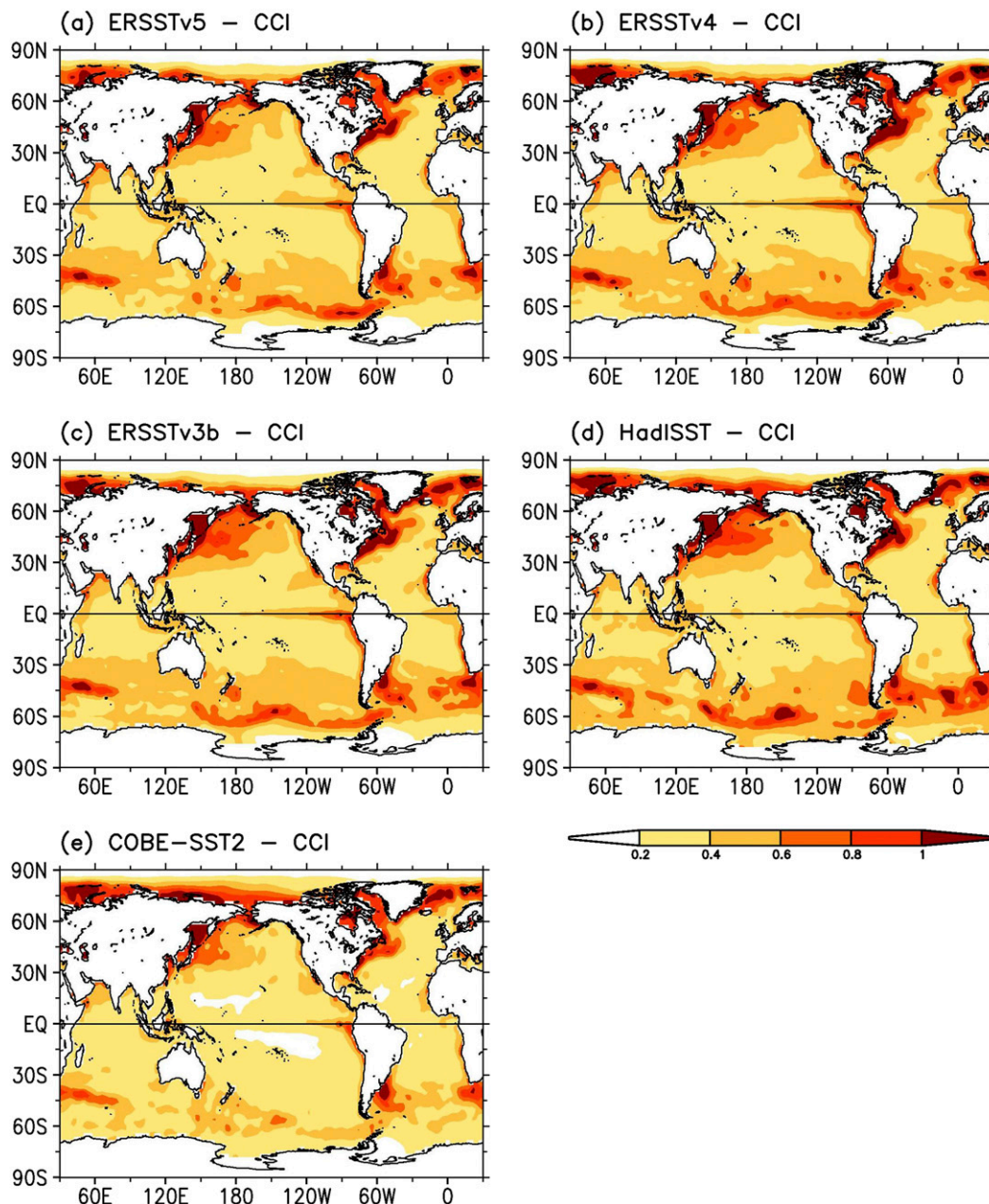
FIG. 11. RMSDs of (a) ERSST.v5, (b) ERSST.v4, (c) ERSST.v3b, (d) HadISST, and (e) COBE-SST2. RMSDs are calculated relative to CCI from 1992 to 2010. Shading intervals are 0.2°, 0.4°, 0.6°, 0.8°, and 1°C.

https://www.ncdc.noaa.gov/data-access/marineocean-data/extended-reconstructed-sea-surface-temperature-ersst); HadISST analysis available online at http://www.metoffice.gov.uk/hadobs/hadisst); COBE-SST2 analysis (available online at https://climate.mri-jma.go.jp/pub/ocean/cobe-sst2/); CCI SST data (available online at http://browse.ceda.ac.uk/browse/neodc/esacci/sst/data/lt/Analysis/L4/v01.1, https://doi.org/10.5285/79229cee-71ab-48b6-b7d6-2fceccead938); and the WOD data. The opinions expressed in this paper are those of the authors

alone and do not necessarily reflect official NOAA, U.S. Department of Commerce, or U.S. government policy.

REFERENCES

Argo, 2000: Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANOE, accessed 4 January 2017, https://doi.org/10.17882/42182.

Barale, V., J. F. R. Gower, and L. Alberotanza, Eds., 2010: *Oceanography from Space: Revisited.* Springer Science & Business Media, 237–238.

Boyer, T. P. and Coauthors, 2016: World Ocean Database 2013 (NCEI accession 0117075), version 1.1. NOAA/National Centers for Environmental Information, accessed 4 January 2017, doi:10.7289/V54Q7S16.

Carella, G., J. J. Kennedy, D. I. Berry, S. Hirahara, C. J. Merchant, S. Morak-Bozzo, and E. C. Kent, 2018: Estimating sea surface temperature measurement methods using characteristic differences in the diurnal cycle. *Geophys. Res. Lett.*, **45**, 363–371, https://doi.org/10.1002/2017GL076475.

Cheng, L., J. Zhu, R. Cowley, T. Boyer, and S. Wijffels, 2014: Time, probe type, and temperature variable bias corrections to historical expendable bathythermograph observations. *J. Atmos. Oceanic Technol.*, **31**, 1793–1825, https://doi.org/10.1175/JTECH-D-13-00197.1.

——, and Coauthors, 2016: XBT science: Assessment of instrumental biases and errors. *Bull. Amer. Meteor. Soc.*, **97**, 924–933, https://doi.org/10.1175/BAMS-D-15-00031.1.

Cowley, R., S. Wijffels, L. Cheng, T. Boyer, and S. Kizu, 2013: Biases in expendable bathythermograph data: A new view based on historical side-by-side comparisons. *J. Atmos. Oceanic Technol.*, **30**, 1195–1225, https://doi.org/10.1175/JTECH-D-12-00127.1.

Dare, R. A., and J. L. McBride, 2011: The threshold sea surface temperature condition for tropical cyclogenesis. *J. Climate*, **24**, 4570–4576, https://doi.org/10.1175/JCLI-D-10-05006.1.

DiNezio, P. N., and G. J. Goni, 2010: Identifying and estimating biases between XBT and Argo observations using satellite altimetry. *J. Atmos. Oceanic Technol.*, **27**, 226–240, https://doi.org/10.1175/2009JTECHO711.1.

Donlon, C. J., T. J. Nightingale, T. Sheasby, J. Turner, I. S. Robinson, and W. J. Emergy, 1999: Implications of the oceanic thermal skin temperature deviation at high wind speed. *Geophys. Res. Lett.*, **26**, 2505–2508, https://doi.org/10.1029/1999GL900547.

Dutton, J. F., C. J. Poulsen, and J. L. Evans, 2000: The effect of global climate change on the regions of tropical convection in CSM1. *Geophys. Res. Lett.*, **27**, 3049–3052, https://doi.org/10.1029/2000GL011542.

Folland, C. K., and D. E. Parker, 1995: Correction of instrumental biases in historical sea surface temperature data. *Quart. J. Roy. Meteor. Soc.*, **121**, 319–367, https://doi.org/10.1002/qj.49712152206.

Freeman, E., and Coauthors, 2017: ICOADS release 3.0: A major update to the historical marine climate record. *Int. J. Climatol.*, **37**, 2211–2237, https://doi.org/10.1002/joc.4775.

Fu, R., A. D. Del Genio, and W. B. Rossow, 1990: Behavior of deep convection clouds in the tropical Pacific deduced from ISCCP radiations. *J. Climate*, **3**, 1129–1152, https://doi.org/10.1175/1520-0442(1990)003<1129:BODCCI>2.0.CO;2.

Gille, S. T., 2009: Diurnal varying wind forcing and upper ocean temperature: Implications for the ocean mixed layer. *16th Conf. on Air–Sea Interaction*, Phoenix, AZ, Amer. Meteor. Soc., 4A.4, https://ams.confex.com/ams/pdfpapers/149958.pdf.

Goes, M., M. Baringer, and G. Goni, 2015: The impact of historical biases on the XBT-derived meridional overturning circulation estimates at 34°S. *Geophys. Res. Lett.*, **42**, 1848–1855, https://doi.org/10.1002/2014GL061802.

Gouretski, V., and K. P. Koltermann, 2007: How much is the ocean really warming? *Geophys. Res. Lett.*, **34**, L01610, https://doi.org/10.1029/2006GL027834.

——, and F. Reseghetti, 2010: On depth and temperature biases in bathythermograph data: Development of a new correction scheme based on analysis of a global ocean database. *Deep-Sea Res. I*, **57**, 812–833, https://doi.org/10.1016/j.dsr.2010.03.011.

——, J. Kennedy, T. Boyer, and A. Köhl, 2012: Consistent near-surface ocean warming since 1900 in two largely independent observing networks. *Geophys. Res. Lett.*, **39**, L19606, https://doi.org/10.1029/2012GL052975.

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

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

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

——, and Coauthors, 2017: Extended Reconstructed Sea Surface Temperature version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Climate*, **30**, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1.

Ishii, M., and M. Kimoto, 2009: Reevaluation of historical ocean heat content variations with time-varying XBT and MBT depth bias corrections. *J. Oceanogr.*, **65**, 287–299, https://doi.org/10.1007/s10872-009-0027-7.

Johnson, N. C., and S.-P. Xie, 2010: Changes in the sea surface temperature threshold for tropical convection. *Nat. Geosci.*, **3**, 842–845, https://doi.org/10.1038/ngeo1008.

JPOTS, 1991: *Processing of Oceanographic Station Data.* UNESCO, 138 pp.

Karl, T. R., and Coauthors, 2015: Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, **348**, 1469–1472, https://doi.org/10.1126/science.aaa5632.

Kennedy, J. J., 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, https://doi.org/10.1029/2010JD015218.

——, ——, ——, ——, and ——, 2011b: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *J. Geophys. Res.*, **116**, D14104, https://doi.org/10.1029/2010JD015220.

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, https://doi.org/10.1002/jgrd.50152.

——, and Coauthors, 2017: A call for new approaches to quantifying biases in observations of sea surface temperature. *Bull. Amer. Meteor. Soc.*, **98**, 1601–1616, https://doi.org/10.1175/BAMS-D-15-00251.1.

Levitus, S., J. I. Antonov, T. P. Boyer, R. A. Locarnini, H. E. Garcia, and A. V. Mishonov, 2009: Global ocean heat content 1955–2008 in light of recently revealed instrumentation problems. *Geophys. Res. Lett.*, **36**, L07608, https://doi.org/10.1029/2008GL037155.

Lumpkin, R., and M. Pazos, 2007: Measuring surface currents with Surface Velocity Program drifters: The instrument, its data, and some recent results. *Lagrangian Analysis and Prediction of Coastal and Ocean Dynamics*, A. Griffa et al., Eds., Cambridge University Press, 39–67.

Matthews, J. B. R., 2013: Comparing historical and modern methods of sea surface temperature measurement—Part 1: Review of methods, field comparisons and dataset adjustments. *Ocean Sci.*, **9**, 683–694, https://doi.org/10.5194/os-9-683-2013.

McTaggart-Cowan, R., L. F. Bosart, C. A. Davis, E. H. Atallah, J. R. Gyakum, and K. A. Emanuel, 2006: Analysis of Hurricane Catarina (2004). *Mon. Wea. Rev.*, **134**, 3029–3053, https://doi.org/10.1175/MWR3330.1.

Merchant, C. J., and Coauthors, 2014: Sea surface temperature datasets for climate applications from phase 1 of the European Space Agency Climate Change Initiative (SST CCI). *Geosci. Data J.*, **1**, 179–191, https://doi.org/10.1002/gdj3.20.

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

Reverdin, G., F. Marin, B. Bourlés, and P. Lherminier, 2009: XBT temperature errors during French research cruises (1999–2007). *J. Atmos. Oceanic Technol.*, **26**, 2462–2473, https://doi.org/10.1175/2009JTECHO655.1.

Reynolds, R. W., and T. M. Smith, 1994: Improved global sea surface temperature analyses using optimum interpolation. *J. Climate*, **7**, 929–948, https://doi.org/10.1175/1520-0442(1994)007<0929:IGSSTA>2.0.CO;2.

——, 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, https://doi.org/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, https://doi.org/10.1175/2007JCLI1824.1.

Roemmich, D., J. Church, J. Gilson, D. Monselesan, P. Sutton, and S. Wijffels, 2015: Unabated planetary warming and its ocean structure since 2006. *Nat. Climate Change*, **5**, 240–245, https://doi.org/10.1038/nclimate2513.

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