# An ensemble data set of sea-surface temperature change from 1850: the Met Office Hadley Centre HadSST.4.0.0.0 data set

**Enter authors here: J.J. Kennedy[1], N.A. Rayner[1], C.P. Atkinson[1], R.E. Killick[1]**

[1]Met Office Hadley Centre, Exeter, UK.

Corresponding author: John Kennedy (john.kennedy@metoffice.gov.uk)

**Key Points:**

- We describe the construction of HadSST.4.0.0.0, a climate data set of sea-surface temperature change from 1850 to 2018.
- A range of bias adjustments was generated to create an ensemble of SST data sets with the ensemble spread partly constrained by oceanographic profile measurements.
- New estimates reduce discrepancy between data sets during the mid 20th century and the recent slowdown in warming, but highlight a divergence in the early 1990s.

**Abstract (max 250 words)**

One of the largest sources of uncertainty in estimates of global temperature change is that associated with the correction of systematic errors in sea-surface temperature (SST) measurements. Despite recent work to quantify and reduce these errors throughout the historical record, differences between analyses remain larger than can be explained by the estimated uncertainties.

We revisited the method used to estimate systematic errors and their uncertainties in version 3 of the Met Office Hadley Centre SST data set, HadSST. Using comparisons with oceanographic temperature-profiles, we make estimates of biases associated with engine-room measurements and insulated buckets and constrain the ranges of two of the more uncertain parameters in the bias estimation: the timing of the transition from uninsulated to insulated buckets in the mid-20th century and the estimated fractions of different measurement methods used.

Here, we present HadSST.4.0.0.0, based on release 3.0.0 and 3.0.1 of the International Comprehensive Ocean-Atmosphere Data Set supplemented by drifting buoy measurements from the Copernicus Marine Environmental Monitoring Service. HadSST.4.0.0.0 comprises a 200-member “ensemble” in which uncertain parameters in the SST bias scheme are varied to generate a range of adjustments. The evolution of global average SST in the new data set is similar to that in other SST data sets and the difference between data sets is reduced during the mid-20$^{th}$ century. However, the changes also highlight a discrepancy in the global-average difference between adjusted SST and marine air temperature in the early 1990s and hence between HadSST.4.0.0.0 and, the NOAA SST data set, ERSSTv5.

## 1 Introduction

Sea-surface temperature (SST) has been designated an essential climate variable (ECV, Bojinksi et al. 2014) that "critically contributes to the characterization of Earth's climate" (WMO https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables). The Global Observing Systems Information Centre (GOSIC) website states, "*Together with air temperature over land, sea-surface temperature is the most important variable for determining the state of the climate system*". It is a key variable for detection of climate change and assessing the relative importance of anthropogenic and natural influences on global climate (http://www.gosic.org/content/gcos-oceanic-surface-ecv-sea-surface-temperature). Atmospheric and oceanic reanalyses (e.g. Kalnay et al. 1996, Hersbach et al. 2015, Carton and Geise 2008, Compo et al. 2011), which are some of the most widely used and cited tools in weather and climate science, typically use SST data sets to provide a lower (or upper) boundary condition. Consequently, there is some value in understanding the long-term evolution of SST and its uncertainties.

Historical SST measurements are to be found, digitized, in great numbers in the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) alongside many other marine meteorological variables. The two most recent major releases of ICOADS are release 2.5 (Woodruff et al. 2010), which contains 261 million records and covers 1662-2007 and release 3.0 (Freeman et al. 2016), which contains over 455 million individual marine reports and covers 1662-2014. Although SST measurements are few in the very early record, they become much more numerous in the latter half of the nineteenth century. Thus far, only ERSSTv5 (currently 1880-2018, Huang et al. 2017) has made use of the far greater number of measurements available in ICOADS release 3.0. Other major historical gridded SST data sets, which run from the latter half of the 19th century, such as COBE-SST-2 (1880-2010, Hirahara et al. 2014), ERSSTv4 (currently 1880-2018, Huang et al. 2015, Liu et al. 2015), and HadSST.3.1.1.0 (currently 1850-2018, Kennedy et al. 2011a; Kennedy et al. 2011b) use ICOADS release 2.5.

One particular difficulty associated with making long data sets based on SST measurements for use in climate analyses is that the technology used to measure SST has changed so much over the past one and a half centuries (Kent et al. 2010, Kent et al. 2017). Even subtle changes in the way that measurements are carried out can lead to systematic errors in the measured trends and the historical changes have not been especially subtle. The magnitude of the estimated errors are of order 0.1-1.0°C, similar to climatic variations over the same period (Hartmann et al. 2013). It is therefore necessary to correct these systematic errors and quantify the residual uncertainties to better understand what actually happened. Each of the previously mentioned data sets – COBE-SST-2, ERSSTv4/v5, and HadSST.3.1.1.0 - applies adjustments to correct systematic errors in the data and provides some estimate of the uncertainty.

Kennedy et al. (2011b) generated an ensemble of one hundred members which comprise the HadSST.3.1.1.0 data set. They calculated a range of corrections by varying poorly-constrained parameters in their bias-adjustment scheme. They used metadata from a number of sources (for example, instructions to marine observers) to assign a measurement method to each observation and took estimates of the systematic errors associated with each measurement method from the literature. The residual uncertainty was combined with uncertainties from other sources of error such as sampling and local measurement errors (Kennedy et al. 2011a and b). An ensemble approach to quantifying uncertainties was also used for the ERSSTv4 data set (Liu et al. 2015, Huang et al. 2016). The parameters they varied to generate the ensemble were associated with all steps in the data-set creation process and only a subset relates directly to the correction of systematic errors. The ERSST SST adjustments are based on comparisons with gridded Night Marine Air Temperature (NMAT) data from the HadNMAT2 data set (Kent et al. 2013) and on estimated differences between SSTs measured by ships and buoys. Hirahara et al. (2014) did not use an ensemble approach, but did provide statistical estimates of the uncertainties in the adjustments applied to the COBE-SST-2 data set. In addition, they used the data to improve the estimates of some of the uncertain parameters identified in Kennedy et al. (2011b).

Despite the efforts of these researchers, significant differences remain between the data sets even at a global scale. The most notable differences (Kent et al. 2017) are between HadSST.3.1.1.0 and ERSSTv5 in the period around the Second World War and since the late 1990s. Larger differences earlier in the record are within the joint uncertainty range.

The period around the Second World War is a period marked by profound uncertainty in the SST record. The war years saw a prolonged El Niño (from late 1939 to 1942, Brönnimann 2005) bringing a peak in global temperatures. The end of the war coincided with a shift in the phase of the Pacific Decadal Oscillation (Newman et al. 2016) with an ensuing period of relatively stable or declining global temperature. A change in both the pattern of international shipping and the composition of data sources available in ICOADS (Thompson et al. 2008) occurs at the same time, confounding a straightforward understanding of the events. The following decades, from 1950 to 1970, witnessed large and poorly documented changes in the way that measurements were made with the development of high-tech insulated buckets and a long-term shift towards measurements being made in the engine rooms of ships: also known as ERI measurements. The "I" has been variously taken to mean intake, inlet and injection.

Even prior to the Second World War, when buckets were the primary means of sampling seawater to measure SST on ships, there are geographical and seasonal differences between the adjustments in ERSSTv5 and HadSST.3.1.1.0. The differences arise from the assumed dependence of the biases on weather conditions. Pre war, ERSST uses adjustments that depend on the air-sea temperature difference alone. The adjustments used in HadSST.3.1.1.0 and COBE-SST-2 (Folland and Parker 1995, Rayner et al. 2006), which assume evaporative cooling from the wet surfaces of the bucket, depend not only on the air-sea temperature difference, but also on solar radiation and, critically, the wet-bulb depression (Carella et al. 2017b). The Folland and Parker (1995) model has recently been assessed in the laboratory by Carella et al. (2017b) who found that the model performed well when conditions were known and controlled, but noted that measurement conditions on board ship were typically neither of these things.

Differences between data sets in the modern period, marked by a slow dwindling of the Voluntary Observing Ship (VOS) fleet and the widespread deployment of drifting buoys, are largely within the joint uncertainties of the various SST data sets. However, the period from 2000 to 2013, during which global temperatures increased at a lower rate than some expected, has been intensively studied (Medhaug et al. 2017) and high demands for accuracy have been made of the SST data sets. It would be advantageous to have a more reliable estimate of SSTs during this well-observed period (Karl et al. 2015, Hausfather et al. 2017) the better to establish exactly what happened.

No part of the SST record is simple to understand or without some little mystery of its own. Therefore, the aims of this paper are to revisit and improve the adjustments applied to HadSST.3.1.1.0 and explore the residual uncertainties, paying particular attention to how the measurements were made. We start by describing the data sources used in the analysis in Section 2. Section 3 explains how the data are aggregated onto a regular grid and how uncertainties associated with the uncorrelated measurement errors and under-sampling are estimated. The bias adjustments, which deal with other, correlated measurement errors and the creation of the ensemble, are the focus of Section 4 with some of the technical details included in the Appendix. Results are in Section 5 before we finish up in Section 6 with the presentation of the new HadSST.4.0.0.0 data set, some discussion and general conclusions.

Throughout this paper, we frequently cite Kennedy et al. (2011a, 2011b and 2011c) as well as Rayner et al. 2006 as this data set builds directly on these papers. We refer to these as K11a, K11b, K11c and R06 for brevity. Also frequently cited are Folland and Parker (1995), hereafter FP95, and Smith and Reynolds (2002), which we shorten to SR02.

## 2 Data

We use various data sets in the analysis and for comparison and validation; they are described in the following subsections. The main analysis is based on the International Comprehensive Ocean-Atmosphere Data Set (ICOADS, Freeman et al. 2017) and water temperatures from HadIOD.1.2.0.0, the Met Office Hadley Centre Integrated Ocean Database (Atkinson et al.

2014). For inter comparison, we use ERSSTv4 (Huang et al. 2015), ERSSTv5 (Huang et al. 2017) and COBE-SST-2 (Hirahara et al. 2014) which have already been introduced. We also use HadNMAT.2.0.1.0 (Kent et al. 2013). For validation we use independent satellite SST retrievals from the ATSR (Along Track Scanning Radiometer) Reprocessing for Climate (ARC) data set (Merchant et al. 2012) as well as some instrumentally homogeneous (Hausfather et al. 2017) subsets of the HadIOD and ICOADS data sets. First, however, it is useful to determine exactly what it is we mean by sea-surface temperature.

### 2.1 Which sea-surface temperature?

Traditionally, long-term *in situ* SST data sets have been considered to be representative of a loosely defined "bulk" SST, which covers a range of measurements made in the upper 10m or so of the water column. However, the daily formation and erosion of a stably-stratified near-surface warm layer in the oceans, particularly during calm, sunny conditions, can lead to strong temperature gradients in the upper 10m (Kawai and Wada 2007) and make it harder to reconcile measurements of water temperature made at different depths. Satellite retrievals of SST are especially prone to this as they are sensitive to temperatures in a very shallow layer where diurnal warming is most pronounced. For instruments measuring in the infra red, this layer is measured in micrometres and its temperature is referred to as the "skin" temperature. Uncertainties associated with modern measurement systems are now sufficiently small that temperature variations with depth are readily detectable in the aggregate and need to be accounted for where detailed comparisons are made (see e.g. Merchant et al. 2012).

Donlon et al. (2007) recommend that all SST measurements be accompanied by an estimate of the depth at which the measurement was made. However, this information is rarely available for historical observations and, where it is available, it does not usually take into account changes due to variable loading of the ship. Even drifting buoys from a single manufacturer, where the design and materials are identical, will measure at a varying depth owing to continual movement of the water and the potential loss of the drogue.

In this paper, rather than specify that SSTs are estimated for a particular depth, we will instead use SST measurements from drifting buoys as our reference. This is conventionally reckoned equivalent to an SST measurement at an approximate depth of 20cm. We make use of near-surface water temperatures measured at a range of depths, as described in the following subsections, but the aim throughout is to adjust these measurements so that they would closely match coincident observations from drifting buoys even if they occurred long before drifting buoys were first dreamed of.

A final note on nomenclature. Throughout this paper we use the words “error” and “uncertainty” as they are defined in Annex B of the Guide to the Expression of Uncertainty in Measurement (JCGM 2008).

**uncertainty** (of measurement): *parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand.*

**error** (of measurement): *result of a measurement minus a true value of the measurand.*

A *measurand* is a “particular quantity subject to measurement”.

## 2.2 Surface meteorological data

The data and metadata used in this analysis are from Release 3.0 of the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) for the period 1850-2014 (Freeman et al. 2017; downloaded version Research Data Archive 2016). An update to 2017 uses the near-real time data from ICOADS release 3.0.1. We will use ICOADS release 3.0.1 as the basis for monthly updates of the data set.

Due to a drop off in drifting buoy observations in ICOADS release 3.0.1 which followed the switch of data transmission codes from TAC (Traditional Alphanumeric Codes) to BUFR (Binary Universal Form for the Representation of meteorological data) in November 2016, we supplement the ICOADS data from January 2016 onwards with near-real-time drifting buoy observations downloaded from CMEMS (Copernicus Marine Environmental Monitoring Service, ftp://nrt.cmems-du.eu/Core/INSITU_GLO_NRT_OBSERVATIONS_013_030/monthly/drifter/). The initial download was made on 5 April 2018 and then regular downloads are made each month to gather data for the preceding month. The drifting buoys from this near-real-time source completely replace the drifting buoys from ICOADS release 3.0.1 in our analysis in the overlapping months. This increases the data volume and observational coverage significantly from November 2016.

We filtered the data to remove coastal stations, non-standard moored buoys (principally around the coast of the US) and other non-standard platforms like oil rigs and research stations. These sources cover a relatively small area and vary widely in design making the biases heterogeneous and difficult to assess. C-MAN and many US Coastal moored buoys were excluded because, as well as being of diverse type, they are often found in estuaries and lagoons that are not representative of a wider area and certainly not of the areas typically sampled during the climatological base period 1961 to 1990. Many coastal moored buoys also produce very high data volumes – sometimes making several measurements an hour – and can potentially skew the processing. A list of the excluded ICOADS platform IDs is available along with the HadSST.4.0.0.0 data set.

Oceanographic measurements from the World Ocean Database (WOD) were also removed by excluding ICOADS deck 780 (1850-2014). Excluding WOD measurements from the processing means that we can use sub-surface measurements as an independent data set for assessing biases and for validation (e.g. Gouretski et al. 2012, Huang et al. 2018). The remaining data were quality controlled (QC'd) to remove outliers and low-quality measurements (an update of R06, https://github.com/ET-NCMP/MarineQC). Figure 1 shows the number of observations passing QC for each month from January 1850 to December 2018.

R06 describes the creation of the climatology we use to calculate the gridded anomalies. It is based on *in situ* measurements made between 1961 and 1990. The climatology has a resolution of 1° of latitude, 1° of longitude and 5 days. A standard 5-day period is a pentad. The first pentad of each year is 1-5 January. The calendar is divided into pseudo months. Each pseudo-month has six pentads except August, which has seven (hence the annual peaks in Figure *1*). Leap days are accommodated by extending the pentad in which they fall.

### 2.2.1 Initial metadata assignment

We assign a measurement method to each observation in ICOADS. Any particular report could be either a buoy measurement, a bucket measurement, an ERI measurement, a measurement made with a hull sensor or else unknown. Where we could not definitively assign a single method, a fractional assignment was attempted based on the fraction of the recruiting country's fleet that used each method (see also K11b). Fractions were estimated based on ships recorded in WMO Publication 47 (WMO Pub. 47, Kent et al. 2007) for that year. An assignment from the ICOADS metadata (SI or SIM indicating bucket, ERI or hull sensor) was preferred. Fractional assignments are always incorrect at the level of individual reports, but should give representative averages when aggregating large numbers of observations. The assignments are uncertain, even where there is a definitive assignment, and we refine the estimates using comparisons between the ship data and oceanographic profile data in Section 4.1.4.

The procedure for assigning metadata to a particular ICOADS report is as follows. The procedure terminates as soon as an assignment is made (abbreviations in brackets refer to the variable names in the IMMA, International Maritime Meteorological Archive format, documentation for ICOADS release 3.0.0 http://icoads.noaa.gov/e-doc/imma/R3.0-imma1.pdf):

1. If the ICOADS platform type (PT) was 6 or 7, we assign the observation to be a moored or drifting buoy measurement respectively.
2. We assigned US ships (C1 = 2) from Deck (DCK) 128 to be ERI in 1968, 1969, 1972 and 1973.

3. UK Royal Navy data, Deck 245, and Russian data from Deck 732 (following Carella et al. 2018) were assigned to be ERI.
4. If an SST measurement method (SI) was present in ICOADS and indicated a bucket, ERI or hull measurement, we used it.
5. If an SST measurement method (SIM) was present in the ICOADS metadata attachment and indicated a bucket, ERI or hull measurement, we used it.
6. Between 1939 and 1945, we set all reports that had not been assigned a measurement method in steps 1-5 to unknown.
7. Before 1939, all reports that had not been assigned a measurement method in steps 1-6 were set to bucket.
8. If a recruiting country (C1) was present in the ICOADS attachment, we used it to assign weights to each report for bucket, ERI, hull and unknown according to the fraction of ships that took each type of measurement in WMO Pub. 47 for that country and year. We assumed that US ships with unknown measurement method were ERI. If WMO Pub. 47 was not available for that year, we used the next available year after the year of interest.
9. If the deck (DCK) could be linked to reports from a particular country (see K11b for details) between 1956 and 1996, then we used it to assign weights to each report of bucket, ERI, hull and unknown according to the fraction of ships that took each type of measurement in WMO Pub. 47 for that country and year. We assumed that US ships with unknown measurement method were ERI. If WMO Pub. 47 was not available for that year, the next available year was used.
10. Any report not assigned a measurement method in steps 1-9 we set to unknown.

At the end of this process, each SST measurement has an assigned measurement method, or fractional assignment. The assignments are provisional and uncertain. In some cases, observations from a ship will be listed as being made with a bucket, when they were in fact ERI measurements. In section 4.1.4, we attempt to estimate what fraction of bucket measurements are mis-identified in this way.

### 2.3 Near-surface sub-surface measurements from HadIOD

HadIOD.1.2.0.0 (Atkinson et al. 2014) is an integrated database of temperature and salinity measurements from oceanographic sources at various depths (from the EN version 4.2.0 data set; Good et al. 2013) combined with surface observations from ICOADS release 2.5.1 (Woodruff et al. 2011). Oceanographic measurements in HadIOD were made with a variety of instruments including Conductivity-Temperature-Depth (CTD), eXpendable BathyThermograph (XBT), Mechanical BathyThermograph (MBT) and Argo devices. For a review of the characteristics of these devices, see Abraham et al. (2013). Each observation has an overall quality flag, an estimated uncertainty and a bias adjustment or bias adjustments. We only use the oceanographic profile measurements from HadIOD between 1930 and 2018 and of these we use only those that were made in the upper 10m of the water column. Measurements in the upper 10m can provide a reasonable approximation for the sea-surface temperature (see Section 2.1). We set aside measurements from Argo floats so that we can use them for independent validation (see Section 5.2 and Appendix 3.1).

A number of XBT and MBT adjustments are included in the HadIOD database. The adjustments are required to correct for known biases in these measurements independent of the SST biases we explore in this paper. We used MBT and XBT adjustments from four analyses (Gouretski and

Reseghetti 2010, Gouretski 2012, Cowley et al. 2013 and Levitus et al. 2009). Two of the analyses provide estimates of both MBT and XBT adjustments (Levitus et al. 2009 and Gouretski and Reseghetti 2010) and we use these as given. Two analyses provide only XBT adjustments (Gouretski 2012 and Cowley et al. 2013). We combined the two XBT-only analyses with each of the two MBT adjustments from the combined analyses to provide four new XBT-MBT combinations, bringing the total number of sets of corrections to six. The six sets of adjusted XBT-MBT data together with measurements from Conductivity-Temperature-Depth (CTD) instruments and bottles are gridded as for the SSTs (Section 3) to make six near-surface reference data sets that we use to provide a set of baselines – albeit uncertain baselines – for the SST bias assessment. We assume that, once corrected, the oceanographic data are less biased than the SST measurements are. However, we note that the adjustment of oceanographic data is also an area of ongoing research with uncertainties all of its own (Abraham et al. 2013, Cheng et al. 2016).

### 2.4 Marine air temperature

The air temperature above the oceans is physically related to the underlying sea-surface temperature. Consequently, sea-surface temperature changes are often used as a proxy for marine air temperature (MAT) changes, for example, in the calculation of global average temperature anomalies (e.g. Morice et al. 2012). Huang et al. (2015) argued, based on the behavior of a particular climate model (the GFDL coupled model, CM2.1), that the difference between MAT and SST anomalies at a global scale was more or less constant, changing by less than 0.1°C in a century. However, other authors have noted that trends in MAT and SST anomalies can diverge, albeit by a relatively small amount on multidecadal time scales (Cowtan et al. 2015). At smaller scales, MAT and SST anomalies can diverge by tens of degrees particularly close to land and sea ice.

Nonetheless, over longer periods and at larger scales, changes in MAT and SST are closely related. Indeed, SR02, FP95 and other related papers used this relationship to adjust for biases in the SST record. However, MAT measurements are not without problems of their own. The principal problem is solar heating of the ship, which biases MAT measurements during the day (Berry et al. 2004) and is typically solved by restricting the use of MAT measurements to those made at night: so called night marine air temperature (NMAT) measurements. The increasing size of ships and the height of temperature sensors above the sea-surface has led to a creeping cold bias in MAT measurements. Warm biases have also been detected during the 19$^{th}$ century and the Second World War, and are thought to be caused by non-standard sensor exposure; for example, reading the thermometer inside rather than on deck. In the creation of the HadNMAT2 data set (1880-2010), Kent et al. (2013) adjusted the data, or excluded certain periods, regions, and data subsets, to account for these biases.

### 2.5 Instrumentally homogeneous data sets

Hausfather et al. (2017) used the term “instrumentally homogeneous” data sets to describe SST data sets that are based on a single type of instrument or group of closely-related instruments, which they considered to be more homogeneous than the general *in situ* SST record. They used these data sets to assess the stability of global-average SST records over the period 1995-2016.

Records that consist of a single type of instrument minimize artificial drifts, shifts or jumps caused by changes in instrumentation. Ideally, measurements from the instruments should also be of high quality, with demonstrably good accuracy and stability. Finally, the measurements should be as independent as possible from the record they are being used to assess. We use three instrumentally homogeneous records based on: Argo floats, Along-Track Scanning Radiometer SST retrievals and buoys. We describe each of these in turn.

#### 2.5.1 Argo

Argo floats are autonomous profiling floats, which move with the prevailing currents at a typical "parking depth" of 1000m descending at regular intervals – usually on a ten-day cycle – to a depth of 2000m and then ascending to the surface taking temperature and salinity measurements. Since around 2007, Argo floats have provided quasi-global sampling of the oceans. The temperature sensors are calibrated before the float is released and the manufacturer's stated accuracy is 0.002°C with a stability of 0.0002°C/year (Abraham et al. 2013). Whether this accuracy is realized in the field is difficult to assess, but floats that have been recovered remained within the manufacturer's stated limits (ibid).

The good coverage of Argo floats, combined with the accuracy of the measurements they make presents a dilemma for the data set creator. On the one hand, it would seem sensible to use high quality measurements like these in the analysis (e.g. Huang et al. 2017). On the other hand, there is much to be gained from using the Argo measurements as independent validation. The latter approach is common in the satellite SST community (see e.g. Merchant et al. 2012, Berry et al. 2018), where drifting buoy data are often used for calibration and thus cannot be used for validation. We adopt the same approach and reserve Argo for validation of the final product. The relatively infrequent sampling provided by Argo – one profile every ten days – when compared to say drifting buoys, which provide one measurement every hour, or even ships, which usually measure once every six hours, means that including Argo in our analysis would have a relatively small impact on the gridded anomalies. However, including Argo leads to a reduction of measurement and sampling uncertainty of up to 30%.

#### 2.5.2 ATSR

The ATSR (Along Track Scanning Radiometer) Reprocessing for Climate (ARC) data set (Merchant et al. 2012) is a "climate-quality" analysis of SST retrievals from the ATSR satellite-based instruments. The ATSR instruments were designed to make climate quality measurements of SST and had a number of features to help achieve this. First, each instrument had a blackbody onboard that allowed for continual calibration checks. Second, the satellites had a dual view configuration, with observations made directly downwards (Nadir) and forwards (55° off vertical) relative to the satellite's motion. The two views allow the satellite to observe the same area of the surface via two different paths through the atmosphere. By comparing the two, it is easier to identify contamination arising from dust in the air or sulfurous volcanic particles. The instruments also had three infrared channels, which allow more sensitive SST retrievals than is possible using the two-channels available on the Advanced Very High Resolution Radiometer (AVHRR) instruments (Merchant et al. 2014).

The ARC project reprocessed SST retrievals from the ATSR instruments. The SSTs we use here are representative of a nominal measurement depth of 20cm (see Section 2.1) and an observation time of 10:30 am and pm local time. Comparisons with Argo and drifting buoys show that there is a minimal residual bias at a global scale (although locally there are deviations of order 0.1°C), and that the uncertainty estimates provided with the data set are reliable. Three-way comparisons with other satellites and buoys show that the individual ATSR retrievals have a typical uncertainty of around 0.15°C (O'Carroll et al, 2008, Lean and Saunders 2013). The stability of the ARC record has been demonstrated in the Tropical Pacific by comparison to moored buoys in the Tropical Atmosphere-Ocean (TAO) array to be of order 0.01°C/decade (Merchant et al. 2012).

The production of the ARC data set is almost entirely independent of SST measurements made *in situ*. There is an indirect dependence as ARC uses reanalysis profiles to estimate the optimal retrieval coefficients and the reanalysis used is driven using SST data sets that incorporate *in situ* measurements.

We use the more reliable dual-view three-channel retrievals of SST from Version 1.1.1 of the ARC data set, which span the entire period of the ATSR record 1991-2012. However, the period from 1991 to 1995 is only intermittently covered due to the failure of one of the infrared channels on the ATSR1 instrument shortly after launch.

### 2.5.3 Buoys

Although drifting and moored buoys are not independent of the data set that we develop in this paper, they do consist of a single type of instrument or closely-related instruments and they are of demonstrably higher quality than ship data (Kennedy 2014). There have been changes to drifting buoy design over the years, but the largest changes had occurred by the early 1990s. There is still some diversity in the design of drifting buoys and they are produced by a number of different manufacturers. The nominal measurement depth varies, but is typically in the range 20-50cm. Most sensors are of a nominal 0.1°C accuracy (Sybrandy et al. 2008), but estimates of measurement uncertainty made in the field vary somewhat. Nevertheless, the buoy record is considerably more homogeneous and stable than the unadjusted ship record over the period 1991 to present.

Moored buoys come in a variety of forms. The measurement depth is typically around 1m. Sensor accuracy is also variable. Representative uncertainty values are given in Table 1. Some moored buoys perform better than the average. In particular, measurements from the moored buoys in the tropical Pacific from the TAO/TRITON array are of generally higher quality than the US coastal arrays (K11c).

## 3 Gridding and basic data preparation

We averaged the SST measurements from the individual reports onto a 5° latitude by 5° longitude monthly grid in a two-step process (R06). First, we sorted the observations into 1° latitude by 1° longitude by pentad bins. Each SST observation was then converted into an anomaly by subtracting the climatological average (for the period 1961-1990) for that 1° latitude by 1° longitude pentad bin. We rejected anomalies with magnitude exceeding 8°C and calculated the Winsorised (a form of trimmed mean, see Wilcox 2001) average of the remaining anomalies. We then sorted the 1° latitude by 1° longitude pentad “super observations” into 5° latitude by 5° longitude pseudo-month bins and took the Winsorised averages of the super-observations in each of the larger bins. Figure 1(b, d, f) shows the number of "super observations" available per month.

The contribution of each observation to the grid-box average has a weight, *w*, equal to

$$w = \frac{1}{ab} \qquad \text{Equation 3.1}$$

where *a* is the number of observations in the same super observation and *b* is the number of super observations in the larger 5° pseudo-month bin. The sum of the weights of all observations in a 5° grid box equals one. Using these weights, we calculated the fractional contributions of different measurement methods to the grid-box average and used these to estimate the fractional contribution of each measurement method to the global and hemispheric averages (Figure 2). Before 1915, bucket measurements have a weight of one in the global average. The weighting does not perfectly reflect the influence of a single observation on the average because of the Winsorisation process.

There are interesting changes in the influence of different measurement methods on the global and hemispheric averages on all time scales. The Second World War stands out because of the abrupt changes at the start and end, but there are rapid changes at other times such as the early 1960s (note that measurements of unknown type were assumed to be from ships). From around 1980 to 2005, the number and influence of buoys increases steadily after a brief peak in the late 1970s arising from a mass deployment of buoys during the First GARP (Global Atmospheric Research Project) Global Experiment (FGGE, Garrett 1980). Despite the large numbers of observations made by buoys since then, amounting to around 90% of all observations in the past ten years (see e.g. Woodruff et al. 2011), the influence on the global average (i.e. the area average of $w$ for buoys as shown in Figure 2) in HadSST.4.0.0.0 has not consistently exceeded 50%. Although buoys make large numbers of observations, they typically do so in a limited area. Observations from ships tend to be fewer, but more widely spread so the effective sampling-per-observation is higher for a ship than it is for a buoy, particularly a moored buoy. This can also be seen in the count of super-observations (Figure 1) which is largest in the 1970-1990 period when the VOS fleet was at its peak.

In addition to the main gridded data set, we also gridded a number of subsets of the data which we used to estimate biases for different measurement methods:

1. moored and drifting buoy observations;
2. all ship measurements;
3. measurements likely to have been ERI measurements (observations identified as an engine room measurement, or taken by a ship from a country where more than 90% of the fleet used engine room measurements at that time);
4. measurements likely to have been bucket measurements using a similar criterion;
5. hull sensor measurements; and
6. observations to which we could not attach a near-definitive measurement method.

### 3.1 Uncertainty estimation for gridded averages

Consider an SST measurement $O_{ij}$ taken by agent $i$ (either a ship or buoy) at space-time point $j$ to be the combination of the true SST, $T_{ij}$, together with a set of error components. The three error components are: $U_{ij}$ associated with uncorrelated errors, $B_{ij}$ associated with "micro-bias" errors that were correlated for all measurements made by the same agent, but uncorrelated between agents, and $G_{ij}$ a "macro-bias" error common to and correlated across all agents of a particular kind e.g. all ships taking ERI measurements.

$$O_{ij} = T_{ij} + U_{ij} + B_{ij} + G_{ij} \qquad \text{Equation 3.2}$$

We calculate uncertainties in the gridded averages, $\sigma_{\text{gridbox}}$, using a variation of the formula from K11a,

$$\boldsymbol{\sigma^2_{gridbox}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \boldsymbol{w^2_{ij}\sigma^2_{u_i}} + \sum_{i=1}^{m} \boldsymbol{w^2_i \sigma^2_{b_i}} + \sum_{h=1}^{o} \boldsymbol{w^2_h \sigma^2_{G_h}} + \boldsymbol{\sigma^2_{sampling}} \qquad \text{Equation 3.3}$$

where $w_{ij}$ is the weight (as defined in Equation 3.1) in the grid-box average of the observation at point $j$ made by agent $i$ in that grid box. There are a total of $m$ agents taking $n_i$ measurements in the grid box. $\sigma_{u_i}$ is the uncertainty associated with uncorrelated measurement errors and $\sigma_{b_i}$ is the uncertainty associated with correlated "micro-bias" measurement errors for agent $i$. $\sigma_{G_h}$ is the uncertainty associated with large-scale correlated errors from a particular measurement method, $h$, of which there are $o$ different types and $w_h$ is the weight assigned to measurements made using method $h$ in that grid box. The values used for $\sigma_u$ and $\sigma_b$ are given in Table 1 and come from K11c. $w_i$ is the sum

$$\boldsymbol{w_i = \sum_{j=1}^{n_i} w_{ij}} \text{ and } \boldsymbol{\sum_{i=1}^{m} w_i = 1} \qquad \text{Equation 3.4}$$

The sampling uncertainty, $\sigma_{sampling}$ is given by:

$$\boldsymbol{\sigma^2_{sampling} = \frac{1}{n_s}\sigma^2_s[1 - \bar{r}]} \qquad \text{Equation 3.5}$$

Where $n_s$ is the number of super observations in the 5° pseudo-month grid box. $\sigma_s$ is the variance at a space-time point within the grid cell (here assumed not to vary across the grid cell) and $\bar{r}$ is the average correlation of space-time points within the grid cell (for details of the calculation see

K11a). We use the number of super observations rather than the number of observations as some agents such as moored buoys make many hundreds of observations in a single location and it is clear that the sampling uncertainty cannot be endlessly reduced by making more observations in the same place. A count of super observations gives a better idea of the number of independent space-time points sampled, though it remains less than perfect. The sampling uncertainty parameters were estimated as in K11a but using the 5° gridded ICOADS release 3.0 data between 1961 and 2016.

## 3.2 Correlated errors

The covariance, *R(p,q)*, associated with correlated "micro bias" errors from the same ships visiting grid boxes *p* and *q* was

$$\boldsymbol{R(p,q) = \sum_{i} w(p)_i w(q)_i \sigma_{b_i}^2} \qquad \text{Equation 3.6}$$

where *i* sums over all the ships that visited grid boxes *p* and *q*.

Correlated errors between grid-box averages can also occur when two different ships using the same measurement method visit different grid cells. In estimating the biases and uncertainties for ERI data (see Appendices A1 and A2) we assumed that $G_{ij}$ (Equation 3.2) was non-zero and equal for all ships making ERI measurements. The covariance is calculated in a similar way to that of the micro-bias errors above.

$$\boldsymbol{R(p,q) = \sum_{h} w(p)_h w(q)_h \sigma_{G_h}(p) \sigma_{G_h}(q)} \qquad \text{Equation 3.7}$$

where $\sigma_{G_h}(p)$ is an estimate of the uncertainty in the large-scale bias associated with measurement method *h* at location *p*. For ERI measurements, $\sigma_{G_h}(p)$, was assumed to be constant, but for bucket measurements, $\sigma_{G_h}(p)$ was assumed to be equal to the bucket adjustment field at that location. In other words, we assume that the shape of the bucket adjustments is known but not the exact magnitude. In effect, the biases are estimated by regression (Appendices A1 and A2), with the covariances defined in Equation 3.6 and 3.7 specifying both the shape of the predictor and its prior variance.

### 3.3 Use of the error model

We use the error model in two distinct ways in this paper. First, it is used to estimate some of the error terms described in Equation 3.2, that is to determine the actual sizes of biases such as $G_{ij}$. The details of this are given in Appendix A1 and A2. Second, the error model is used to estimate uncertainties in derived quantities such as grid box averages (Equation 3.3) as well as regional and global averages. In order to calculate regional and global averages, Equation 3.3 and Equation 3.6 are combined to produce a total error covariance matrix and the uncertainties are propagated using the propagation of uncertainties formula for a weighted average where the weights are the grid-cell areas. The specific form is described in Section 3.3 of K11a.

### 3.4 Example fields

Figure 3 shows example fields for June 2003. The number of observations exceeds 100/month in many grid boxes and is somewhat homogeneous thanks to the widespread use of drifting buoys. In contrast, the number of super observations varies greatly, with the highest numbers in the northern hemisphere shipping lanes, demonstrating again the more efficient sampling per observation afforded by ships. The very localized sampling of moored buoys is also clear.

The uncorrelated error component of the uncertainty is relatively small except for at the edges of the observational coverage, where there are very small numbers of observations contributing to each grid-box average. Uncertainties associated with micro-bias errors are typically larger as they are related to the number of platforms contributing to the average. They are lowest over the north Pacific and North Atlantic where there is a great diversity of measurement platforms.

The greater weight given to ship observations, which are generally less reliable, is potentially a weakness in the simple approach adopted here. A more sophisticated method such as an optimal interpolation scheme (Karspeck et al. 2012) might give a greater weight to the more reliable buoy observations. However, the current method uses Winsorisation, which minimize the effect of outliers and makes it much easier to keep track of the correlation structure of the errors through the full uncertainty calculation.

## 4 Bias Adjustments

The estimation of biases has two basic steps. First, we create grids of the fractional contribution of each measurement type to the gridded averages. Second, we assign biases to each measurement method and calculate an overall bias in each grid cell. The bias B in a grid cell is equal to

$$\boldsymbol{B = f_e E + f_c B_{tc} + f_w B_{tw} + f_r B_{tr} + f_d D} \qquad \text{Equation 4.1}$$

Where $f_e$ is the fractional contribution of ERI and hull sensor measurements, $f_c$ is the fractional contribution of measurements made with canvas, or otherwise-uninsulated, buckets; $f_w$ is the fractional contribution from wooden buckets; $f_r$ is the fractional contribution of measurements made with rubber, or otherwise-insulated modern buckets; and $f_d$ is the fractional contribution of measurement from buoys. E, $B_{tc}$, $B_{tw}$, $B_{tr}$ and D are the biases associated with each of these measurement methods respectively.

In practice, we do not have estimates of the true historical bucket biases, $B_{tc}$ and $B_{tw}$ (the *t* in the subscript stands for "true"). What we do have are the corrections $B_c$ and $B_w$ (from FP95) which adjust a bucket measurement to be consistent with the average bias in the period 1961 to 1990 (rather than adjusting a bucket measurement to the true SST). To convert $B_c$ to $B_{tc}$, it is necessary to calculate:

$$\boldsymbol{B_{tc}} = \frac{\boldsymbol{B_c} + \overline{\boldsymbol{f_r}}\boldsymbol{B_{tr}} + \overline{\boldsymbol{f_e E}}}{\left(\mathbf{1} - \overline{\boldsymbol{f_c}}\right)} \qquad \text{Equation 4.2}$$

where an overbar denotes the average for that value over the period 1961-1990. The derivation, which is a modified from of that in K11b, can be found in Appendix A4. Unlike in K11b, we will make a direct estimate of $B_{tr}$ from the data (Section 4.1.3) so the formula shown here is somewhat different.

Another difference from K11b is that we assume D is zero. In other words, we believe that buoys provide both an accurate measurement of SST and a benchmark (as noted in Section 2.1) for estimating the biases associated with other measurement methods. A similar approach, using drifting buoys as a baseline, is taken in ERSSTv5 (Huang et al. 2017) and drifting buoys are widely used as "ground truth" in the satellite SST community (e.g. Lean and Saunders 2013 and Embury et al. 2012). In contrast, in K11b it was assumed that the time series E was known and the adjusted ERI measurements were used as the baseline. A similar approach was taken for ERSSTv4. Note that the choice of whether to use ships or drifters as a baseline, in so far as this constitutes a constant offset, does not affect the estimation of trends or anomalies. It can, however, affect the estimation of actual SSTs, which are important for some applications. It also contributes to a more intuitive presentation of the time series, with smaller uncertainties in the drifter-rich period (Figure *8*).

In practice, when we come to estimate the biases in the data where we do not know the assignments perfectly, the bias in an average of a group of observations will be some linear combination of biases from both bucket and ERI measurements. For example, if we compare a collection of observations labeled as buckets to a set of unbiased drifting buoy data then the empirical bias B* seen in the bucket-labelled data (in this case assuming that all the buckets are rubber) would be:

$$\boldsymbol{B}^{*} = \boldsymbol{f_{correct}}\boldsymbol{B_{tr}} + (\boldsymbol{1} - \boldsymbol{f_{correct}})\boldsymbol{E} \qquad \text{Equation 4.3}$$

where $f_{correct}$ is the fraction of correctly-labeled bucket observations.

### 4.1 Implementation

In order to calculate biases in the data using Equation 4.1 described above, we need to assign values to each of the components. In the following subsections, we describe how the biases are estimated and how we refine some of the metadata assignments. Due to uncertainty in many of those values and the complicated interactions and correlations between them, we take an ensemble approach and generate many different sets of possible biases that span some part of the overall uncertainty in the adjustments. The process of estimating the biases is broken down into a number of steps:

1. Generate a set of correction fields for canvas and wooden buckets ($B_c$ and $B_w$) using a modified version of R06 (see Section 4.1.1) and SR02. These will form the basis for the bucket adjustments in the early part of the record.

2. Generate an ensemble of estimates of $E(t)$, the time-and-space-varying biases associated with ERI measurements, and $B^*$ (Equation 4.3), the bias associated with observations labeled as buckets (see Section 4.1.2). Note that the observations labeled as buckets will contain some unknown fraction of mislabeled ERI measurements. These are dealt with in step 4.
3. Estimate the spatially- and seasonally-varying biases for modern insulated buckets and generate an ensemble of bucket biases ($B_{tr}$) (see Section 4.1.3).
4. Using the values generated in steps 1-3, generate an ensemble of start and end dates for the transition from canvas to insulated buckets and estimates of the fraction of observations that are correctly identified as bucket measurements, $f_{correct}$ (see Section 4.1.4). The ensemble of estimates of $f_{correct}$ gives a measure of the uncertainty in the metadata assignments.
5. Generate an ensemble of estimates of how unknown measurement types are to be assigned (Section 4.1.5) and a separate ensemble of parameters for the Second World War, which reflects the fact that shipping and behaviors changed during the war and are somewhat independent from and more uncertain than the periods before and after (see Section 4.1.6).

We describe these steps in the following subsections and a schematic representation is shown in Figure 4. Comparisons with sub-surface profile data were used to inform or constrain some of the parameter ranges. In order to minimize the effect of over-fitting, particularly when data were few, the constrained parameters were not tied too closely to each other and were chosen to represent a reasonable range of uncertainty.

### 4.1.1 Generating wooden and canvas bucket correction fields, 1850-1941

The method described in K11b (in turn a modified version of R06) was the basis for generating an ensemble of 200 bucket correction fields for the data prior to 1942. A number of changes were made to the processing where uncertainties were likely to have been underestimated in the previous version.

First, we treated errors in the monthly correction fields as correlated: a single number was drawn from a standard normal distribution, multiplied by the uncertainty from R06 and applied to all monthly fields for a particular ensemble member. In K11b, a separate draw was made for each of the twelve months, effectively treating the errors as uncorrelated. However, in the original version of the method (R06) although they were independent draws and hence uncorrelated, the 95% uncertainty ranges were calculated separately for each month and then combined. This step effectively treated them as if the errors were fully correlated. This change brings the two methods back into line and increases the uncertainty in the bucket biases at annual and longer time scales relative to K11b.

Second, in the calculation of the difference in annual tropical average SST and NMAT – used to fix the fractions of wooden and canvas buckets from 1850 to 1920 – the estimated uncertainties arising from measurement errors were increased to account for the correlated micro-bias errors. We calculated the uncertainties assuming the errors were uncorrelated and then multiplied the resulting uncertainties by 5.08 (using the conversion factor for the tropics from K11a which accounts for the spatial correlation of the errors (= 2.2) multiplied by the square root of 12/2.25 for the temporal correlations). This again has the effect of increasing the uncertainty in the estimated bucket biases prior to 1920 relative to K11b and R06 where these errors were treated as if they were uncorrelated.

Third, we generated half of the 200-member ensemble as in R06, by assuming a linear transition from wooden to canvas buckets. The other 100 ensemble members were generated by assuming a step-change with different constant fractions of wooden and canvas buckets before 1906 and after 1910, with a linear change between these states from 1906 to 1910. The step change provides a qualitatively better fit to the noisy tropical SST-NMAT temperature differences used to estimate the transition (not shown). Rather than being a consequence of a step change in the fractions of canvas and wooden buckets, the step change in the SST-NMAT difference might instead be due to an increase in ship speeds around this time (Carella et al. 2017a). The change that we apply would be about the same in either case because an increase in the speed of the ship would also increase the necessary correction.

Fourth, every even ensemble member (2, 4 … 198, 200) was a blended average of R06 style adjustments and SR02 style adjustments (Equation 4.4). SR02 use patterns of SST-NMAT differences to estimate their adjustments. For our SR02-style adjustments, we estimated fields of adjustments associated with canvas buckets by taking the SR02 corrections for 1941. Wooden bucket corrections were calculated by assuming that the 1850 SR02 corrections were 80% wooden buckets and 20% canvas buckets. We estimated an uncertainty of 10% for the SR02-style bucket corrections i.e. the bucket corrections were multiplied by a number ($A_0$) drawn from a normal distribution with mean 1 and standard deviation of 0.1. For every even ensemble member, a weighted average of the R06 and SR02 adjustments was taken with the weight of the

R06-style correction, $A_1$ drawn from a uniform distribution in the interval [0,1] and the weight of the SR02-style correction equal to 1-$A_1$. In the extreme cases, the corrections look either entirely like R06, or entirely like SR02. For odd ensemble members, $A_1$ was set to 1. For each ensemble member a new value for $B_{cR06}$ was drawn and combined like so:

$$\boldsymbol{B_c = A_1 B_{cR06} + (1 - A_1) A_0 B_{cSR02}} \qquad \text{Equation 4.4}$$

The use of SR02 adjustments in addition to R06-style adjustments means it is possible to use a parametric framework to explore some of the structural uncertainties (Thorne et al. 2005) that would ordinarily only be accessible by comparing different data sets. The SR02 adjustments tend to follow the climatological pattern of sensible heat fluxes, whereas the R06 adjustments more closely follow latent heat fluxes. By using a weighted average of the two for half of the ensemble members, we explore different combinations of the two very different approaches and hence a wider spectrum of adjustments with different relationships to specific and latent heat fluxes. Time series of the resulting corrections are shown in Figure 5.

**4.1.2 Estimating biases for individual measurement methods, 1940 onwards**

We generated a subset of the data after 1930 using only measurements that we identified as ERI or that had a fractional assignment to ERI greater than or equal to 0.9. Error covariances were estimated as described in Section 3, Equation 3.6. The uncertainty associated with the large-scale correlated errors arising from using ERI measurements, $\sigma_G$, was set to be 0.2°C (the estimated mean bias from K11b). Monthly ERI biases were then estimated by comparison with drifting buoys and six different versions of the near-surface profile data for 1930-2014, from HadIOD.1.2.0.0, using the method described in the Appendix (Section A2). Figure 6 shows the global average of the combined large-scale systematic errors and micro-bias errors estimated in this way.

The left hand column of Figure 6 shows ERI biases estimated from comparisons with the six different data sets created from near-surface profile data for the globe, southern hemisphere and northern hemisphere.

ERI biases are generally positive apart from in the very earliest years when uncertainties are large. The largest reliably-estimated ERI biases occur between 1955 and 1970 when they range from 0.2 to 0.6°C. After 1970, ERI biases drop to a local minimum in the mid 1990s. They then rise again to a peak in the early 2000s before dropping again, approaching 0°C around 2018. The reasons for the variations in ERI bias with time are not clear, but the fleet of ships making ERI measurements was not designed to make climate quality measurements so the instability itself is unsurprising.

Measurements labeled as buckets and hull sensors were analysed in the same way. For buckets $\sigma_G(p)$ was set to the value of the bucket correction field at location $p$. The bias estimation is effectively a Bayesian regression using the bucket correction field as a predictor. For hull sensors, $\sigma_G$ was set to 0.2°C as it was for ERI measurements. Bucket biases (see left-hand column of Figure *6*) are rather variable but less so than in the ERI set. The bias is often positive, which is contrary to the general expectation that, on average, buckets lose heat (Kent et al. 2017). However, buckets can exhibit a warm bias during the day if the sun shines on the bucket and warms the water sample, and that might offset heat losses at other times (although in FP95, the solar effects in the correction fields were rather small). It is also likely that there is some contamination of the bucket subset by mislabeled ERI measurements. These issues are dealt with further in Section 4.1.3 and 4.1.4. Hull sensors show biases very similar to those of ERI measurements from 1990 onwards, so we combined these two data sources in the bias adjustments.

The ERI biases in each grid box were smoothed in time, but not in space. We generated an ensemble of 200 sets of adjustments. For each ensemble member, we randomly selected one of the six bias-adjusted sub-surface data sets to use as a basis, drew samples from the posterior covariance of the estimated ERI biases and added these to the mean biases (see section A2). The samples for each grid box were then smoothed in time using a LOWESS (LOcally WEighted Scatterplot Smoother) filter (http://flux.aos.wisc.edu/data/code/idl-lib/util/bueilib/lowess.pro) with a width of 121 months (±5 years). Because the ERI biases can only be estimated where there are known ERI measurements, there can be long temporal gaps in individual grid-box series. Where fewer than 30 months of data were available for a particular grid box in a 121-month period centred on a particular month, the grid box value for that month was set to the global average ERI bias. This procedure preserves the spatial structure seen in the ERI biases and their spatial covariances at the expense of some loss in temporal resolution. Spatial structure could arise for a number of reasons; for example, the biases of ships travelling along a common shipping lane could be quite different to the biases of ships travelling in the same region but not following the same routes.

### 4.1.3 Estimating parameters for modern bucket biases, 1970s onwards

Carella et al. (2018) used ICOADS flags SI (SST measurement method) and SIM (SST measurement method from WMO Pub. 47) to identify bucket measurements. They found that their compound flag, SI(M), was confirmed in 90% of cases from 1970 onwards. We selected bucket observations using SI, or SIM where SI was missing or ambiguous, and gridded these separately. This is similar to the assignment made in the previous section, but does not include fractional assignments. Biases were estimated for this data set relative to the sub-surface data from 1970-2005 using the method described in the Appendix (Section A2). For each calendar month, we computed fields of the mean bias and its standard deviation and the fields were filled using simple Kriging (Clark 1979) with a fixed angular length scale of 15° in latitude and longitude. 200 different versions of the adjustments were calculated by drawing samples from the posterior distribution of the Kriged solution.

The resulting fields show a small overall warm bias (Figure 5d) and there is a strong seasonal cycle in the northern hemisphere, which peaks in the summer months (not shown). This suggests that there is a solar heating bias in modern bucket measurements that exceeds the small heat losses found at night under moderate wind conditions by Kent and Kaplan (2006). The warm bias primarily affects the high latitude oceans (poleward of 40°N and S) and there is a small cold bias throughout the tropics. This pattern is similar to that seen by Carella et al. (2018) in which bucket measurements were biased warm relative to ERI measurements at high latitudes from the late 1980s onwards. The pattern identified by Carella et al. (2018) peaked towards local noon, but it is not clear whether the warm bias is due to the bucket sampling a shallow surface layer, or due to direct solar heating of the bucket itself. In either case it is distinct from buoy measurements which are the target for our analysis (Section 2.1)

### 4.1.4 Generating start and end dates for the transition from canvas to modern buckets and estimating the fraction of measurements misidentified as buckets

In the middle to late twentieth century there was a transition from the use of canvas buckets to the use of rubber buckets. This transition occurred at the same time as large numbers of ships adopted the ERI method. Neither transition is well documented and metadata, particularly that implied by deck and country information (Section 2.2.1), is not completely reliable. We can use the estimated biases from the previous sections (Section 4.1.1 to 4.1.3) together with proposed start and end dates for the canvas to rubber bucket transition to infer the fraction of incorrect metadata. Where this fraction takes an impossible value – outside the range [0,1] – we can reject that combination of start and end dates and thus narrow the uncertainty range for these parameters. Once we have narrowed down the range of start and end dates in this way, we can generate a best estimate of the time series of the fraction of correct metadata and a plausible range within which it can be varied to generate an ensemble.

We started with a wide range of dates for the transition from canvas to insulated buckets. Start dates were initially in the interval 1930 to 1960 and end dates between one year after the start date and 1980. The transition from canvas to insulated buckets was assumed to be linear between the start and end dates and the same everywhere.

Similarly to Equation 4.3, if we assume that $B*$, the bias estimated from observations labelled (perhaps incorrectly) as buckets, is a combination of biases associated with buckets, $B$, and biases associated with ERI, $E$, then we can write $B*$ as

$$B^{*} = f_{correct}B + (1 - f_{correct})E \qquad \text{Equation 4.5}$$

$$B = f_{canvas}B_{tc} + (1 - f_{canvas})B_{tr} \qquad \text{Equation 4.6}$$

where $f_{correct}$ is the unknown fraction of measurements labeled as bucket measurements that were correctly identified, $f_{canvas}$ is the fraction of canvas buckets changing linearly from one to zero between the start and end date of the transition. Rearranging, we get

$$f_{correct} = \frac{B^{*} - E}{B - E} \qquad \text{Equation 4.7}$$

With an estimate of $B$, $B^*$ and $E$ it is therefore possible to get an estimate of $f_{correct}$. $B_{tc}$ was estimated by subtracting the climatological-average ship bias from the mean bucket correction $B_c$ (Section 4.1.1) with the ship bias being calculated relative to each of the six different sub-surface data sets (Section 2.3) and then averaged across them. $B_{tr}$ was calculated as in Section 4.1.3.

For each pair of start and end dates the estimates of $B^*$, $E$ and $B$, for each month were used to derive an estimate of $f_{correct}$ for each month. Note that values of $f_{correct}$ slightly in excess of one are possible due to measurement errors in $B^*$, $E$ or $B$. A simple uncertainty range on $f_{correct}$ was estimated by increasing and decreasing the bias of rubber buckets by 0.05°C (a lower bound on the uncertainty of the method, see Appendix A3.3). If the uncertainty range in annual average $f_{correct}$ did not overlap the range [0.5, 1.0] during the transition period from 1955 to 1962, the start and end dates were rejected. In practice, no combination was consistent with a value less than 0.5. The period 1955 to 1962 was found to be particularly sensitive to the choice of start and end points (the spread in Figure 7a and b is particularly wide during this period) while also having reliable sub-surface data for estimating the biases. Prior to 1955, data coverage of the sub-surface data is much more sparse and therefore the uncertainties are larger and provide a much less useful constraint. After 1962, the spread is already well-constrained and nothing is gained by extending the constraint period beyond this point.

Figure 7 shows the accepted and rejected start and end dates for the transition and the associated time series of $f_{correct}$. Few start dates were rejected by this method, but end dates for the transition before 1961 were inconsistent with the data. This still gives a wide range of possible start and end dates, including those which describe a relatively rapid transition starting in the late 1950s.

The time series of $f_{correct}$ is broadly consistent with an independent assessment of the metadata made in Carella et al. (2018). Metadata are largely reliable from the early 1980s to the early 1990s. Another period of higher reliability is seen from around 1955 to 1962. Between 1965 and 1980, there is a period of less reliable metadata. Carella et al. (2018) argue for a relatively rapid transition from uninsulated to insulated buckets between the mid 1950s and mid 1960s, which is consistent with the results shown (Figure 7c). For example, such a transition starting in 1955 and ending in 1965 is accepted.

As with the ERI adjustments, a set of perturbed realisations of $f_{correct}$ were calculated that were similar to, but not tightly constrained by the best estimate. We calculated a time series of central values for $f_{correct}$ by averaging the time series of $f_{correct}$ across all allowed pairs of start and end dates (Figure 7d). Before 1952, the best estimate for $f_{correct}$ was set to 0.5 and after 1978 it was set to 0.95. This gives a continuous series that was then smoothed with a LOWESS filter with a width of 4 years. The one-sigma uncertainty was set to: 0.15 before 1952; 0.1 from 1952 to 1978; and 0.05 from 1978 onwards. This uncertainty range encompasses the majority of annual averages except in the post 1978 period. After 1978, there are some sharp variations, but these occur during a period when bucket measurements made a smaller and smaller contribution to the global average.

200 random time series with a lag-1 correlation of 0.99 and values in the range [-1,1] were generated, scaled by the estimated uncertainty in $f_{correct}$ and added to the best estimate. Values were capped at one with generated values above one set to one. Several example series are shown in Figure 7d.

### 4.1.5 Measurements with unknown method

Some measurements cannot be assigned a measurement method (Figure 2). A fraction of these unknown measurements was randomly reassigned to be either bucket or engine room measurements. A monthly time series was created which varied randomly between zero and one with an autocorrelation of 0.99 as in K11b. For a given month, the contribution of unknown measurements in each grid box was multiplied by this number and added to the contribution from ERI measurements. The remainder was added to the contribution from bucket measurements.

### 4.1.6 Parameters for the Second World War, 1941-1945

During the Second World War, there was widespread disruption to shipping. There are also discontinuities in ICOADS data sources and rapid changes in bias at the start of the war (R06, Huang et al. 2017) and at the end (Thompson et al. 2008). FP95 suggested that the Second World War saw the hasty (and even permanent) adoption of ERI measurements, as they were safer to make. The rapid changes seen in Figure 2 during this period are also partly due to our choices for the initial metadata assignment. In order to reflect the greater uncertainty during the war years, we generated a separate set of parameters with increased uncertainty for the period between January 1941 and August 1945.

The fraction of measurements that were labeled as buckets, but which were really ERI, was set to a value selected from a uniform distribution in the range [0, 1] to reflect the possibility that ERI measurements were temporarily taken in preference to bucket measurements across a large fraction of the fleet (Figure 7d). The ERI bias was chosen from a uniform distribution in the range [0, 0.5] and the fraction of unknown measurements set to buckets was drawn from a uniform distribution in the range [0, 0.25]. These values were chosen to give a broad range of possibilities – and hence large uncertainty in the adjustment – and reflect the likely prevalence of ERI measurements.

## 5 Results

Figure 8 shows the estimated biases in global and hemispheric averages for all data sources including ships and buoys. The biases shown in (Figure 8 a,c,and e) are relative to our reference buoy SST at a nominal depth of 20cm (Section 2.1). The bias relative to the average bias in the climatology period, 1961-1990, is also shown (Figure 8 b,d,and f). The difference between the two can be sizeable because of the large absolute biases in the climatology period when the vast majority of measurements were made by ships, which were affected by warm biases associated with ERI measurements.

From 1850 to 1939, the bias becomes increasingly negative (Figure 8 a,c,and e) reflecting the transition from wooden to canvas buckets. This is accompanied by an increase in the seasonal cycle of the biases, which is particularly clear in the Northern Hemisphere and arises from the seasonal drivers of the bucket biases: sensible, latent and solar heat fluxes (FP95). Between 1935 and 1939, the bias becomes slightly less negative as ERI measurements start to appear in ICOADS. Between 1939 and 1941, the bias becomes rapidly more positive as ERI measurements, principally from US sources, enter the record in large numbers (see Figure 6 of R06). From 1941 to 1945, during the Second World War, uncertainties are larger and there is a net positive bias that reflects the assumed increase in the prevalence of ERI measurements. The bias falls briefly in the post war years, reflecting a partial reversion to bucket measurements. However, the change is not so marked as it was in HadSST.3.1.1.0 because the ERI biases are here estimated to be larger than was previously assumed and the reliability of the initial method assignments (Section 2.2.1) is estimated, at times, to be worse than was assumed in HadSST.3.1.1.0. The bias increases to around 0.2 to 0.3°C between 1955 and 1970. During this period, ERI biases remain high and, towards the end, many observations that were initially flagged as buckets were reassigned to be ERI measurements. In HadSST.3.1.1.0, a constant fraction (30±10%) of measurements initially labelled as buckets were reassigned to be ERI with a different fraction in each ensemble member. In HadSST.4.0.0.0, as well as being different for each ensemble member, the reassignment is time-varying and is constrained based on the biases estimated from comparisons with the sub-surface oceanographic profiles (Figure 7, Section 4.1.4).

From 1970 to 1980, the average bias drops reflecting a reduction in ERI bias (see also Figure 6). From around 1990, buoys start to have a significant effect on the record. This reduces both the bias in the combined data set and the uncertainty in the bias. Despite the increasing prevalence of buoys, changes in bias from 1990 to the present continue to reflect changes in the bias of the ERI and hull sensor measurements, which are the predominant means by which SST measurements were made from ships in this period and which still have a weight of around 50% in the global average (Figure 2).

The fall in bias in the ship data since the early 2000s is not confined to ERI measurements and is consistent with the change in bias estimated by Huang et al. (2015, 2017) using comparisons with NMAT and buoy data. It is seen in VOSClim ships (http://sot.jcommops.org/vos/vos.html), which are a select subset of the full VOS fleet for which better quality metadata are available. It is also seen in measurements from ships that are not part of VOSClim, and in both bucket measurements and hull sensor measurements, which suggests that this is a pervasive reduction in the overall bias in the ship data from the early 2000s to present.

The bias relative to 1961-1990 has a similar evolution (Figure 8 b,d,and f). However, there are two things to note. First, as the absolute bias shows, there is a large warm bias in the data from 1961-1990 and the uncertainty is larger than it is for the later data. When biases in the buoy-rich period from 1990 onwards are expressed relative to the 1961-1990 average, their uncertainty increases. This happens because we are comparing an accurate measurement to an inaccurate baseline and the overall uncertainty in the difference is therefore high. Second, the continued use of buckets in the climatology period means that there is a residual seasonal cycle in the absolute biases. When those biases are expressed relative to 1961-1990, the seasonal cycle is reduced during the climatology period and in the pre-Second World War period, but increased outside of it.

Figure 9 shows series of global and regional average SST anomalies for HadSST.4.0.0.0 and for the unadjusted gridded measurements. The adjustments have the largest effect prior to the Second World War. However, even after the war, the biases affect the long-term changes in the data. The apparent step change in the mid-1970s is slightly larger in the adjusted data, with the adjustments cooling the series in the late 1960s and warming it in the late 1970s. The difference between the 2010s and the 1961-1990 period is also increased. This is caused by the overall decrease in the bias over the past 50 years, which is principally due to the decreasing ERI bias and the increasing influence of unbiased drifting buoys. There is a strong annual cycle in the northern hemisphere average from around 2005 onwards. The origin of this is not clear, but may be due in part to reduced sea-ice extent during summer months.

In the following subsections, we compare the new HadSST.4.0.0.0 data set and bias adjustments with other marine temperature (SST and MAT) data sets to identify points of similarity and difference. In Section 5.1 we compare subsets of data made using buckets and engine room measurements and show that the adjustments reduce the relative bias between them. Section 5.2 details comparisons with instrumentally homogeneous data records from 1991 to 2018, demonstrating the stability of the HadSST.4.0.0.0 record in the modern era. Comparisons with longer MAT and near-surface oceanographic profile measurements series are described in Section 5.3 and highlight an interesting discrepancy with HadNMAT.2.0.1.0. In Section 5.4 we compare HadSST.4.0.0.0 to HadSST.3.1.1.0 and in Section 5.5 we extend the comparison to other SST data sets showing good long term agreement between them.

## 5.1 Internal consistency

Two data sets were created. One contained only observations identified as likely bucket measurements. The other contained only observations identified as likely ERI measurements. The two data sets were gridded separately and their measurement and sampling uncertainties estimated. Each of the two subsets of the data was bias adjusted using the relevant estimated biases. Figure 10 shows the global average anomaly for the two collocated data sets before and after adjustment.

Prior to adjustment, there is a clear time-varying offset between the two data sets between 1940, the first year in which ERI measurements are present in significant numbers, and 2014. This is consistent with various estimates (see K11b) of the relative biases between the two measurement methods. There is similar inter-annual variability in the two data sets. The drop in the combined, unadjusted SST anomalies in 1945 (Figure 8 and Figure 9) is not evident in either of the two subsets. This suggests that the abrupt drop in estimated global temperature highlighted by Thompson et al. (2008) is largely a result of a rapid change in the relative proportions of ERI and bucket measurements in ICOADS at that point. On the other hand, the two subsets are each noisier than the combined series and the discontinuity occurs at the point at which the estimated uncertainty changes markedly so further analysis is warranted.

The adjustments improve the agreement between the two data sets. This indicates that the bias adjustments are functioning as expected at a hemispheric scale. The agreement is expected because adjustments to both strands of data and the parameter choices are set to ensure loose consistency with the same reference data sets. The divergence post-2005 is likely due to the scarcity of bucket data after this point, which is also reflected in the much larger estimated uncertainties. The uncertainties in the adjusted data are also broader than in the unadjusted data. This may seem counterintuitive and arises because bias errors are not included in the uncertainty range for the unadjusted data. If they were, they might amount to a few tenths of a degree reflecting the large differences between ERI and bucket data.

### 5.2 Comparison with instrumentally homogeneous data sets

In the modern period, the data set can be compared to three instrumentally homogeneous data sets (Argo, ARC and buoys, see Section 2.2) that combine higher accuracy with good global coverage.  Hausfather et al. (2017) used similar instrumentally homogeneous data sets to assess the effect of bias adjustments on trends in the modern period and detect drifts in the global averages of combined series such as HadSST3 and ERSSTv3/v4 that were smaller than 0.05K/decade. Our comparison builds upon the Hausfather et al. (2017) analysis in two ways. First, all the data sets are gridded using a common procedure (Section 3), which minimizes the possibility of differences due to processing. Second, the same 1961-1990 climatology is used to calculate anomalies for all of the data sets. Using a common climatology, rather than force each series to average to zero across the recent period of overlap, means that any absolute differences in SST will also be highlighted along with any trend differences.

We constructed gridded data sets of the three comparison data sources using the method described in Section 3. Figure *11* shows global averages of the three data sets compared to

HadSST.4.0.0.0 at the locations where there is common data coverage. The agreement between these data sets is very good with differences largely falling within the uncertainty ranges.

The least interesting comparison is between HadSST.4.0.0.0 and the buoys (Section 2.5.3) as they are used in the construction of the HadSST.4.0.0.0 data set. Nevertheless, it is useful to compare the two to ensure that we have not introduced a bias by the addition of the ship data. It is clear that the adjustments greatly reduce the difference between the combined data set and the buoys alone. There is, however, a period in 1992/1993 where the two diverge by much more than the estimated uncertainty. The early 1990s are rather sparsely observed by drifting buoys and, in this case, the buoy record is likely to be erroneous with local large deviations associated with poor-quality drifting buoy data. Spatially (Figure *12*(a)), we generally see small average differences (within ±0.1°C for most areas) except in coastal regions around the US and in the northern hemisphere western boundary currents.

A more stringent test is to compare the HadSST.4.0.0.0 data set with the independent ARC satellite retrievals (Section 2.5.2). The adjustments bring the *in situ* SSTs more closely in to line with the ARC SSTs. There are some discrepancies here, most notably in the early 1990s. During this period, the ARC SSTs are based on data from the ATSR1 instrument (1991-1995), which is less well understood than its successors, ATSR2 (1995-2003) and AATSR (2002-2012). Differences are larger at an individual ocean basin level and locally (Figure *12*(b)). HadSST.4.0.0.0 is warmer than ARC south of 40°S and along the climatological ice edge in the Arctic. HadSST.4.0.0.0 is cooler than ARC through much of the tropics, particularly in the Indian Ocean and Maritime Continent, but differences are typically less than 0.2°C and very widely less than 0.1°C. Future versions of HadSST may benefit from using more highly-resolved adjustments. Not shown in Figure *11* is the notional 0.1°C uncertainty associated with the large-scale correlated errors in the ATSR series that arise from uncertainty in the retrieval process. It should also be noted that local differences of order 0.1-0.2°C remain between ATSR and collocated buoy measurements (Merchant et al. 2012), so differences seen here are likely to be a combination of errors in ARC and errors in HadSST.4.0.0.0.

The comparison with Argo shows that the adjustments continue to be effective from 2012 through 2017 after the failure of AATSR (Figure *11*). Indeed, agreement with the independent Argo data is excellent back to 2007 when the Argo array reached its design coverage. Even before this, when coverage was less than global, the agreement remains good, suggesting that the bias adjustments are reliable at smaller scales. The map of differences between Argo and HadSST.4.0.0.0 is noisier than for the comparison with ARC due to the sparser sampling of Argo (Figure *12*(c)). Some patterns are perhaps common to the two – the cooler Indian Ocean, for example – but  the large differences at high latitudes are not seen in the comparison with Argo suggesting that this is due either to problems with the ARC data or is related to sampling errors of some kind.

Although agreement with the instrumentally homogeneous series is good overall, there are some months when the discrepancies exceed the estimated uncertainties. This is to be expected from time to time as the uncertainty range represents a 95% confidence interval. The discrepancies in the comparisons suggest an overall 1-sigma uncertainty in SST changes seen through this period of around 0.05°C, which corresponds to a stability of a few hundredths of a degree per decade. In contrast, the difference between the instrumentally homogeneous series and the unadjusted data approaches a maximum of 0.2°C (Figure *11*(c)) in the global average (and locally more Figure *12*(d)) over the same period highlighting the importance and effectiveness of the adjustments in the modern period. The average difference between HadSST.4.0.0.0 and the instrumentally homogeneous data sets is much smaller than the applied adjustments.

### 5.3 Comparison with other long time series

Over longer periods, it is necessary to use other data sets for comparison. We use two data sets here. The first is HadNMAT.2.0.1.0 (Kent et al. 2013) which is a data set of Nighttime Marine Air Temperatures (NMAT). Anomalies in NMAT are thought to closely track anomalies in SST over long periods and large scales (see Huang et al. 2015 for an example using climate models). The second is based on oceanographic profiles from HadIOD.1.2.0.0, excluding Argo (Atkinson et al. 2014) and adjusted using the Levitus et al. (2009) adjustments for MBTs and XBTs.

In order to make a direct comparison between HadSST.4.0.0.0 and HadIOD.1.2.0.0, anomalies from HadSST.4.0.0.0 were adjusted using the absolute bias rather than the relative bias so that the SSTs could be directly compared. Anomalies were then calculated for both data sets using an unadjusted climatology. However, HadSST.4.0.0.0 is not then directly comparable to HadNMAT.2.0.1.0 as HadNMAT.2.0.1.0 is provided as actuals or relative to its own adjusted 1961-1990 climatology and not relative to a biased SST climatology. This will lead to a constant annual offset between the SST and NMAT series, which is approximately the size of the average climatological bias in the SST. Consequently, we shifted HadNMAT.2.0.1.0 by 0.15°C in Figure

13. The offset was chosen by eye to approximately align the two series; none of the conclusions depend on the choice of offset.

Except for a period in the late 1940s and early 1950s, differences between HadNMAT.2.0.1.0 and HadSST.4.0.0.0 anomalies (Figure 13b) on a decadal time scale are constant between 1920 and 1990. Outside this period, the differences exceed the estimated uncertainties in the HadSST.4.0.0.0 data set. One notable difference occurs around 1991-1993, when HadNMAT.2.0.1.0 apparently cools relative to HadSST.4.0.0.0 (or HadSST.4.0.0.0 warms). Further investigation shows that the cooling occurs in the tropics, partly offset by warming in the northern extratropical Pacific.

Christy et al. (2001) previously remarked on the cooling of air temperature relative to SST in the tropics. The cause of these differential rates of warming is unknown. They hypothesized that a large scale change in circulation might have caused a persistent change in air-sea temperature differences but could not rule out the effects of biases in either the SST or NMAT data sets which, at the time, had not been studied in detail. Both HadNMAT.2.0.1.0 and HadSST.4.0.0.0 are now bias adjusted and are independent of one another in the relevant period. This suggests that the differences represent a real change in the air-sea temperature difference across this transition.

However, other hypotheses could explain the change. There could still be an undetected bias in either the SST data, the NMAT data or both. For example, automation during the early 1990s might have allowed air temperature sensors to be placed in better-exposed locations with a consequent drop in the measured air temperature. This change would likely have been accompanied by a move to electric sensors. It could also be that the near-surface oceanographic profile data that we use as a basis for estimating the biases in ERI and modern bucket measurements change in the early 1990s in a way that is not captured by the corrections to the data. Huang et al. (2018) found a change in the average depth of profile measurements in the near-surface layer, but their criteria for selecting the profiles were different from those used here and the step change in the NMAT-SST difference is there even when unadjusted SSTs are used.

### 5.4 Comparison with HadSST.3.1.1.0

Figure 14 shows a comparison between HadSST.4.0.0.0 and HadSST.3.1.1.0. The overall evolution of these is similar, although differences at some times are larger than the estimated uncertainties. HadSST4 runs colder than HadSST3 in the period following the Second World War to 1970. From the late 1970s to the early 2000s, HadSST4 is warmer. This change from cooler to warmer, leads to a slight sharpening of an apparent step change in global average SST around 1975. This is most distinct in the Northern Hemisphere with the transition being somewhat smoother in the Southern Hemisphere. The differences between the data sets in this period are due to two factors. First, the ERI biases are now estimated from the data (Section 4.1.2) and they are larger in the 1960s in HadSST4 than they were assumed to be in HadSST3. The ERI biases are also outside the uncertainty range for ERI biases (0.2±0.1°C) used in HadSST3. Second, the fraction of cooler bucket measurements is lower overall in HadSST4 than in HadSST3. This is largely due to the new method of inferring the fraction of incorrectly assigned metadata (Section 4.1.4).

The estimated uncertainties in the global and hemispheric averages are for the most part larger in HadSST4 than HadSST3 prior to around 1970. This is due to the wider range allowed for ERI biases (Section 4.1.2) and to changes made to the bucket corrections (Section 4.1.1).

### 5.5 Comparison with other SST data sets

The latest versions of ERSST, COBE-SST and HadSST all now apply adjustments to the whole SST record. Figure 15 and Figure 16 show global and regional averages from HadSST.4.0.0.0, COBE-SST-2 and ERSSTv5 (with the ensemble from ERSSTv4) calculated where the data sets have common coverage. The overall evolution of the three data sets and the interannual variability in each are very similar.

The adjustments applied in each of the three data sets decrease the overall temperature change seen from the nineteenth century (and especially since 1900) relative to the unadjusted data. Of the three data sets, HadSST4 has a marginally higher trend from 1900 (estimated using ordinary least squares) but the difference between the trends in the three data sets is not larger than the estimated uncertainty (estimated using the ensemble with each ensemble member additionally perturbed by a sample from the measurement and sampling errors). ERSST and COBE-SST warm at a similar rate to the unadjusted data from the 1940s, 50s and 60s, but HadSST4 warms somewhat faster than the other data sets due to the adjustments applied to account for the general decline in ERI biases over that period (Figure *6*). From start dates in 1970, 1980 and 1990, COBE-SST-2 warms faster than the unadjusted data and, from 1980 and 1990, faster than either HadSST4 or ERSSTv4 by a significant margin.

From 2000-2012, the rates of warming in all three data sets are very similar and consistent within their uncertainty ranges. All three warm faster than the unadjusted data, which has a trend close to zero. During this period, there are two important factors. First, there is a large increase in the relatively cooler drifting buoy measurements and, second, there is a decrease in the average ship bias. The analysis of HadSST4 supports ERSSTv4 and ERSSTv5 in this period (Karl et al. 2015) and is consistent with instrumentally homogeneous reference series, supporting the analysis of Hausfather et al. (2017).

Kent et al. (2017) showed that there were significant differences between HadSST.3.1.1.0 and ERSSTv4 at other times. The period of the largest global differences was found to be during 1945-1970 when HadSST3 was warmer than ERSSTv4. HadSST4 is much closer to ERSSTv4 during this period. This change is due to the new ERI bias estimates being larger than assumed in HadSST3 during this period. However, from 1960 or 1970, HadSST4 warms faster than ERSSTv4. The long-term bias adjustments in ERSSTv4 are derived from assuming a constant relationship with HadNMAT2 that, as we have already shown in Section 5.3, warms less than HadSST4 over this period with much of the difference arising from a step-like change of unknown origin in the early 1990s.

## 6 Summary

In this paper, we have estimated biases associated with different methods for making SST measurements by comparison to near-surface oceanographic measurements and buoys. The estimated biases were combined with other metadata to bias adjust a composite SST data set. Because many of the parameters in the bias adjustment scheme are uncertain and give rise to complicated covariance structures, we present the data set as an ensemble in which we vary uncertain parameters to understand their impact on the indicators that can be derived from the data, such as the global average, or changes in temperatures.

The method builds on that used to create HadSST.3.1.1.0. We now have improved estimates of the biases associated with different measurement methods – including engine room measurements and insulated buckets – and we are better able to constrain poorly known parameters such as the timing of the transition from canvas to rubber buckets and the fraction of incorrect metadata on measurement method. Some of the newly constrained parameters, particularly biases associated with engine room measurements, were outside the previously estimated ranges at some times. This highlights the difficulties, expounded at greater length in Kent et al. (2017), of working with historical meteorological data, particularly when trying to make data sets that are useful for climate research. Nonetheless, by paying careful attention to the data, quantifying the biases and estimating the uncertainties, we can produced a climate data record of SST back to 1850 that is consistent with independent information.

The method relies on comparisons with sub-surface data. This raises two possible difficulties. First, the depth of near-surface subsurface measurements is usually slightly greater than the depth at which drifting buoys make measurements and may have changed systematically over time (Huang et al. 2018). This could lead to a cool bias in the earlier data where sub-surface measurements are used to estimate biases. On the other hand, there is no clear signal that this is the case where we can compare ship measurements to both buoys and sub-surface measurements. Second, sub-surface temperature measurements also exhibit biases. While these biases are expected to be smaller than those in SST measurements from ships, they are nonetheless significant on longer time- and space- scales. Adjustments for these biases are themselves uncertain and an active area of research (Abraham et al. 2013, Cheng et al. 2016).

The small adjustments that Huang et al (2015, 2017) applied to ERSSTv4/v5 in the post war years were somewhat puzzling because they suggested a small net bias during a period that saw a transition from canvas to insulated buckets and from widespread bucket use to widespread ERI use, factors that led to the larger adjustments applied to HadSST.3.1.1.0. The work we present here is in closer agreement with ERSST and suggests that the smaller net corrections are due to a greater prevalence of ERI measurements – supporting the conclusions of Carella et al. (2018) – partly offsetting larger biases associated with uninsulated buckets and an earlier change to insulated bucket use.

Important uncertainties likely remain. In the unadjusted data, a rapid drop in global average SST marks the end of the Second World War (Thompson et al. 2008). The drop is seen in both ERSSTv5 and HadSST.3.1.1.0, though it is less marked in the latter. It coincides with a large change in the areas sampled by the global fleet, which likely explains some of the fall. The question of how much of the remainder is artificial is still open. The separated bucket and engine room data sets considered in Section 5.1 suggest that some of the drop arises from a change in the mix of the two rather than a globally coordinated drop in actual SST. However, it is important to note that the sampling of both these data sets changes at this point. The drop is most pronounced in ERSSTv5 and coincides with a similar rapid drop seen in HadNMAT.2.0.1 at that point. However, comparisons between HadNMAT.2.0.1.0 and CRUTEM4 at common coastal grid cells (Cowtan et al. 2018) suggests that the NMAT, and hence ERSST, is artificially warm during the war years despite the adjustments that have been applied for non-standard exposure (Kent et al. 2013). These lines of evidence suggest that at least some of the drop is artificial, but they do not help to understand which of the data sets provides a better estimate. Consequently, considerable uncertainty remains regarding SST during the Second World War. This uncertainty is partly reflected in the wide uncertainty ranges given in HadSST.4.0.0.0, but a more satisfactory solution is needed. Users of the data set should be wary of drawing strong conclusions based on trends that start or end during the war years until this is resolved.

From 2000 to 2012, the period studied in detail here, ERSSTv5 and HadSST.4.0.0.0 have trends that are consistent with each other and with COBE-SST-2. In addition, both HadSST.4.0.0.0 and ERSSTv5 compare well with independent and instrumentally homogeneous data over the period 1991 to 2017 (see also Hausfather et al. 2017). We highlight the importance of changing ship biases as well as the shifting balance of ship and buoy measurements for understanding this period.

Although HadSST.4.0.0.0 and ERSSTv5 show reasonable agreement in the overall evolution of global average SST, there are some interesting differences between the trends estimate from these data sets. In particular, warming since the 1950-1970 period is higher in HadSST.4.0.0.0. This is associated with a cooling of ERSSTv5 relative to HadSST.4.0.0.0 in the early 1990s. This discrepancy is also seen in a comparison with HadNMAT.2.0.1.0, the data set used to adjust ERSSTv5. The discrepancy between HadNMAT.2.0.1.0 and HadSST.4.0.0.0 suggests that there is either a large-scale change in atmospheric circulation in the early 1990s that modified the air-sea temperature difference throughout the tropics or that undetected biases remain in one or the other of the marine temperature (SST or NMAT) data sets considered here.

Huang et al. (2015) showed that NMAT-SST differences exhibited little variability at annual time scales in a climate model between 60°S and 60°N, although there was a long-term warming of MAT relative to SST of around 0.1°C. Hawkins et al. (2015) likewise found that MAT warmed faster globally than SST in a range of climate models. Neither paper shows specific step-change behaviour, but both show spikes in the mean model response of SST-MAT following large tropical volcanic eruptions such as Mount Pinatubo in 1991. While the step change is in the same direction as this model response, the subsequent recovery and long-term warming of MAT is not apparent in Figure 13. Hawkins et al. (2015) note that the size of the SST-MAT differences are comparable to the uncertainties in the SST and MAT data sets used, HadSST3 and HadNMAT2 in their case, but it is also true for HadSST.4.0.0.0.

Because of the strong links between SST and MAT and between systematic errors in SST and MAT, a fuller understanding of marine temperatures in general can only be achieved by studying both in greater detail along with metadata and other relevant marine variables such as humidity (Willett et al. 2008) and winds. While measurements of SST are now more numerous than ever thanks to the wealth of satellite data and autonomous platforms such as drifting buoys, there has been a marked continuing decline in the MAT observing system which relies on ship-borne instruments and is currently far below the level of adequacy as judged by a number of criteria (Berry and Kent 2017).

Recently attention has been drawn to spatial as well as temporal heterogeneity in SST biases and how these affect the interpretation of climate variability (Huang et al. 2013). Although the methods we describe account for the spatial heterogeneity that arises from geographically-varying ERI biases and changing numbers of buckets and ERI measurements, some of the factors in the adjustments – for example, the scaling for the patterns of bucket biases and how measurements without a method are assigned – are only specified at a global level and thus might not be as effective at a basin scale.

One factor that might vary locally is the type of bucket used for measurement. Buckets issued by different countries are of varied design and the design can affect the rate of heat loss (Carella et al. 2017b) as well as other properties of the measurements (Kent and Taylor 2006). This might be of particular importance during the early and middle decades of the 20$^{th}$ century when bucket use was widespread. The period 1900-1940 saw an overall increase in shipping, large changes in ship routes allowed by the opening of the Panama Canal, two World Wars, and large biases associated with uninsulated buckets deployed from fast ships. The results for modern bucket biases (Section 4.1.3) suggest a potentially important role for solar heating of the bucket and water sample. In contrast, Kent and Kaplan (2006) and Carella et al. (2017b) focused on situations in which solar heating was negligible. Recently, Chan and Huybers (2019) showed that there are relative biases between bucket measurements made by ships from different countries and between bucket measurements found in different decks in ICOADS. They argue that correcting for biases between decks and nations should improve estimated SSTs.

After a long analysis and discussion of the problems with the data, it can be easy to forget the enormous value that the voluntary observing ships (VOS) provide. This would be a mistake. For much of the historical record, reports from ships are all we have and, although it is the outliers of the distribution which often draw our attention, most ships have provided useful, reliable measurements. At those times where we have used other sources to correct the ship data (oceanographic profile measurements from the 1950s to present and drifting buoy measurements from the early 1990s) the ship data provide vital spatial detail in the large areas not covered by these measurements (Figure *18* and Section A2). Even in the modern period, when the coverage of drifting and moored buoy data is quasi-global, the density of shipping, particularly in the northern hemisphere, adds additional useful information (see e.g. Figure 2 and Figure 3). As researchers extend climate data records of SST derived from satellites further back in time, they will need to rely on ships to provide a “ground truth” against which their products can be tested. Last of all, ships also measure variables other than SST – air temperature (Kent et al. 2013, Berry et al. 2004), humidity (Willett et al. 2008), cloud, pressure and wind (Berry and Kent 2011) – that are essential for understanding the continual fluxes of heat and water between the atmosphere and oceans.

Challenges remain for building models – be they statistical or physical – which can adequately describe and constrain the spectrum or hierarchy, of errors that exist in *in situ* marine measurements of sea-surface temperature and air temperature. The methods detailed here – which can in some cases extract useful information about the error characteristics of individual ships – could be extended to include more detailed error models which track weather-dependent biases associated with ships from a particular country or which use a certain kind of bucket.

Finally, we reiterate the recommendations made in Kent et al. (2017), "A Call for New Approaches to Quantifying Biases in Observations of Sea Surface Temperature", in particular the need to:

- add more data and metadata to ICOADS;
- reprocess existing ICOADS records;
- improve information on observational methods;
- improve physical and statistical models of SST bias;
- maintain and extend the range of different estimates of SST bias; and
- expand data sources for validation and extend the use of measures of internal consistency in validation.

## Data availability

The HadSST.4.0.0.0 data set, and supporting information, is available from http://www.metoffice.gov.uk/hadobs/hadsst4. The following listed data sets were used in this analysis. Links to the data sets are provided where applicable.

International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 3, Individual Observations. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/D6ZS2TR3. (Research Data Archive et al. 2016)

Drifting buoy data were collected and made freely available by the Copernicus project and the programs that contribute to it. Data downloaded (6 April 2018) from http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=INSITU_GLO_NRT_OBSERVATIONS_013_030

HadNMAT.2.0.1.0 was downloaded from https://www.metoffice.gov.uk/hadnmat2

HadIOD is available from the corresponding author of Atkinson et al. (2014)

The operational version of ERSSTv4 was downloaded from http://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v4/netcdf/

The ERSSVTv4 ensemble was downloaded (16 June 2015) from https://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v4/ensemble/

The operational version of ERSSTv5 was downloaded (9 February 2018) from https://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v5/netcdf/

COBE SST 2 was provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/ (https://www.esrl.noaa.gov/psd/data/gridded/data.cobe2.html, accessed 23 October 2014)

ARC data were downloaded (5 July 2017) from http://catalogue.ceda.ac.uk/uuid/e6497acddf9cd8345ffbd0643c0d9729

**Acknowledgments**

The authors were supported by the Joint UK BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101). Thanks to Boyin Huang of NOAA NCEI for making ERSSTv5 and the ERSSTv4 ensembles available and Shoji Hirahara of JMA for providing the COBE-SST-2 data set. This research would not have been possible without the vital ongoing work of the ICOADS team. This study has been conducted using E.U. Copernicus Marine Service Information. Many thanks to the reviewers for their helpful comments.

# Appendices

## A1 Interpolation

In order to estimate biases for different measurement types during the modern period, we used a simple statistical interpolation scheme based on Gaussian Processes (Rasmussen and William, 2006). The scheme is described in this section. It requires as input the error covariances calculated in Section 3.2. In addition, an estimate of the covariance matrix of the actual SST fields is needed. The method for estimating this matrix is described in Section A1.1. The method is applied in Section A2 and we present some tests of the method in Section A3.

The variability of gridded SST anomalies was modelled as a multivariate normal distribution with mean zero and covariance matrix *C*. For a vector of observations *y* with error covariance *R* (see Section 3.1 and 3.2), a globally complete reconstruction of the SST anomaly field, $\mu$, can be obtained using

$$\mu = CH^T(HCH^T + R)^{-1}y \qquad \text{Equation A1}$$

where *H* is a matrix consisting of zeroes and ones that selects points from *C* at the measured locations in *y*. The posterior distribution for the reconstruction is a multivariate normal distribution with mean, $\mu$, and covariance *P*

$$P = C - CH^T(HCH^T + R)^{-1}HC \qquad \text{Equation A2}$$

Equations A1 and A2 are equivalent to Equations 2.23 and 2.24 from Rasmussen and Williams (2006) but with a non-diagonal error covariance. As well as producing a reconstruction of the SST field, this framework can also be used to get an improved estimate of the errors in the data.

For example, considering $R$ as the sum of several error components $R_1$, $R_2$ … $R_n$ then an improved estimate of the component $R_1$ is given by

$$\mu_{R_1} = R_1(HCH^T + R)^{-1}y \qquad \text{Equation A3}$$

with posterior covariance

$$P_{R_1} = R_1 - R_1(HCH^T + R)^{-1}R_1 \qquad \text{Equation A4}$$

This breakdown of the estimated errors into individual components is used in Section A3 to test that the interpolation method is working as expected.

**A1.1 Estimating the prior covariance**

In order to get a reasonable interpolation, a good estimate of $C$ is needed. Here we used a time-invariant estimate of $C$ calculated from monthly data. $C$ was built from a set of local covariance kernels, in which the covariance at each point was modeled as a simple local covariance kernel. The method is based on that used in Karspeck et al. (2012), assuming an exponential kernel and that the length scales are strictly zonal and meridional. The covariance between two points is equal to

$$C(x,x') = \sigma\sigma' \frac{|\Sigma|^{\frac{1}{4}}|\Sigma'|^{\frac{1}{4}}}{\left|\overline{\Sigma}\right|^{\frac{1}{2}}} exp(-\overline{\tau}) \qquad \text{Equation A5}$$

Where unprimed and primed variables indicate the values at the two different points and τ, the Mahalanobis distance, can be written as:

$$\overline{\tau}(x,x') = \sqrt{(x-x')^T\overline{\Sigma}^{-1}(x-x')} \qquad \text{Equation A6}$$

and,

$$\Sigma = \begin{bmatrix} L_x^2 & 0 \\ 0 & L_y^2 \end{bmatrix};\ \bar{\Sigma} = \frac{1}{2}(\Sigma + \Sigma') \qquad \text{Equation A7}$$

Here $x$ and $y$ are the angular separations in longitude and latitude, $L_x$ and $L_y$ are the length scales. The length scales and variance of the process can vary from point to point. The primes indicate the values of $\Sigma$ and $\sigma$ at location $x'$ and the unprimed variables indicate the values at $x$.

The variance was estimated first using ARC data for each grid cell by assuming that the anomalies in a grid cell were normally distributed with mean zero and variance equal to the sum of the variance of the process and the error variance (provided with the ARC data). The value of $\sigma^2$ that maximized the likelihood of the data given the parameters was found using a simple line search with increments of $0.005$ K$^2$. Values of the standard deviation above 1.2°C were set to 1.2°C.

The length scales were then estimated for each grid box separately. $x$ and $y$ distances were measured as angular distances in latitude and longitude multiplied by 6400km with the implicit assumption that the Earth is a cylinder. This is more numerically stable than assuming a spherical Earth and the geographically-varying length scales allow for the geometrical fact that one degree of longitude corresponds to different lengths at different latitudes as well as changes in the physical length scale. Time series of anomaly differences were calculated between the target grid box and all neighbors within 10 000km and detrended using a 5$^{th}$ order polynomial. The true SST anomalies at the two locations were assumed to be normally distributed (mean zero and standard deviations $\sigma$ and $\sigma'$) and correlated with each other with a correlation of $\exp(-\tau)$. The uncertainty on the anomalies were $\sigma_{error}$ and $\sigma'_{error}$. The differences between the two series are then normally distributed with mean zero and variance $\sigma_{diff}$ given by

$$\sigma_{diff}^2 = \sigma_{error}^2 + \sigma'^2_{error} + \sigma^2 + \sigma'^2 - 2\sigma\sigma' \exp(-\tau) \qquad \text{Equation A8}$$

The values of $L_x$ and $L_y$, which maximized the likelihood of the data given the parameters were found using the using the downhill simplex method of Nelder and Mead (1965) (as implemented in the IDL 8.2 AMOEBA function). All other values – the error variances ($\sigma_{error}$) and the variance of the process (σ and σ') – were fixed from earlier calculations. Missing values of $L_x$ were set to 2500km. Missing values of $L_y$ were set to 1000 km. The resulting fields of σ, $L_x$ and $L_y$ (Figure 17) were then stitched together to produce a single covariance matrix using equation A5.

The covariances vary in character from place to place (see Figure 17). In the Tropical Pacific, zonal length scales are long and meridional scales are limited. Variances are also higher in the tropical Pacific associated with ENSO variability. In the North Pacific, length scales are shorter in general. In the North Atlantic and Indian Ocean, the covariances are more isotropic with similar zonal and meridional length scales. Over the western boundary currents, the length scales are short and variability is high.

### A2 Estimating biases

Engine Room biases and other measurement method biases were estimated using the simple interpolation scheme in two or three steps. In the first step, gridded (see Section 3) drifting buoy observations (if they were available) were interpolated using the formulas (Equations A9 and A10) below to get an improved estimate of the global SST field. The mean and covariance of the posterior distribution were used as the prior for the second step, in which near-surface sub-surface measurements (when these were available) were interpolated using the output from the drifting-buoy interpolation as input (Equations A11 and A12). In the third step, gridded Engine Room measurements (a similar thing can be done for other measurement types) were interpolated using the output from the previous interpolation (Equations A13 and A14). In contrast to the first and second, the aim in the third step was to estimate the correlated errors in the gridded ERI measurements rather than to get an improved estimate of the SST field (although this can be obtained as well).

First the buoy data are assimilated.

$$\mu_{buoy} = C_{ARC} H^T \left( H C_{ARC} H^T + R_{buoy} \right)^{-1} y_{buoy} \qquad \text{Equation A9}$$

$$C_{buoy} = C_{ARC} - C_{ARC} H^T \left( H C_{ARC} H^T + R_{buoy} \right)^{-1} H C_{ARC} \qquad \text{Equation A10}$$

Where $C_{ARC}$ is the prior covariance calculated in A1.1 using the ARC data. Then the sub-surface data

$$\mu_{sub} = C_{buoy} H^T \left( H C_{buoy} H^T + R_{sub} \right)^{-1} \left( y_{sub} - \mu_{buoy} \right) + \mu_{buoy} \qquad \text{Equation A11}$$

$$C_{sub} = C_{buoy} - C_{buoy} H^T \left( H C_{buoy} H^T + R_{sub} \right)^{-1} H C_{buoy} \qquad \text{Equation A12}$$

Finally, the bias in the ERI measurements and its uncertainty are estimated.

$$\mu_{ERIbias} = R_{ERI} (H C_{sub} H^T + R_{ERI})^{-1} (y_{ERI} - \mu_{sub}) + \mu_{sub} \qquad \text{Equation A13}$$

$$R_{ERIbias} = R_{ERI} - R_{ERI} (H C_{sub} H^T + R_{ERI})^{-1} R_{ERI} \qquad \text{Equation A14}$$

The estimated ERI bias and its uncertainty was calculated for each month using a value of $\sigma_G$ of 0.2°C (the mean ERI bias used in Kllb). The resulting time series along with similar series for buckets are shown in Figure 6. Example input and outputs for the interpolation are shown in Figure 18a-d. Common patterns can be seen between the anomaly fields estimated using ship data (a) and buoy data (b), but there are also some differences. For example, there are features in the ship data that follow common shipping routes which could correspond to measurement errors, particularly micro biases. In the analyzed anomaly field (c), the patterns that are common to both ship and buoy data have been picked out and some of the errors that can be identified by eye in the ship data have been successfully separated (d). The largest estimated errors are in the tropics and Southern Hemisphere where ship traffic is less frequent and individual ships can have a larger effect. However, in the North Pacific and Atlantic, the ship data provide vital detail where there is no buoy coverage and the aggregate biases are less pronounced.

In the bias-adjustment algorithm (Section 4.1.2), a time-varying, temporally-smoothed field of $\mu_{ERI}$ was used. This ensures that the ERI measurements are unbiased relative to the drifting buoys and oceanographic data, on longer space- and time-scales, but preserves individual ship biases, which are described by the error covariances.

**A3 Tests of the interpolation method**

We tested the interpolation method in three different ways, which probe different aspects of the reconstruction.

First, we looked at its ability to reconstruct data that had been deliberately withheld (Section A3.1). Data were withheld in three different ways: at random locations, by reducing coverage in a well-observed period to match that of the 19$^{th}$ century and at the locations of Argo floats. These probe different aspects of the reconstruction such as the ability to fill small gaps, to reconstruct large missing areas, and to estimate SST anomalies from an independent validation system.

Second, we looked at the method's skill in estimating biases for individual ships (Section A3.2). We compared our estimates to those calculated by comparing ships to a satellite-based analysis. Although we do not use the biases calculated for individual ships in HadSST.4.0.0.0, estimating them tests the method's ability to reconstruct biases locally.

Third, we generated realistic synthetic data for which the uncorrelated, micro bias and macro bias errors were known (Section A3.3). We then interpolated the data and reconstructed the micro and macro biases. This test ensures that the method can estimate the large-scale biases in different data sources under a different set of circumstances.

### A3.1 Testing the reconstruction using withheld data

In order to test the reconstruction method, fields from 2000 to 2014 were reduced in coverage. The coverage was reduced in three ways. First, coverage in the period 2000-2014 was reduced to that of 1850-1864. In the second test, half of the grid boxes were removed randomly. In the third test a quarter of grid cells for which Argo measurements were available were removed. The reduced-coverage SST fields were then reconstructed and the reconstruction was compared to the data that had been withheld, and in the third case also to Argo data in those grid cells that had been withheld. Two tests were then made.

In the first test the chi-squared statistic was calculated using,

$$\chi^2 = (\mu - w)^T (P + R)^{-1} (\mu - w) \qquad \text{Equation A15}$$

where $\mu$ is a vector containing the reconstructed SSTs at the locations of the withheld data, $w$. $P$ is the posterior covariance of the reconstruction and $R$ is the error covariance of the withheld data. For a good reconstruction – one for which the estimated fields and uncertainties are consistent – the statistic should follow a chi-squared distribution for which the number of degrees of freedom is equal to the rank of the covariance matrix, in this case equal to the number of withheld grid boxes (Povey et al. 2015).

The second test compared the withheld data to samples from the posterior distribution of the reconstruction, *P*, which were combined with samples drawn from the estimated error covariance of the withheld data, *R*. The combined samples give a set of fields that should resemble the withheld data – in both true SST variability and in the spectrum of observational error. The residual differences between the samples and the reconstruction were divided by the estimated uncertainty (scaled samples) as were the differences between the withheld data and the reconstruction (scaled observations). For a good reconstruction with well-specified observational uncertainties, the normalization step would, in the long run, yield distributions that are close to Normal with unit variance. However, for individual months, the spatial autocorrelation represented in *P* and *R* leads to distributions that do not look Normal.

The overall goodness of fit was assessed by examining how the distribution of scaled observations diverged from a distribution calculated from the scaled samples. A histogram of the scaled observations was calculated (with 0.1 unit bins) for all withheld data between 2000 and 2014. An equivalent histogram of the scaled samples was calculated for the same period and then repeated 500 times with different samples. If the histogram of scaled observations falls within the range of the 500 histrograms calculated from the scaled samples then the fit is considered a good one.

Figure 19(a) shows the distribution of the probability of the calculated chi-squared for each month between 2000 and 2016 and for each of the three different data-reduction schemes. For a good reconstruction, the probability ought to be approximately uniformly distributed. However, in this case there is a predominance of probabilities clustered towards one and zero, which suggests that the uncertainties are often underestimated and sometimes overestimated, or that the distribution of errors is not normal.

Figure 19 (b-d) compares the observed and theoretical distributions using the sampling method. For the case where data are "missing at random", the observed discrepancies are most likely to fall above the range defined by the samples where deviations are small - within 0.5 standard deviations – or where deviations are very large – in excess of 3 standard deviations. To compensate, the observed discrepancies are below the expectation in most other places. Such sharp-peaked, long-tailed distributions are characteristic of observational errors in ship and buoy data (K11c) and the shape would explain the poor chi-squared distribution. Considering the simplicity of the model, the agreement between the modeled and observed distributions is rather good. In the more challenging case where the coverage is reduced to that of the 19th century, the distribution is slightly broader than the theoretical case, suggesting a slight underestimate of the uncertainty.

The Argo test is also encouraging. The bias in the reconstruction relative to the higher accuracy Argo measurements that arises when using the warm-biased unadjusted SST data is removed when using the bias-adjusted SSTs (Figure 19d). The distribution falls below the expected range between ±2 standard deviations. There is a small excess at higher deviations which is associated with measurements made in the western boundary currents, marginal sea ice areas and other areas where sampling uncertainty is typically higher (assessed from maps of mean absolute residuals and root-mean-squared residuals, not shown). This suggests that either the sampling uncertainty is underestimated in these regions, or that the reconstruction technique cannot resolve the small scale variability in these regions.

An additional test was done based on the Argo data. Data were removed at the locations of the selected Argo data, but instead of comparing the reconstruction to the Argo data, it was compared to the withheld SST data. The shape of the distribution is somewhere between that seen in the missing-at-random case and that seen in the 19th Century coverage case. Given the good agreement between Argo and the reconstruction, the implication is that the narrow-peak, long-tail shape of the distribution of differences between SST and reconstruction arises purely from the distribution of observational errors in the SST data.

K11c investigated the effect of errors that are not normally distributed and which vary from agent to agent. They derived representative uncertainties that provide a good overall fit to the spectrum of errors seen in real data and it is these that we use in our analysis. Some care is needed when interpreting the standard deviations obtained from the data in terms of an actual distribution of errors because they cannot be assumed to be Normal. However, these tests indicate that there are no severe biases in the method.

### A3.2 Testing the reconstruction by estimating individual ship biases

It is possible to extract a posterior distribution for the error characteristics of each identifiable ship, which can then be compared to other estimates as a way of checking the reliability of the reconstruction method. This is done by noting that the error covariance, $R$, in the above equations is a simple sum of contributions from individual ships. Each of these individual contributions can be used to get an improved mean and covariance for the error characteristics of individual ships and buoys. Of particular interest are the micro biases, which are an important component of the measurement uncertainty at a global level.

It is worth pausing here momentarily to remind ourselves of the discussion and definitions of “error” and “uncertainty” from Section 2.1 and Section 3 because the nomenclature can get horribly confusing at this point. What we are attempting to estimate are the values of $B_{ij}$ from Equation 3.0. $B_{ij}$ is that part of the error (defined as the difference between the true SST and the measured SST) that is a persistent offset associated with a particular ship and for simplicity’s sake we shall assume that $B_{ij}$ is constant for a particular ship, $i$, so we can write $B_i$. To start with we assume that $B_i$ is zero for all ships, with a large uncertainty, $\sigma_{b_i}$. Using the method described below, we can make an improved estimate of the size of the error, $B_i'$, and its uncertainty $\sigma_{b_i}' < \sigma_{b_i}$.

For each ship, $i$, the individual contribution, $R_i$ to the overall error covariance (individual terms in the summation in Equation 3.4) was estimated and then the posterior mean and covariance of the errors for that ship were estimated using.

$$\mu_i = R_i(HCH^T + R)^{-1}y \qquad \text{Equation A16}$$

$$R_i' = R_i - R_i(HCH^T + R)^{-1}R_i \qquad \text{Equation A17}$$

The mean, $\mu_i$, found in this way is the weighted contribution of a particular error to the gridded averages, so it needs to be divided by the appropriate weight ($w_i$, Equation 3.2) to obtain an updated estimate of the bias, $B_i'$,

$$B_i' = \frac{\mu_i}{w_i}$$

Likewise the covariance $R_i'$ can be processed to obtain an updated estimate of $\sigma_{b_i}'$.

We applied Equations A16 and A17 for every month from 2000 to 2012 with: *y* being the median adjusted HadSST.4.0.0.0 for that month; the large-scale bias covariances in *R* set to zero; and C as derived in Section A1.1. We then extracted an estimate of $B_i$ for every uniquely identifiable ship.

For ships in well-travelled regions, the resulting micro biases, $B_i'$, remain close to the prior: the mean is zero and the uncertainty almost equal to $\sigma_{b_i}$. This is expected because many ships are averaged together in well-travelled grid boxes making the individual contributions impossible to separate. However, if a ship visits several, poorly-populated grid cells, the estimated micro bias will take some more-definite value and the posterior variance will be small.

Figure 20 shows the monthly estimated ship micro biases for those ships where the estimated uncertainty in the micro bias was less than 0.25°C compared to estimated micro biases taken from the IQUAM (in situ QUAlity Monitoring, Xu and Ignatov 2010) tool. In the IQUAM analysis, *in situ* data were compared to a daily background SST field derived from a combination of satellite and in *in situ* data. The correlation between the estimates from IQUAM and the

HadSST.4.0.0.0 analysis is around 0.7 where the uncertainty in the HadSST.4.0.0.0-estimated bias is less than 0.25°C. Differences are expected as IQUAM uses a different QC system and therefore the micro biases are not estimated from exactly the same observations. The figure shows the regression of the IQUAM estimates on HadSST.4.0.0.0 estimates and vice versa (diagonal blue lines). The two regression lines encompass a line that is parallel to y=x (the red line) and passes through 0.19°C, which can be interpreted as the approximate average bias between the two analyses during this period: the HadSST.4.0.0.0 estimate is adjusted to remove large-scale biases, but this is not done for the IQUAM data. A rough estimate of the large-scale average ship bias in HadSST.4.0.0.0 for this period can be calculated from the combined ship and buoy biases in the three regions shown in Figure 8 by dividing the bias by the fractional contribution of ships to the average. This gives a range of values from 0.14 (in the Northern Hemisphere) to 0.20°C (in the Southern Hemisphere) which encompasses 0.19°C. The analysis suggests that the magnitudes of the systematic errors for individual ships have not been systematically underestimated and that the reconstruction can reliably estimate the size of systematic errors in the data even for the challenging case of individual ship micro biases.

### A3.3 Testing the reconstruction using synthetic data

In the final test, a set of synthetic observations were generated from a globally-complete high-resolution (1/20° grid resolution) daily SST data set, called OSTIA (Operational Sea Surface Temperature and Sea Ice Analysis, Donlon et al. 2012). The idea is to create a synthetic, but realistic data set with known SSTs and known measurement errors. The techniques developed in this paper can then be applied to estimate the errors and compare them to the prescribed values.

The OSTIA SST fields were sampled at the locations of ICOADS observations and synthetic measurement errors were added to the data. If no OSTIA SST was available, no synthetic observation was produced. The errors were drawn from the error model described in Section 3.1. Each synthetic observation had an uncorrelated measurement error added to it which was drawn from a normal distribution with a standard deviation equal to the estimated uncorrelated error uncertainty for that platform (ship, drifting buoy or moored buoy, see Table 1). In addition, each individual agent, identified by its ID in ICOADS, was assigned a micro-bias error that was added to every observation made by the agent. The micro-bias errors were drawn from a normal distribution with a standard deviation equal to the estimated micro-bias error uncertainty for that platform (Table 1). Finally, characteristic biases were added to all observations made using a particular measurement method. The characteristic biases were drawn from a normal distribution with standard deviation of 0.2°C. New error values were drawn for each month for all components.

The synthetic observations were processed in the same way as the actual observations. The synthetic observations were gridded and the uncertainties in the gridded data were calculated (Section 3). The large-scale biases (Section A2) and micro-biases for individual ships were estimated (Section A3.2) for each month.

Figure 21 shows the difference between the estimated and prescribed synthetic biases for each measurement type and the estimated uncertainty envelope. The uncertainties in the biases are reasonably well estimated for the bucket and hull sensor measurements. However, there is some evidence of a slight bias in the estimation of the ERI biases, which appears to be consistently "warm" prior to 2009. This is odd, because the input biases have a mean of zero and are symmetrically distributed, and suggests that residual errors of around 0.05°C cannot be reliably eliminated using this method. Consequently, a lower limit on uncertainties estimated in this way is set at 0.05°C.

Figure 22 shows the comparison of the assigned and estimated micro-biases for ships. For the majority of ships, the estimated micro-biases are close to zero (Figure 22(a)), which is the mean of the prior estimate for the micro-biases. This happens because the estimates are based on coarsely gridded data and there is insufficient information to estimate the micro-biases if many ships contribute to the same grid-box average. This is reflected in the uncertainties attached to each estimate of the micro-bias. Selecting only those ships where the uncertainty is significantly lower than the prior value (Figure 22(b)-(d)) shows a closer correlation between the estimated and assigned micro-biases. This further demonstrates the ability of the method to extract individual ship biases.

### A4 Derivation of bucket biases

The bias for a grid cell can be written (Equation 4.1) as

$$\boldsymbol{B = f_e E + f_c B_{tc} + f_w B_{tw} + f_r B_{tr} + f_d D} \qquad \text{Equation A18}$$

The bucket correction for a canvas bucket, $B_c$, which adjusts the grid-box average bias to be consistent with the climatological average, can be written as

$$\boldsymbol{B_c = B_{tc} - \overline{B}} \qquad \text{Equation A19}$$

Where the overbar denotes the 1961-1990 average. Expanding this out:

$$\boldsymbol{B_c = B_{tc} - \overline{f_e E} - \overline{f_c} B_{tc} - \overline{f_w} B_{tw} - \overline{f_r} B_{tr} - \overline{f_d} D} \qquad \text{Equation A20}$$

The overbars cover only those terms of the equation that are time-varying. In addition, $f_w$ is zero during the climatology period (there are considered to be no wooden buckets) and $D$, the drifter bias, is zero at all times, so

$$\boldsymbol{B_c = B_{tc} - \overline{f_e E} - \overline{f_c} B_{tc} - \overline{f_r} B_{tr}} \qquad \text{Equation A21}$$

And, rearranging:

$$\boldsymbol{B_{tc} = \frac{B_c + \overline{f_e E} + \overline{f_r} B_{tr}}{\left(1 - \overline{f_c}\right)}} \qquad \text{Equation A22}$$

We get the formula for calculating the true bias of a canvas bucket given the bucket correction for a canvas bucket and the true biases for ERI and rubber bucket measurements.

## A5 Comparison to AR5

The IPCC AR5 (Hartman et al. 2013) showed trends over particular periods for a number of different SST data sets: HadISST1.1 (Rayner et al. 2003), HadSST2 (R06), COBE-SST (Ishii et al. 2005) and ERSSTv3 (Smith et al. 2008). Figure *23* shows the HadSST.4.0.0.0 and ERSSTv4 ensemble estimates of the trends over the same periods (1880-2012, 1901-2012, 1951-2012 and 1979-2012). Also shown are trends over the period 1998-2012 (used in Karl et al. 2015) and 2002-2012.

There are a number of interesting things to note. First, the spread from the HadSST.4.0.0.0 ensemble, which incorporates uncertainty in the bias adjustments only, is a major part of the uncertainty at long time scales (>30 years, Figure *23* (a)-(d)). At shorter time scales, however, the ensemble spread contributes only a small part to the overall uncertainty, the remainder coming from measurement and sampling errors, particularly the effect of micro-biases. This is particularly clear for the periods 1998-2012 and 2002-2012 where this component dominates

(Figure *23* (e) and (f)). Second, the ERSSTv4 and HadSST.4.0.0.0 ensembles overlap at all times, implying that the two estimates are consistent for this measure. Third, over the largest time periods (Figure *23*(a) and (b)) the net effect of the adjustments is to reduce the trend relative to the unadjusted observations.

In the period 1998-2012, the unadjusted data have a trend close to zero. In the period 2002-2012, the unadjusted trend is negative. In contrast, all the adjusted data sets indicate more warming (or less cooling) than in the unadjusted data. This is consistent across the three data sets. For other periods, 1979-2012, for example, the sign of the correction is not clear with some adjustments increasing the trend and others reducing it even within one ensemble.

Finally, the estimates from the three current, fully-adjusted, data sets all tend to sit at the upper end of the range from the data sets employed in IPCC AR5.

## References

Abraham, J.P., et al. (2013), A review of global ocean temperature observations: Implications for ocean heat content estimates and climate change, *Rev. Geophys.*, 51, 450-483, doi:10.1002/rog.20022.

Atkinson C.P., Rayner, N.A., Kennedy, J.J. and Good S.A. (2014), An integrated database of ocean temperature and salinity observations, *Journal of Geophysical Research: Oceans*, 119 (2014), no. 10, 7139-7163.

Berry, D.I., Kent, E.C. and Taylor, P.K. (2004), An Analytical Model of Heating Errors in Marine Air Temperatures from Ships. *J. Atmos. Oceanic Technol.*, 21, 1198–1215, https://doi.org/10.1175/1520-0426(2004)021<1198:AAMOHE>2.0.CO;2

Berry, D. I. and Kent, E. C. (2017), Assessing the health of the in situ global surface marine climate observing system. *Int. J. Climatol*., 37: 2248-2259. doi:10.1002/joc.4914

Berry, D. I. and Kent, E. C. (2011), Air–Sea fluxes from ICOADS: the construction of a new gridded dataset with uncertainty estimates. *Int. J. Climatol.*, 31: 987-1001. doi:10.1002/joc.2059

Berry, D.I.; Corlett, G.K.; Embury, O.; Merchant, C.J. (2018), Stability Assessment of the (A)ATSR Sea Surface Temperature Climate Dataset from the European Space Agency Climate Change Initiative. *Remote Sens.*, 10, 126.

Bojinski, S., M. Verstraete, T.C. Peterson, C. Richter, A. Simmons, and M. Zemp, (2014), The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy. Bull. Amer. Meteor. Soc., 95, 1431–1443, https://doi.org/10.1175/BAMS-D-13-00047.1

Brönnimann, S. (2005), The global climate anomaly 1940–1942. *Weather*, 60: 336-342. doi:10.1256/wea.248.04

Carella, G., Kent, E.C. and Berry, D.I. (2017a), A probabilistic approach to ship voyage reconstruction in ICOADS. *Int. J. Climatol.*, 37: 2233-2247. Doi:10.1002/joc.4492

Carella, G., Morris, A. K. R., Pascal, R. W., Yelland, M. J., Berry, D. I., Morak-Bozzo, S., Merchant, C. J. and Kent, E. C. (2017b), Measurements and models of the temperature change of water samples in sea-surface temperature buckets. *Q.J.R. Meteorol. Soc.*, 143: 2198–2209. doi:10.1002/qj.3078

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

Carton, J.A. and Giese, B.S. (2008), A Reanalysis of Ocean Climate Using Simple Ocean Data Assimilation (SODA). *Mon. Wea. Rev.*, 136, 2999–3017, https://doi.org/10.1175/2007MWR1978.1

Chan, D. and Huybers, P., (2019) Systematic differences in bucket sea surface temperature measurements amongst nations identified using a linear-mixed-effect method. *J. Climate*, https://doi.org/10.1175/JCLI-D-18-0562.1

Cheng, L., Abraham, J., Goni, G., Boyer, T., Wijffels, S., Cowley, R., Gouretski, V., Reseghetti, F., Kizu, S.,  Dong, S., Bringas, F., Goes, M., Houpert, L., Sprintall, J., and Zhu, J. (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

Christy J.R., Parker, D.E., Brown, S.J., Macadam, I., Stendel, M., and Norris, W.B. (2001), Differential trends in tropical sea surface and atmospheric temperatures since 1979, *Geophysical Research Letters* 28(1), 183-186.

Clark I. (1979), Practical Geostatistics, London (Applied Sci. Publ. Ltd.) ISBN-0-853340843-X

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N. , Allan, R. J., Yin, X. , Gleason, B. E., Vose, R. S., Rutledge, G. , Bessemoulin, P. , Brönnimann, S. , Brunet, M. , Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M. , Mok, H. Y., Nordli, Ø. , Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D. and Worley, S. J. (2011), The Twentieth Century Reanalysis Project. Q.J.R. Meteorol. Soc., 137: 1-28. doi:10.1002/qj.776

Cowley R., Wijffels, S., Cheng, L., Boyer, T., and Kizu, S. (2013), Biases in expendable bathythermograph data: A new view based on historical side-by-side comparisons, *Journal of Atmospheric and Oceanic Technology*, 30(6), 1195-1225.

Cowtan, K., Hausfather, Z., Hawkins, E., Jacobs, P., Mann, M.E., Miller, S.K., Steinman, B.A., Stolpe, M. B., and Way, R. G. (2015), Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures, *Geophys. Res. Lett.*, 42, 6526–6534, doi:10.1002/2015GL064888.

Cowtan, K., Rohde, R., Hausfather, Z. (2018), Evaluating biases in sea surface temperature records using coastal weather stations. *Q J R Meteorol Soc.* 144: 670–681. https://doi.org/10.1002/qj.3235

Donlon, C., et al. (2007), The Global Ocean Data Assimilation Experiment High-resolution Sea Surface Temperature Pilot Project. *Bulletin of the American Meteorological Society*, 88, 11971213. doi: 10.1175/BAMS-88-8-1197

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E. and Wimmer, W. (2012), The operational sea surface temperature and sea ice analysis (OSTIA) system. *Remote Sensing of Environment*, 116, 140-158.

Embury O., Merchant, C.J., and Corlett, G.K. (2012), A reprocessing for climate of sea surface temperature from the along-track scanning radiometers: Initial validation, accounting for skin and diurnal variability effects, *Remote Sensing of Environment* 116, 62-78, Advanced Along Track Scanning Radiometer(AATSR) Special Issue.

Embury, O. (2012), ARC: Level 3 daily sea surface temperature data v1.1. NERC Earth Observation Data Centre, downloaded 5 Jul 2017. http://catalogue.ceda.ac.uk/uuid/e6497acddf9cd8345ffbd0643c0d9729

Folland, C.K. and Parker, D.E. (1995), Correction of instrumental biases in historical sea surface temperature data. Q.J.R. Meteorol. Soc., 121: 319-367. doi:10.1002/qj.49712152206

Freeman, E. , Woodruff, S. D., Worley, S. J., Lubker, S. J., Kent, E. C., Angel, W. E., Berry, D. I., Brohan, P. , Eastman, R. , Gates, L. , Gloeden, W. , Ji, Z. , Lawrimore, J. , Rayner, N. A., Rosenhagen, G. and Smith, S. R. (2017), ICOADS Release 3.0: a major update to the historical marine climate record. Int. J. Climatol., 37: 2211-2232. doi:10.1002/joc.4775

Garrett, J.F. (1980), Availability of the FGGE drifting buoy system data set, *Deep Sea Research Part A. Oceanographic Research Papers* 27(12), 1083-1086 https://doi.org/10.1016/0198-0149(80)90067-9

Good, S.A., Martin, M.J. and Rayner, N.A. (2013), EN4: quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, *Journal of Geophysical Research: Oceans*, 118, 6704-6716, doi:10.1002/2013JC009067

Gouretski, V. (2012), Using GEBCO digital bathymetry to infer depth biases in the XBT data, *Deep Sea Research Part I: Oceanographic Research Papers* 62(C), 40-52.

Gouretski V., Kennedy, J., Boyer, T., and Köhl, A. (2012), Consistent near-surface ocean warming since 1900 in two largely independent observing networks, *Geophysical Research Letters* 39, L19606, doi: 10.1029/2012GL052975.

Gouretski, V. and Reseghetti, F. (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 Research Part I: Oceanographic Research Papers* 57(6), 812 - 833. https://doi.org/10.1016/j.dsr.2010.03.011

Hartmann, D.L., Klein Tank, A.M.G., Rusticucci, M., Alexander, L.V., Brönnimann, S., Charabi, Y., Dentener, F.J., Dlugokencky, E.J., Easterling, D.R., Kaplan, A., Soden, B.J., Thorne, P.W., Wild, M. and Zhai, P.M. (2013), Observations: Atmosphere and Surface. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

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

Hersbach, H., Peubey, C., Simmons, A., Berrisford, P., Poli, P. and Dee, D. (2015), ERA-20CM: a twentieth-century atmospheric model ensemble. *Q.J.R. Meteorol. Soc*, 141: 2350–2375. doi:10.1002/qj.2528

Hirahara S., Ishii, M., and Fukuda, Y. (2014), Centennial-scale sea surface temperature analysis and its uncertainty, *Journal of Climate* 27(1), 57-75. https://doi.org/10.1175/JCLI-D-12-00837.1

Huang, B., L'Heureux, M., Lawrimore, J., Liu, C., Zhang, H., Banzon, V., Hu, Z. and Kumar, A. (2013), Why Did Large Differences Arise in the Sea Surface Temperature Datasets across the

Tropical Pacific during 2012?. *J. Atmos. Oceanic Technol*., 30, 2944–2953, https://doi.org/10.1175/JTECH-D-13-00034.1

Huang B., Banzon, V.F., Freeman, E., Lawrimore, J., Liu, W., Peterson, T.C., Smith, T.M., Thorne, P.W., Woodruff, S.D. and Zhang, H-Min (2015), Extended reconstructed sea surface temperature version 4 (ERSST v4). Part I: Upgrades and intercomparisons, *Journal of Climate* 28(3), 911-930, https://doi.org/10.1175/JCLI-D-14-00006.1

Huang B., Thorne, P.W., Smith, T.M., Liu, W., Lawrimore, J., Banzon, V.F., Zhang, H-Min, Peterson, T.C. and Menne, M. (2016), Further exploring and quantifying uncertainties for extended reconstructed sea surface temperature (ERSST) version 4 (v4), *Journal of Climate* 29(9), 3119-3142, https://doi.org/10.1175/JCLI-D-14-00007.1

Huang B., Thorne, P.W., Banzon, V.F., Boyer, T., Chepurin, G., Lawrimore, J.H., Menne, M.J., Smith, T.M., Vose, R.S., and Zhang, H-Min (2017), Extended reconstructed sea surface temperature version 5 (ERSSTv5): Upgrades, validations, and intercomparisons, *Journal of Climate*, 30, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1

Huang, B., Angel, W., Boyer, T., Cheng, L., Chepurin, G., Freeman, E., Liu, C., and Zhang, H. (2018), Evaluating SST Analyses with Independent Ocean Profile Observations. *J. Climate*, 31, 5015–5030, https://doi.org/10.1175/JCLI-D-17-0824.1

Ishii, M. , Shouji, A. , Sugimoto, S. and Matsumoto, T. (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

JCGM 100 – Evaluation of measurement data – Guide to the expression of uncertainty in measurement (ISO/IEC Guide 98-3) https://www.iso.org/sites/JCGM/GUM/JCGM100/C045315e-html/C045315e.html?csnumber=50461

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D. (1996), The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, 77, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Karl, T.R., Arguez, A., Huang, B., Lawrimore, J.H., McMahon, J.R., Menne, M.J., Peterson, T.C., Vose, R.S. and Zhang, H-Min (2015), Possible artifacts of data biases in the recent global surface warming hiatus, Science, aaa5632. DOI: 10.1126/science.aaa5632

Karspeck, A.R., Kaplan, A. and Sain, S.R. (2012), Bayesian modelling and ensemble reconstruction of mid-scale spatial variability in North Atlantic sea-surface temperatures for 1850-2008, *Quarterly Journal of the Royal Meteorological Society* 138(662), 234-248. doi:10.1002/qj.900

Kawai, Y. and Wada, A. (2007), Diurnal sea surface temperature variation and its impact on the atmosphere and ocean: a review. *Journal of Oceanography*, 2007, 63(5), 721-744. https://doi.org/10.1007/s10872-007-0063-0

Kennedy J.J., Rayner, N.A., Smith, R.O., Saunby, M. and Parker, D.E. (2011a), Reassessing biases and other uncertainties in sea-surface temperature observations measured in situ since 1850, Part 1: measurement and sampling uncertainties, *J. Geophys. Res. Atmospheres*., 116, D14103, doi: 10.1029/2010JD015218.

Kennedy J.J., Rayner, N.A., Smith, R.O., Saunby, M. and Parker, D.E. (2011b), Reassessing biases and other uncertainties in sea-surface temperature observations measured in situ since 1850, Part 2: biases and homogenisation, *J. Geophys. Res. Atmos.*, 116, D14104, doi: 10.1029/2010JD015220..

Kennedy J.J., Smith, R.O. and Rayner, N.A. (2011c), Using AATSR data to assess the quality of in situ sea surface temperature observations for climate studies, *Remote Sensing of Environment*, 116(15), 79-92, https://doi.org/10.1016/j.rse.2010.11.021

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

Kent, E.C. and Taylor, P.K. (2006), Toward Estimating Climatic Trends in SST. Part I: Methods of Measurement. *J. Atmos. Oceanic Technol.*, 23, 464–475, https://doi.org/10.1175/JTECH1843.1

Kent, E.C. and Kaplan, A. (2006), Toward Estimating Climatic Trends in SST. Part III: Systematic Biases. *Journal of Atmospheric and Oceanic Technology*, 23:3, 487-500, https://doi.org/10.1175/JTECH1845.1

Kent, E.C., Kennedy, J.J., Berry, D.I. and Smith, R.O. (2010), Effects of instrumentation changes on sea surface temperature measured in situ, *Wiley Interdisciplinary Reviews: Climate Change*, 1(5), 718-728, https://doi.org/10.1002/wcc.55

Kent, E.C., Woodruff, S.D. and Berry, D.I. (2007), Metadata from WMO publication no. 47 and an assessment of voluntary observing ship observation heights in ICOADS, *Journal of Atmospheric and Oceanic Technology*, 24(2), 214-234, https://doi.org/10.1175/JTECH1949.1

Kent, E.C., Kennedy, J.J., Smith, T.M., Hirahara, S., Huang, B., Kaplan, A., Parker, D.E., Atkinson, C.P., Berry, D.I., Carella, G., Fukuda, Y., Ishii, M., Jones, P.D., Lindgren, F., Merchant, C.J., Morak-Bozzo, S., Rayner, N.A., Venema, V., Yasui, S., and Zhang, H. (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

Kent, E.C., Rayner, N.A., Berry, D.I., Saunby, M., Moat, B.I., Kennedy, J.J., and Parker, D.E. (2013), Global analysis of night marine air temperature and its uncertainty since 1880: The HadNMAT2 data set, *Journal of Geophysical Research: Atmospheres*, 118(3), 1281-1298, https://doi.org/10.1002/jgrd.50152

Lean, K. and Saunders, R.W. (2013), Validation of the ATSR reprocessing for climate (ARC) dataset using data from drifting buoys and a three-way error analysis, *Journal of Climate* 26(13), 4758-4772, https://doi.org/10.1175/JCLI-D-12-00206.1

Levitus S., Antonov, J.I., Boyer, T.P., Locarnini, R.A., Garcia, H.E., and Mishonov, A.V. (2009), Global ocean heat content 1955-2007 in light of recently revealed instrumentation problems, *Geophys. Res. Lett.*, 36, L07608. https://doi.org/10.1029/2008GL037155

Liu W., Huang, B., Thorne, P.W., Banzon, V.F., Zhang, H-Min, Freeman, E., Lawrimore, J., Peterson, T.C., Smith, T.M. and Woodruff, S.D. (2015), Extended reconstructed sea surface temperature version 4 (ERSSTv4): Part II. Parametric and structural uncertainty estimations, *Journal of Climate*, 28(3), 931-951, https://doi.org/10.1175/JCLI-D-14-00007.1

Medhaug I., Stolpe, M.B., Fischer, E.M. and Knutti, R. (2017), Reconciling controversies about the global warming hiatus, *Nature* 545(7652), 41-47.

Merchant C.J., Embury, O., Rayner, N.A., Berry, D.I., Corlett, G.K., Lean, K., Veal, K.L., Kent, E.C., Llewellyn-Jones, D.T. and Remedios, J.J. (2012), A 20 year independent record of sea surface temperature for climate from along-track scanning radiometers, *Journal of Geophysical Research: Oceans*, (1978-2012) 117(C12), https://doi.org/10.1029/2012JC008400

Merchant, C. J., Embury, O., Roberts-Jones, J., Fiedler, E., Bulgin, C. E., Corlett, G. K., Good, S., McLaren, A., Rayner, N., Morak-Bozzo, S. and Donlon, C. (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. doi:10.1002/gdj3.20

Nelder and Mead (1965) Computer Journal, Vol 7, pp 308-313

Newman M., Alexander, M.A., Ault, T.R., Cobb, K.M., Deser, C., Di Lorenzo, E., Mantua, N.J., Miller, A.J., Minobe, S., Nakamura, H., Schneider, N., Vimont, D.J., Phillips, A.S., Scott, J.D. and Smith, C.A. (2016), The Pacifc decadal oscillation, revisited, *Journal of Climate*, 29(12), 4399-4427, https://doi.org/10.1175/JCLI-D-15-0508.1

O'Carroll, A.G., Eyre, J.R. and Saunders, R.W. (2008), Three-Way Error Analysis between AATSR, AMSR-E, and In Situ Sea Surface Temperature Observations. *J. Atmos. Oceanic Technol.*, 25, 1197–1207, https://doi.org/10.1175/2007JTECHO542.1

Povey, A. C. and Grainger, R. G. (2015) Known and unknown unknowns: uncertainty estimation in satellite remote sensing, *Atmos. Meas. Tech.*, 8, 4699-4718, https://doi.org/10.5194/amt-8-4699-2015.

Rasmussen, C.E. and Williams, C.K.I. (2006) Gaussian Processes for Machine Learning, MIT Press, 2006. ISBN-10 0-262-18253-X, ISBN-13 978-0-262-18253-9.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C. and Kaplan, A. (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, D14.

Rayner N.A., Brohan, P., Parker, D.E., Folland, C.K., Kennedy, J.J., Vanicek, M., Ansell, T.J., and Tett, S.F.B. (2006), Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 data set, *Journal of Climate*, 19(3), 446-469, https://doi.org/10.1175/JCLI3637.1

Research Data Archive/Computational and Information Systems Laboratory/National Center for Atmospheric Research/University Corporation for Atmospheric Research, Physical Sciences Division/Earth System Research Laboratory/OAR/NOAA/U.S. Department of Commerce, Cooperative Institute for Research in Environmental Sciences/University of Colorado, National Oceanography Centre/University of Southampton, Met Office/Ministry of Defence/United Kingdom, Deutscher Wetterdienst (German Meteorological Service)/Germany, Department of Atmospheric Science/University of Washington, Center for Ocean-Atmospheric Prediction Studies/Florida State University, and National Centers for Environmental Information/NESDIS/NOAA/U.S. Department of Commerce. 2016, updated monthly. International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 3, Individual

Observations. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/D6ZS2TR3. Accessed 12 August 2016.

Smith, T.M. and Reynolds, R.W. (2002), Bias corrections for historical sea surface temperatures based on marine air temperatures, *Journal of Climate* 15(1), 73-87, https://doi.org/10.1175/1520-0442(2002)015<0073:BCFHSS>2.0.CO;2

Smith, T.M., Reynolds, R.W., Peterson, T.C., and Lawrimore, J. (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

Sybrandy A.L., Niiler, P.P., Martin, C., Scuba, W., Charpentier, E. and Meldrum, D.T. (2008), Global Drifter Programme Barometer Drifter Design Reference, DBCP Report No. 4 REVISION 2.2, August 2009. Published by the *Data Buoy Cooperation Panel*. Downloaded from http://www.jcommops.org/dbcp/community/standards.html 17 February 2018.

Thompson, D.W.J., Kennedy, J.J., Wallace, J.M. and Jones, P.D. (2008), A large discontinuity in mid 20th century global-mean surface temperatures, *Nature* 453, 646-649, http://dx.doi.org/10.1038/nature06982

Thorne, P.W., Parker, D.E., Christy, J.R. and Mears, C.A. (2005), Uncertainties in climate trends - lessons from upper- air temperature records, *Bulletin of the American Meteorological Society* 86(10), 1437-1442, https://doi.org/10.1175/BAMS-86-10-1437

Wilcox, R. (2001), Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy, New York: Springer.

Willett, K.M., Jones, P.D., Gillett, N.P., and Thorne, P.W. (2008), Recent Changes in Surface Humidity: Development of the HadCRUH Dataset. *J. Climate*, 21, 5364–5383, https://doi.org/10.1175/2008JCLI2274.1

Woodruff S.D., Worley, S.J., Lubker, S.J., Ji, Z., Freeman, J.E., Berry, D.I., Brohan, P., Kent, E.C., Reynolds, R.W., Smith, S.R. and Wilkinson, C. (2011), ICOADS release 2.5: extensions and enhancements to the surface marine meteorological archive, *International Journal of Climatology*, 31, 951-967, https://doi.org/10.1002/joc.2103

Xu, F. and Ignatov, A. (2010), Evaluation of in situ sea surface temperatures for use in the calibration and validation of satellite retrievals, *Journal of Geophysical Research: Oceans* (1978-2012), 115, C09022, doi: 10.1029/2010JC006129.

**Tables**

Table 1: Values used for uncertainties arising from uncorrelated random effects ($\sigma_u$) and from random effects which are correlated for a particular agent ($\sigma_b$) for ships, drifting buoys and moored buoys from K11c.

| | $\sigma_u$ | $\sigma_b$ |
|---|---|---|
| **Ships** | 0.74°C | 0.71°C |
| **Drifting buoys** | 0.26°C | 0.29°C |
| **Moored buoys** | 0.30°C | 0.20°C |

## Figure Captions

Figure 1: numbers of observations passing QC (a, c, e) and super-observations (b, d, f, see Section 3 for the definition) per month for the globe (black), Southern Hemisphere (orange) and Northern Hemisphere (blue) for (a,b) 1850-1880, (c, d) 1880-2000 and (e, f) 2000-2018. Note the very different scales for the y-axes.

Figure 2: Fractional contribution of different SST observation methods, 1915-2018, to (a) the Global average, (b) the Southern Hemisphere average and (c) the Northern Hemisphere average. The brown/orange/tan areas indicate ship observations as labeled in panel (a) and the blue areas indicate buoy observations. The pale lilac area represents unknown measurement method (assumed to be from ships). These are the initial assignments (Section 2.2.1) and are not the assignments finally used to calculate the adjustments.

Figure 3: Example fields from the gridding procedure for the 5° by 5° by pentad grid boxes: (a) SST anomalies (°C) for June 2003 relative to the 1961-1990 average; (b) number of observations

contributing to each grid-box average; (c) number of super observations in each grid box; (d) fractional contribution to the grid-box average by ERI measurements; (e) fractional contribution to the grid-box average by drifting buoys (drifters); (f) fractional contribution to the grid-box average by moored buoys (moorings); (g) estimated uncertainty associated with uncorrelated errors; (h) estimated uncertainty (ºC) associated with under sampling; and (i) estimated uncertainty (°C) associated with micro-bias errors.

Figure 4: Schematic diagram showing the flow of information in the parameter and bias estimation and the corresponding sections in the paper. Blue boxes are input data sets and analyses. Pink boxes correspond to steps described in Section 2 and 3. Orange boxes correspond to processing described in Section 4.

Figure 5: Annual global averages of the 200 realizations of the bucket corrections (a) R06-style corrections (b) SR02-style corrections, (c) combined R06 and SR02 corrections. Blue lines in (a)-(c) show ensemble members with a linear transition from wooden to canvas buckets and orange lines show ensemble members generated assuming a step change in the fraction of wooden and canvas buckets. (d) Estimated seasonal cycle of insulated bucket biases.

Figure 6: (left column) Estimated seasonal-average ERI biases (orange) and bucket biases (blue) (°C) 1940-2018 for (a) the globe, (c) the southern hemisphere and (e) the northern hemisphere. (right column) Smoothed estimated monthly ERI biases (°C) for (b) the globe, (d) the southern hemisphere and (f) the northern hemisphere. The smoothed estimate is shown in orange with the seasonal-averages from the left column shown in grey.

Figure 7: Smoothed time series of the estimated fraction $f_{correct}$ of measurements labeled as buckets that were correctly identified as buckets for (a) rejected start and end dates and (b)

accepted start and end dates. (c) accepted (blue) and rejected (red) start and end date combinations. (d) The black line is the annual average inferred fraction of correct bucket assignments which is the average of the unsmoothed blue lines in (b). The red lines indicate the mean and ranges used to draw samples. The mean values are held constant at 0.5 between 1945 and 1952 and at 0.95 after 1978. The blue lines show 10 samples of the full 200 member ensemble from 1945 on. Before 1945 the mean is set to 0.5, but the uncertainty is larger.

Figure 8: (a) Estimated monthly bias (°C, orange) in global average SST 1850-2018 for the full gridded dataset, including ships and buoys and (b) estimated bias in global average SST anomaly (°C, blue) relative to a 1961-1990 period, again for all data. (c and d) as for (a and b) except for the Southern Hemisphere. (e and f) as for (a and b) except for the Northern Hemisphere. The black line is the median bias and the shaded area represents the 95% range of the estimated biases.

Figure 9: Monthly global average SST anomalies (°C) 1850-2018 (a) relative to the unadjusted 1961-1990 climatology and (b) relative to the bias-adjusted 1961-1990 climatology. (c and d) as for (a and b) except for the Southern Hemisphere. (e and f) as for (a and b) except for the Northern Hemisphere. The grey line shows the unadjusted data, the black line is the median of the adjusted data. The blue and orange shading represents the 95% range of the ensemble.

Figure 10: (a) Annual global average unadjusted SST anomalies 1940-2014 (°C relative to unadjusted 1961-1990 climatology) for collocated bucket (blue) and ERI (orange) measurements. The solid line is the best estimate and the shaded area is the 95% uncertainty range (accounting for measurement and sampling errors). (b) Adjusted anomalies with expanded

uncertainty range including bias adjustment uncertainty. The dotted line indicates the best estimate of the unadjusted series from (a). (c) and (d) as for (a) and (b) but for the Southern Hemisphere. (e) and (f) as for (a) and (b) but for the Northern Hemisphere.

Figure 11: Comparisons between sea-surface temperature data sets from different sources 1990-July 2018, different comparison data sets cover different periods. (a) Global average SST anomalies (°C), relative to the 1961-1990 HadSST2 climatology, from Argo floats (purple and purple shading) and HadSST.4.0.0.0 (black solid line and grey shading) each reduced to their common coverage. Shading indicates the 95% uncertainty range. The unadjusted SST data are shown as a black dotted line. (b) Indicator of the number of 5° grid boxes in HadSST.4.0.0.0 (pale grey) and in the Argo data set (purple) Data are plotted only for the overlap. (c) as for (a), but with ARC (in red, no uncertainty range shown) substituted for Argo. (d) as for (b) but with ARC substituted for Argo. (e) as for (a) but with buoys (in blue and blue shading) substituted for Argo. (f) as for (b) but with buoys substituted for Argo.

Figure 12: Average SST difference (°C) between HadSST.4.0.0.0 and the three instrumentally homogeneous data sets (a) buoys 1995-2018, (b) ARC 1995-2012, and (c) Argo 2000-2017. (d) Shows the difference between HadSST.4.0.0.0 and the unadjusted gridded SSTs, 1995-2018.

Figure 13: (a) Collocated global annual average NMAT anomalies (°C) 1900-2010 offset by 0.15°C (blue, relative to 1961-1990) and global annual average SST anomalies from HadSST.4.0.0.0 (black is central estimate and grey shading indicates 95% uncertainty range). (b) Offset NMAT anomalies minus SST anomalies with combined 95% uncertainty range (taking into account the bias errors from the HadSST.4.0.0.0 ensemble, and measurement and sampling errors in the SST). The dashed line indicates zero difference. (c) Collocated global annual

average offset NMAT anomalies (blue), global annual average near-surface water temperature from HadIOD excluding Argo (orange) and SST (black is central estimate and grey shading again indicates 95% uncertainty range). (d) Difference between HadIOD and HadSST.4.0.0.0 (orange) and NMAT and HadSST.4.0.0.0 (blue). The shaded area indicates the 95% uncertainty range.

Figure 14: (a and b) Global, (c and d) Southern Hemisphere and (e and f) Northern Hemisphere annual average SST anomalies (°C) 1850-2018 relative to 1961-1990 for HadSST.4.0.0.0 (black line is the median in the left column and the grey shading in the right column represents the 95% uncertainty range) and HadSST.3.1.1.0 (blue line is the median and the blue shading represents the 95% uncertainty range). Uncertainty estimates combine the bias-adjustment uncertainties from the ensemble with measurement and sampling uncertainties.

Figure 15: (a) Global average SST anomaly 1850-2012 (°C relative to 1961-1990) series from HadSST.4.0.0.0 (black), ERSSTv5 (blue, thick line is operational version and thin-thick dashed lines are ensemble range from the 1000-member ERSSTv4 ensemble), COBE-SST-2 (orange), HadSST.3.1.1.0 (green) and unadjusted SSTs (red). All data sets are averaged on to a 5° grid and reduced to HadSST.4.0.0.0 coverage before comparison. (b) Global-average difference for each data set from HadSST.4.0.0.0. The grey shading shows the 95% uncertainty range from HadSST.4.0.0.0 including effects from measurement, sampling and bias-adjustment errors. The bias-adjustment uncertainty range is shown in darker grey. (c) and (d) as for (a) and (b) but for the Southern Hemisphere. (e) and (f) as for (a) and (b) but for the Northern Hemisphere.

Figure 16: Global and regional average SST anomaly trends to 2012. Median trends from HadSST.4.0.0.0 are indicated by a black horizontal line and the grey shading indicates median

and 95% uncertainty range including effects from measurement, sampling and bias-adjustment errors. The bias-adjustment uncertainty range is shown in darker grey; ERSSTv5 (blue), the lozenge is the operational version and the vertical line is the 95% ensemble range from the 1000-member ERSSTv4 ensemble); COBE-SST-2 (orange); HadSST.3.1.1.0 (green); and unadjusted SSTs (red).

Figure 17: (a)  Standard deviation of SST anomalies (°C), (b) Zonal length scale Lx (km) and (c) meridional length scale Ly (km) used in the interpolation scheme..

Figure 18: Example of the inputs and outputs of the interpolation for July 2003. (a) gridded SST anomalies (°C) from ships. (b) gridded SST anomalies from buoys. (c) interpolated SST anomalies. (d) estimated ERI biases (°C). Note that the ERI data are a subset of the ship data so the coverage is not identical.

Figure 19: Tests of effectiveness of the reconstruction method from 2000-2014. Blue lines show results for the 19th century coverage tests, red lines show the results of the missing-at-random tests and the purple lines show the test where data were removed at selected locations of Argo observations. (a) probability distribution for chi-squared statistics calculated from the withheld data. (b) average of histograms from all scaled, withheld observations in the 19th century coverage test (blue line) compared to the samples drawn from the posterior of the reconstruction (black line and grey shading). (c) as for (b) but for the missing-at-random test (red). (d) as for (b) but for the test where data were removed at the locations of Argo observations (purple). In (d) the dashed purple line shows the offset that occurs when doing the reconstruction based on

unadjusted data and the thin blue line shows the scaled residuals of the adjusted SSTs at the locations of the Argo observations.

Figure 20: Density map of estimated biases for ships from the HadSST.4.0.0.0 analysis (where the estimated uncertainty on the bias was less than 0.25°C) and from the IQUAM analysis between 2000 and 2012. The blue lines show the regression of the IQUAM estimate on HadSST.4.0.0.0 and vice versa. The red line is the y=x line offset by 0.19°C. A small white cross marks [0,0]. White lines show the boundaries of the bins, separated by 0.1K.

Figure 21: Tests of bias estimation using synthetic data. (a) histograms of differences between prescribed and estimated characteristic biases for each month for each measurement type in the synthetic data set including buckets (blue), ERI (orange) and hull sensors (hot pink). (b) time series of difference between assigned and estimated biases for bucket measurements (blue) and the 95% uncertainty range (grey lines). (c) as for (b) but for ERI measurements in orange. (d) as for (b) but for hull sensors in hot pink.

Figure 22 (a) Density map of prescribed vs estimated micro-biases for all ships in the synthetic data set. Darker colours indicate higher densities. (b) as for (a) except it shows only those ships for which the uncertainty in the estimated micro-bias is less than 0.5°C. (c) as for (b) but with an uncertainty less than 0.15°C. (d) as for (b) but for uncertainties less than 0.05°C. The black diagonal line is y=x. White lines show the boundaries of the bins, separated by 0.1K.

Figure 23: Histograms for trend estimates in global average temperatures (°C/decade) for different periods (indicated in the individual plot titles) and different data sets. HadSST.4.0.0.0 is shown in grey (generated from the 200-member ensemble of bias adjustments only) and black (generated from the 200-member ensemble of bias adjustments combined with samples from the

measurement and sampling uncertainties). The ERSSTv4 ensemble is shown in blue, with the operational ERSSTv5 point shown as a blue lozenge. COBE-SST-2 is shown in orange. The unadjusted data are shown in red. In addition, the trends from the IPCC AR5 SST data sets are shown as green lozenges numbered as follows: 1 is HadSST3, 2 is HadSST2, 3 is HadISST1.1, 4 is COBE SST, and 5 is ERSSTv3. They do not appear in panels (e) and (f) as these periods were not considered in Chapter 2.

**Figure 1.**

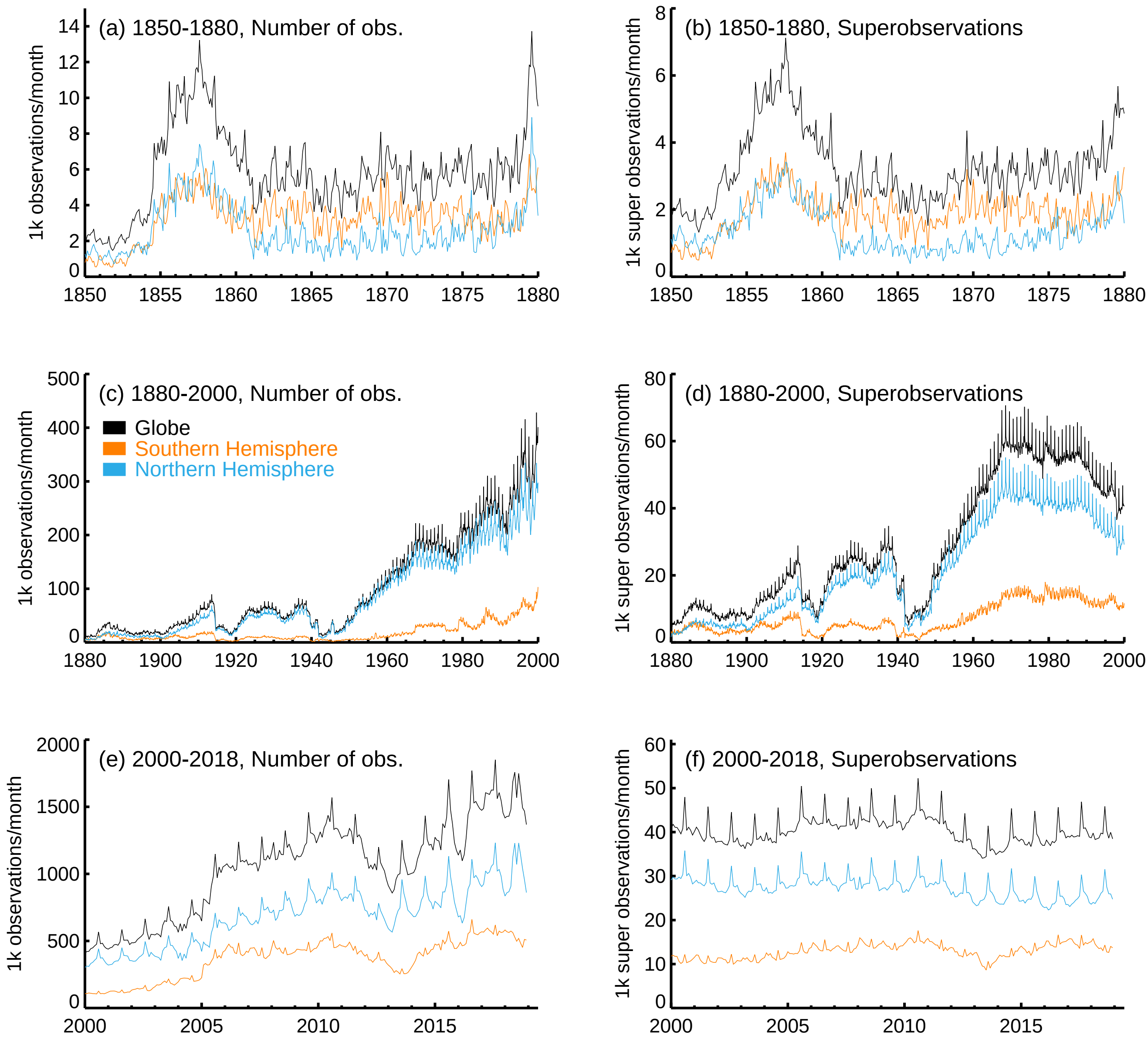
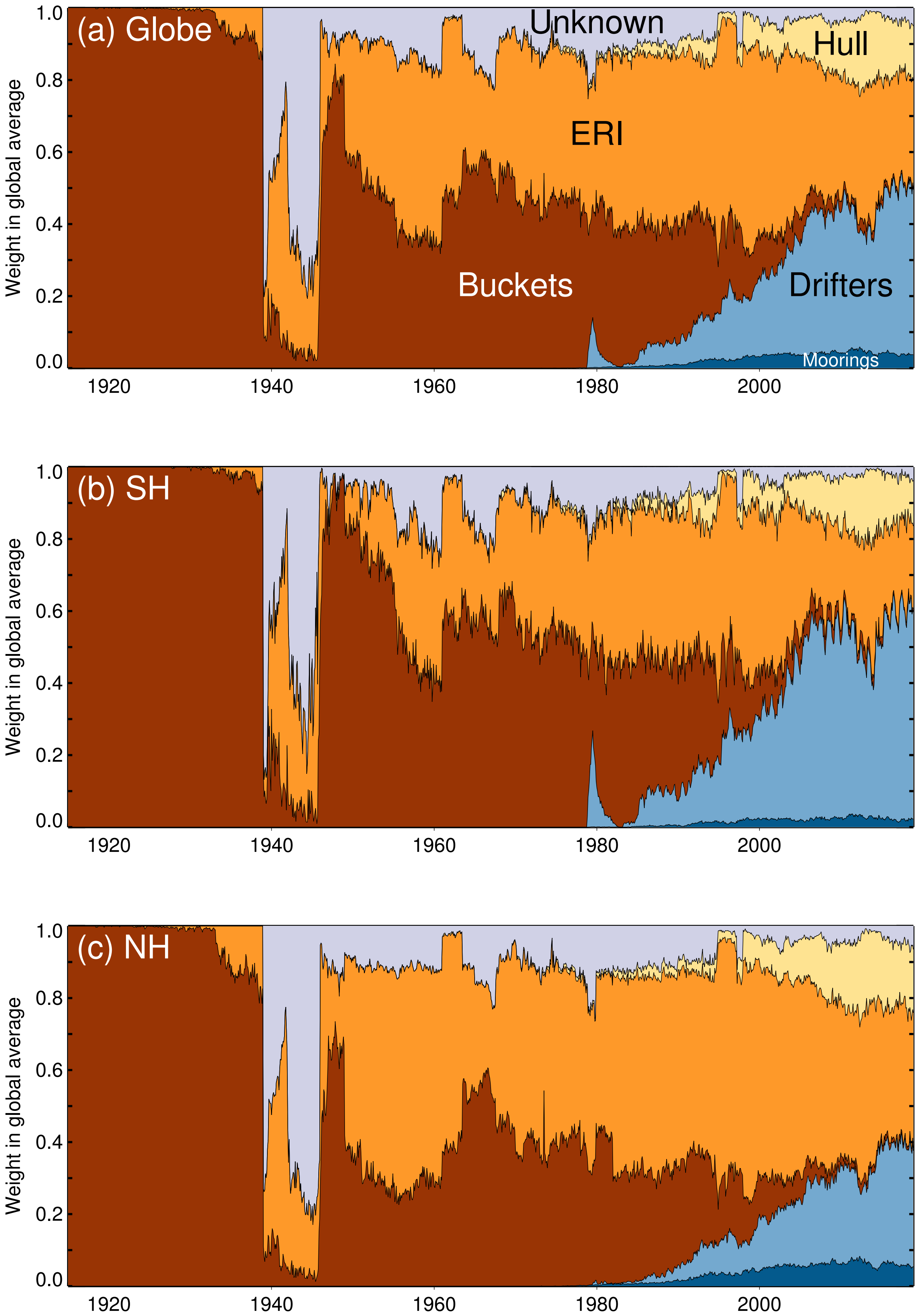

(a) 1850-1880, Number of obs.
1k observations/month
(b) 1850-1880, Superobservations
1k super observations/month
(c) 1880-2000, Number of obs.
Globe
Southern Hemisphere
Northern Hemisphere
(d) 1880-2000, Superobservations
(e) 2000-2018, Number of obs.
(f) 2000-2018, Superobservations

**Figure 2.**



(a) Globe
Unknown
Hull
ERI
Buckets
Drifters
Moorings
(b) SH
(c) NH
Weight in global average
1.0
0.8
0.6
0.4
0.2
0.0
1920
1940
1960
1980
2000

**Figure 3.**

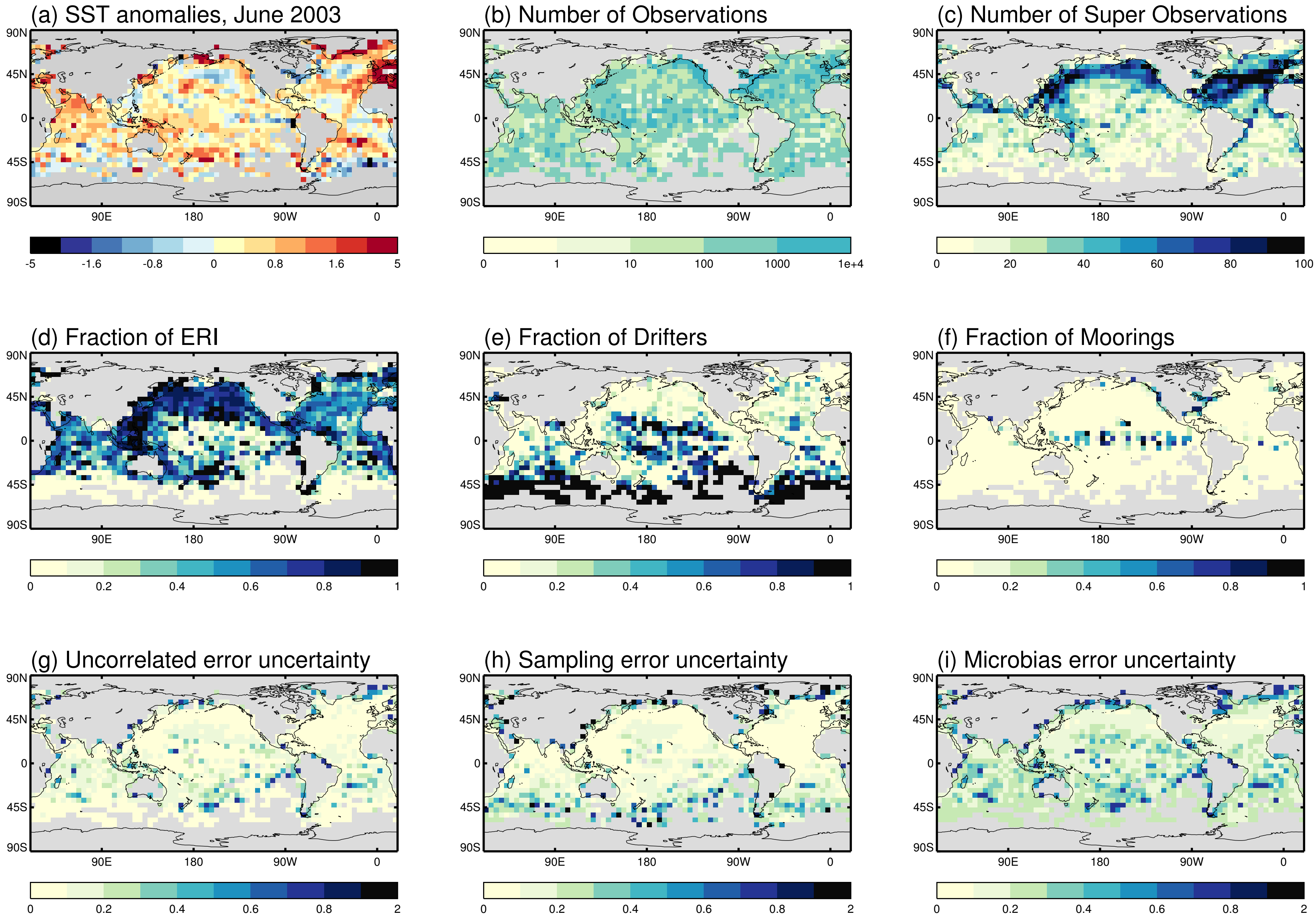

(a) SST anomalies, June 2003
(b) Number of Observations
(c) Number of Super Observations
(d) Fraction of ERI
(e) Fraction of Drifters
(f) Fraction of Moorings
(g) Uncorrelated error uncertainty
(h) Sampling error uncertainty
(i) Microbias error uncertainty
90N
45N
0
45S
90S
90E
180
90W
0
-5
-1.6
-0.8
0
0.8
1.6
5
1
10
100
1000
1e+4
20
40
60
80
100
0.2
0.4
0.6
0.8
1
2

**Figure 4.**

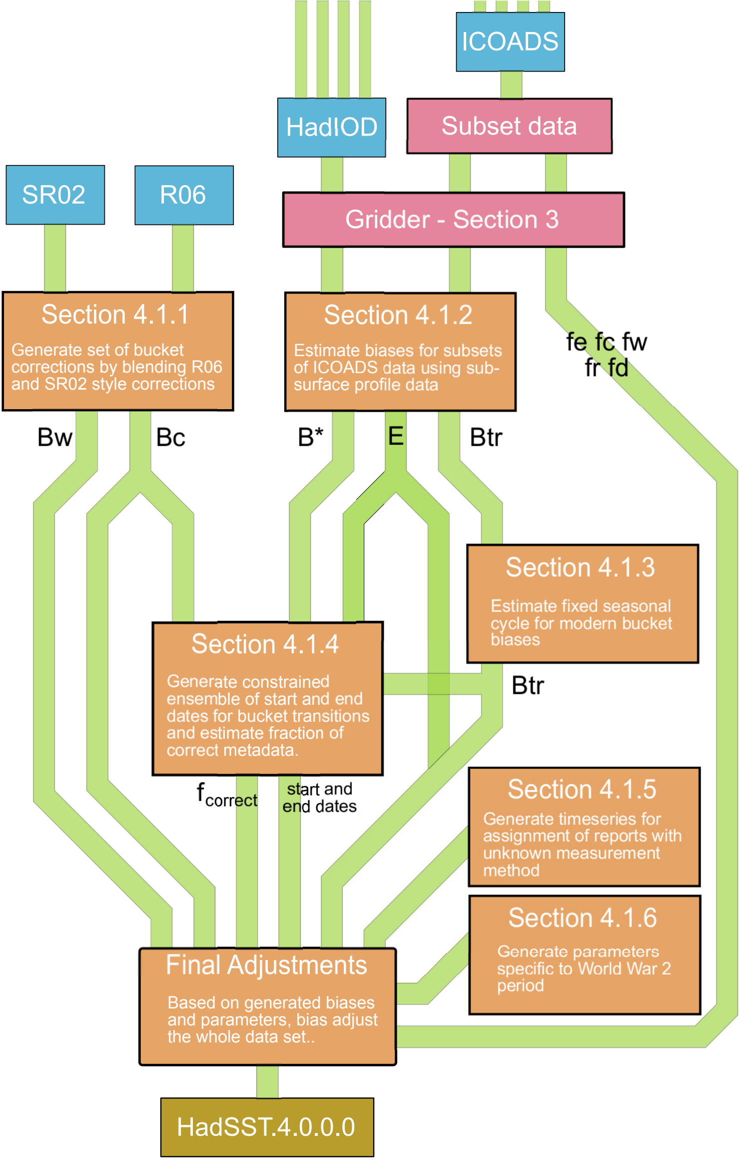
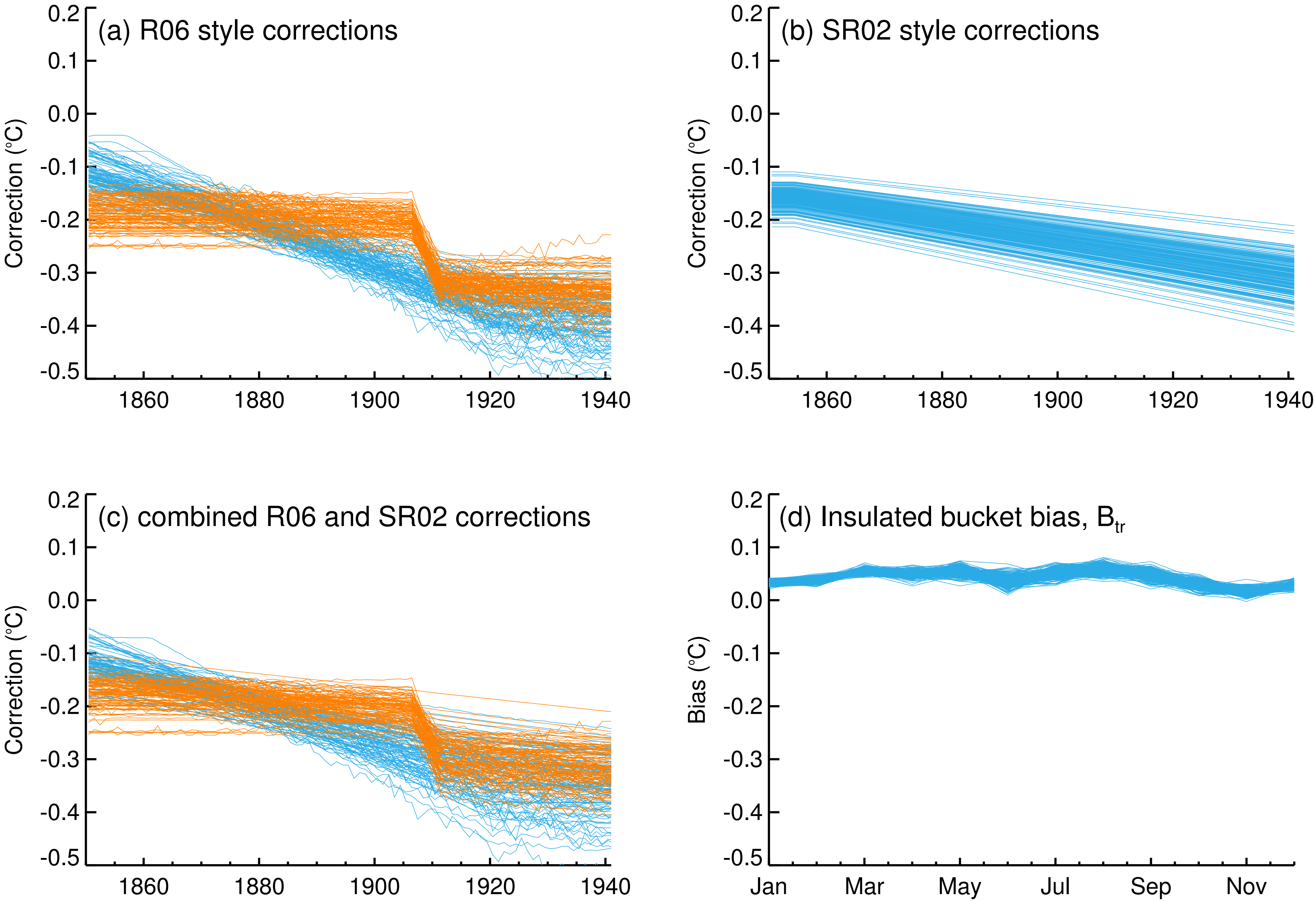

ICOADS
HadIOD
Subset data
SR02
R06
Gridder - Section 3
Section 4.1.1
Generate set of bucket corrections by blending R06 and SR02 style corrections
Section 4.1.2
Estimate biases for subsets of ICOADS data using sub-surface profile data
fe fc fw
fr fd
Bw
Bc
B*
E
Btr
Section 4.1.3
Estimate fixed seasonal cycle for modern bucket biases
Section 4.1.4
Generate constrained ensemble of start and end dates for bucket transitions and estimate fraction of correct metadata.
Btr
f correct
start and end dates
Section 4.1.5
Generate timeseries for assignment of reports with unknown measurement method
Section 4.1.6
Generate parameters specific to World War 2 period
Final Adjustments
Based on generated biases and parameters, bias adjust the whole data set..
HadSST.4.0.0.0

**Figure 5.**



(a) R06 style corrections
(b) SR02 style corrections
(c) combined R06 and SR02 corrections
(d) Insulated bucket bias, $B_{tr}$
Correction (°C)
Bias (°C)
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
1860
1880
1900
1920
1940
Jan
Mar
May
Jul
Sep
Nov

**Figure 6.**

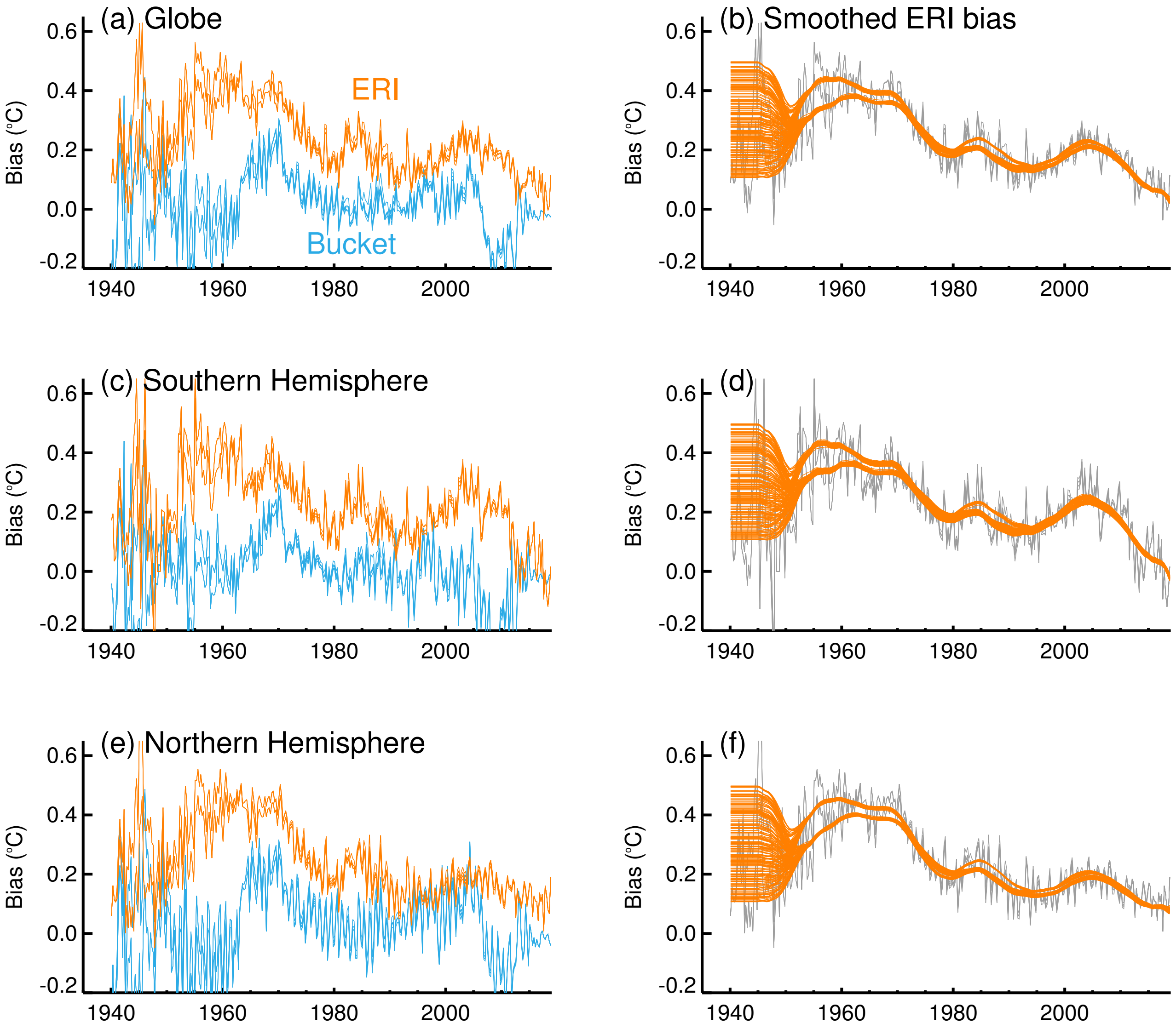
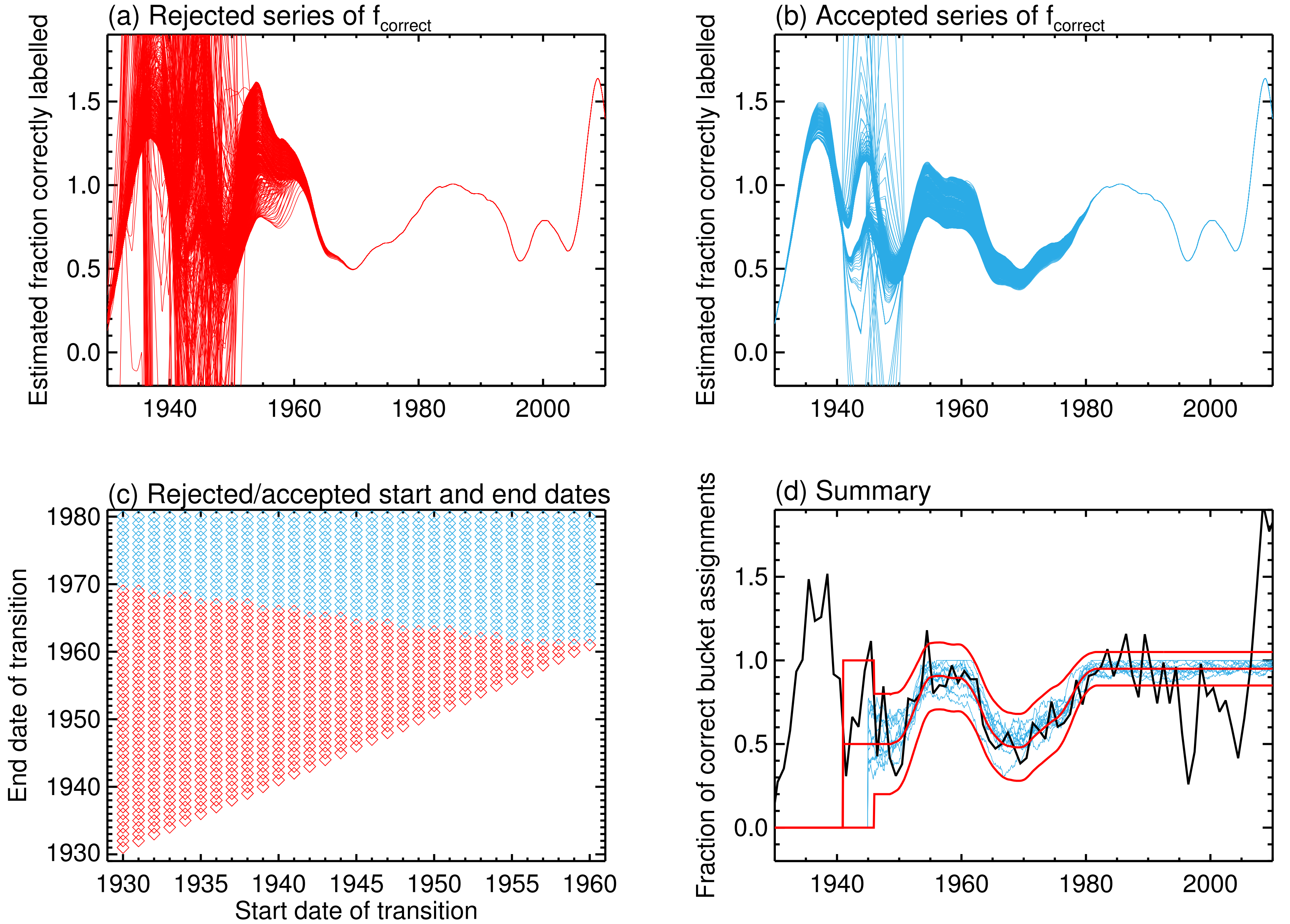

(a) Globe
ERI
Bucket
(b) Smoothed ERI bias
(c) Southern Hemisphere
(d)
(e) Northern Hemisphere
(f)
Bias (°C)
0.6
0.4
0.2
0.0
-0.2
1940
1960
1980
2000

**Figure 7.**



(a) Rejected series of $f_{correct}$
Estimated fraction correctly labelled
1.5
1.0
0.5
0.0
1940
1960
1980
2000
(b) Accepted series of $f_{correct}$
Estimated fraction correctly labelled
1.5
1.0
0.5
0.0
1940
1960
1980
2000
(c) Rejected/accepted start and end dates
End date of transition
1980
1970
1960
1950
1940
1930
1930
1935
1940
1945
1950
1955
1960
Start date of transition
(d) Summary
Fraction of correct bucket assignments
1.5
1.0
0.5
0.0
1940
1960
1980
2000

**Figure 8.**

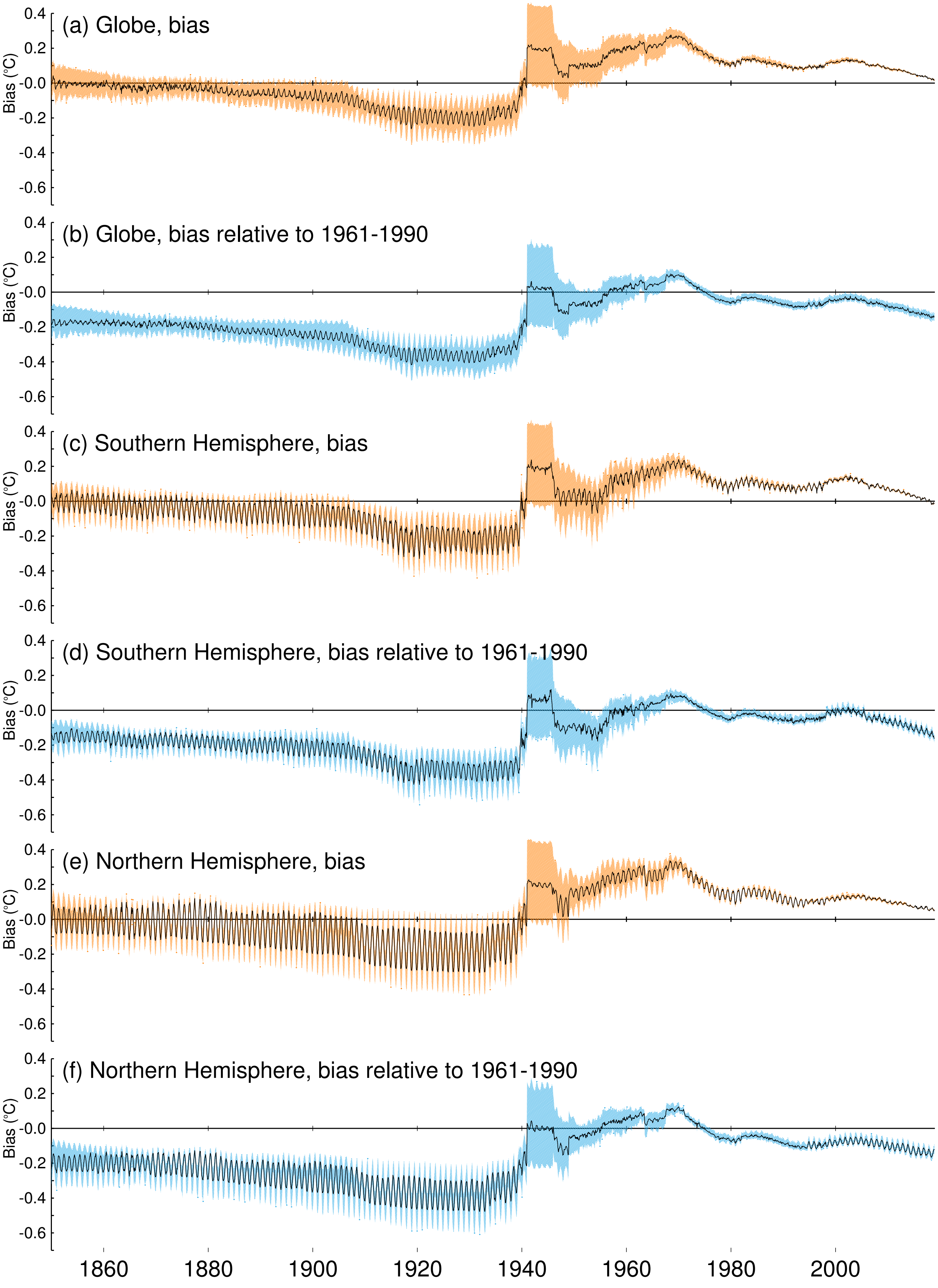

(a) Globe, bias
(b) Globe, bias relative to 1961-1990
(c) Southern Hemisphere, bias
(d) Southern Hemisphere, bias relative to 1961-1990
(e) Northern Hemisphere, bias
(f) Northern Hemisphere, bias relative to 1961-1990
Bias (°C)
0.4
0.2
-0.0
-0.2
-0.4
-0.6
1860
1880
1900
1920
1940
1960
1980
2000

**Figure 9.**

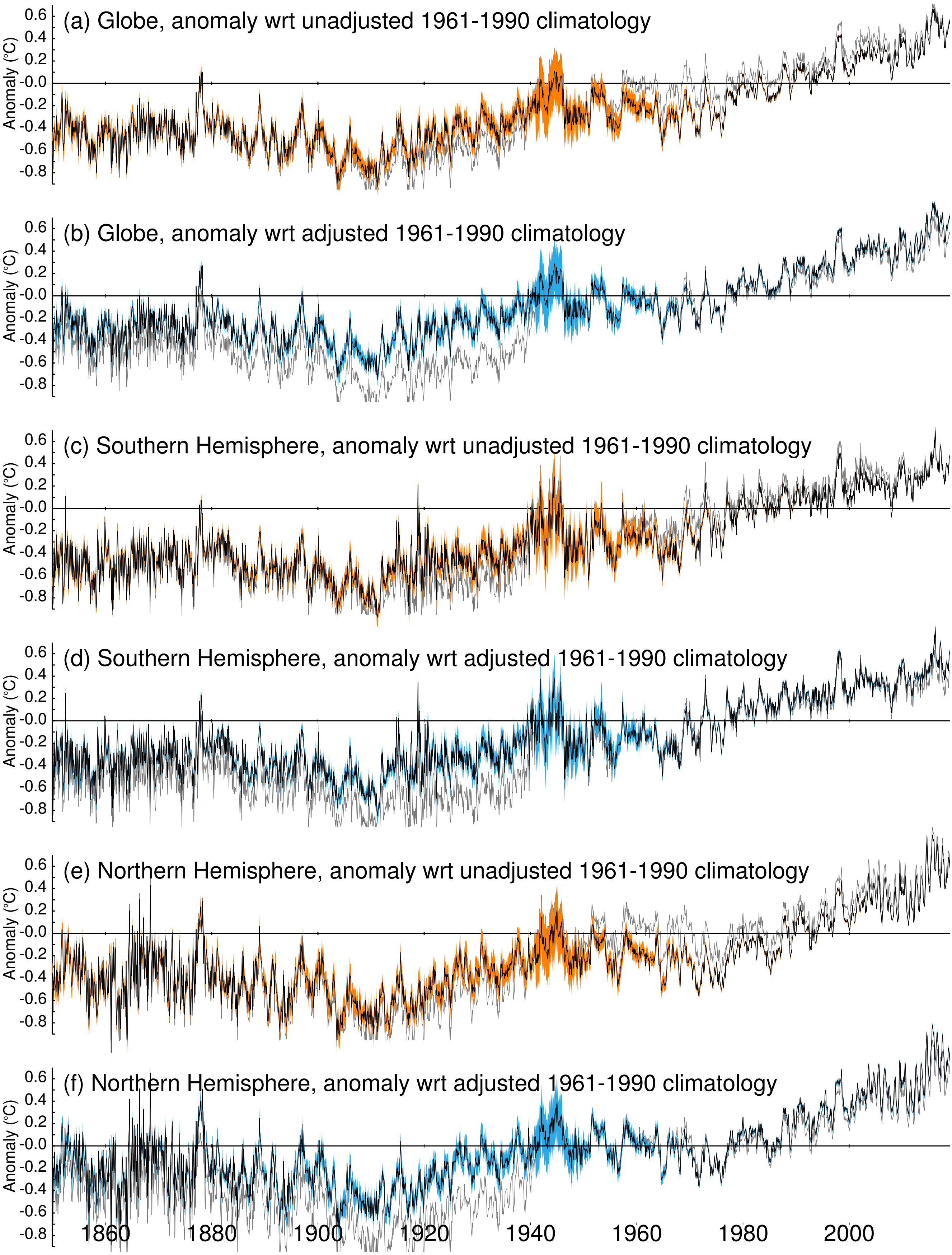
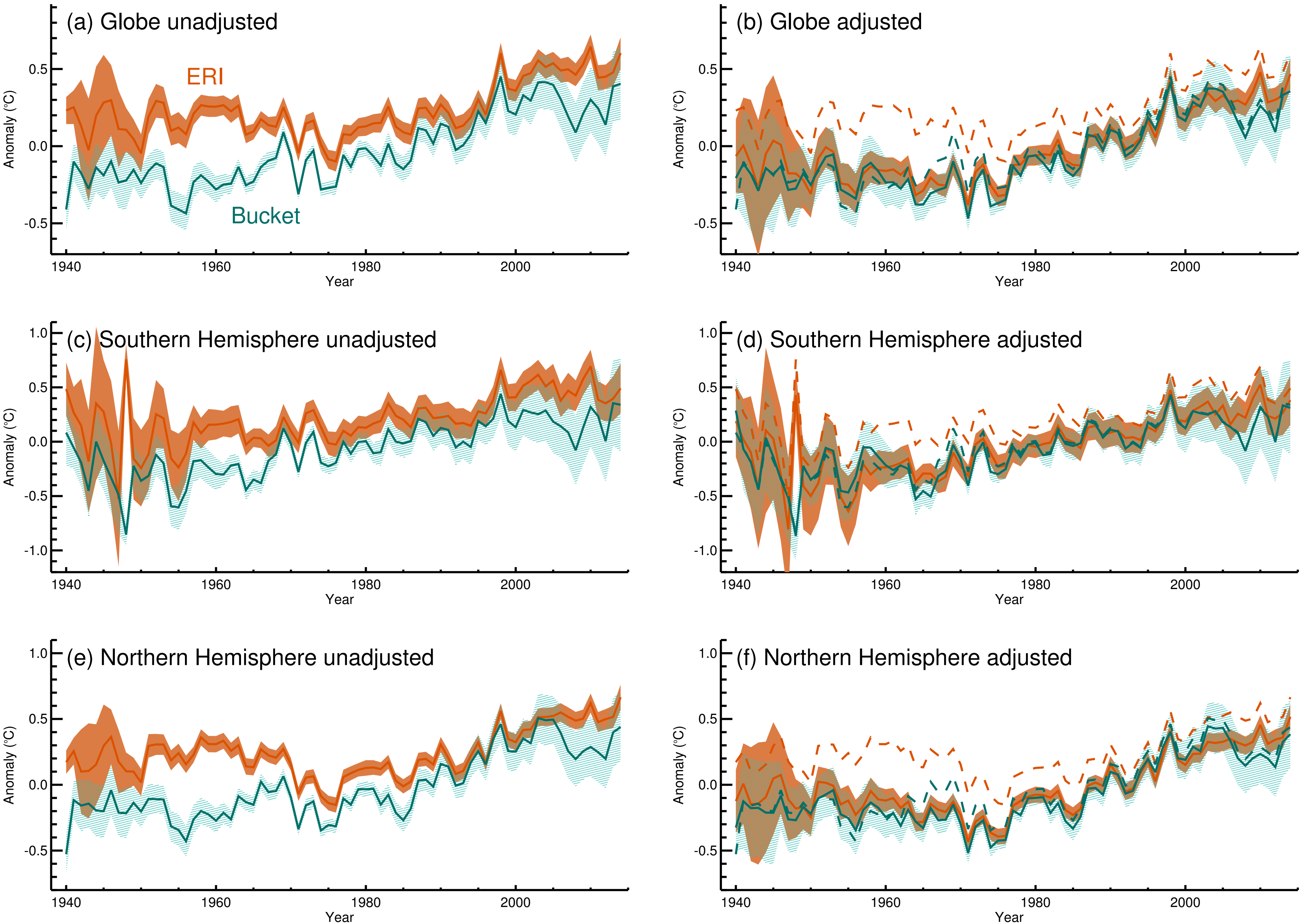

(a) Globe, anomaly wrt unadjusted 1961-1990 climatology
(b) Globe, anomaly wrt adjusted 1961-1990 climatology
(c) Southern Hemisphere, anomaly wrt unadjusted 1961-1990 climatology
(d) Southern Hemisphere, anomaly wrt adjusted 1961-1990 climatology
(e) Northern Hemisphere, anomaly wrt unadjusted 1961-1990 climatology
(f) Northern Hemisphere, anomaly wrt adjusted 1961-1990 climatology
Anomaly (°C)
0.6
0.4
0.2
-0.0
-0.2
-0.4
-0.6
-0.8
1860
1880
1900
1920
1940
1960
1980
2000

**Figure 10.**



(a) Globe unadjusted
ERI
Bucket
(b) Globe adjusted
(c) Southern Hemisphere unadjusted
(d) Southern Hemisphere adjusted
(e) Northern Hemisphere unadjusted
(f) Northern Hemisphere adjusted
Anomaly (°C)
Year
1940
1960
1980
2000
-1.0
-0.5
0.0
0.5
1.0

**Figure 11.**

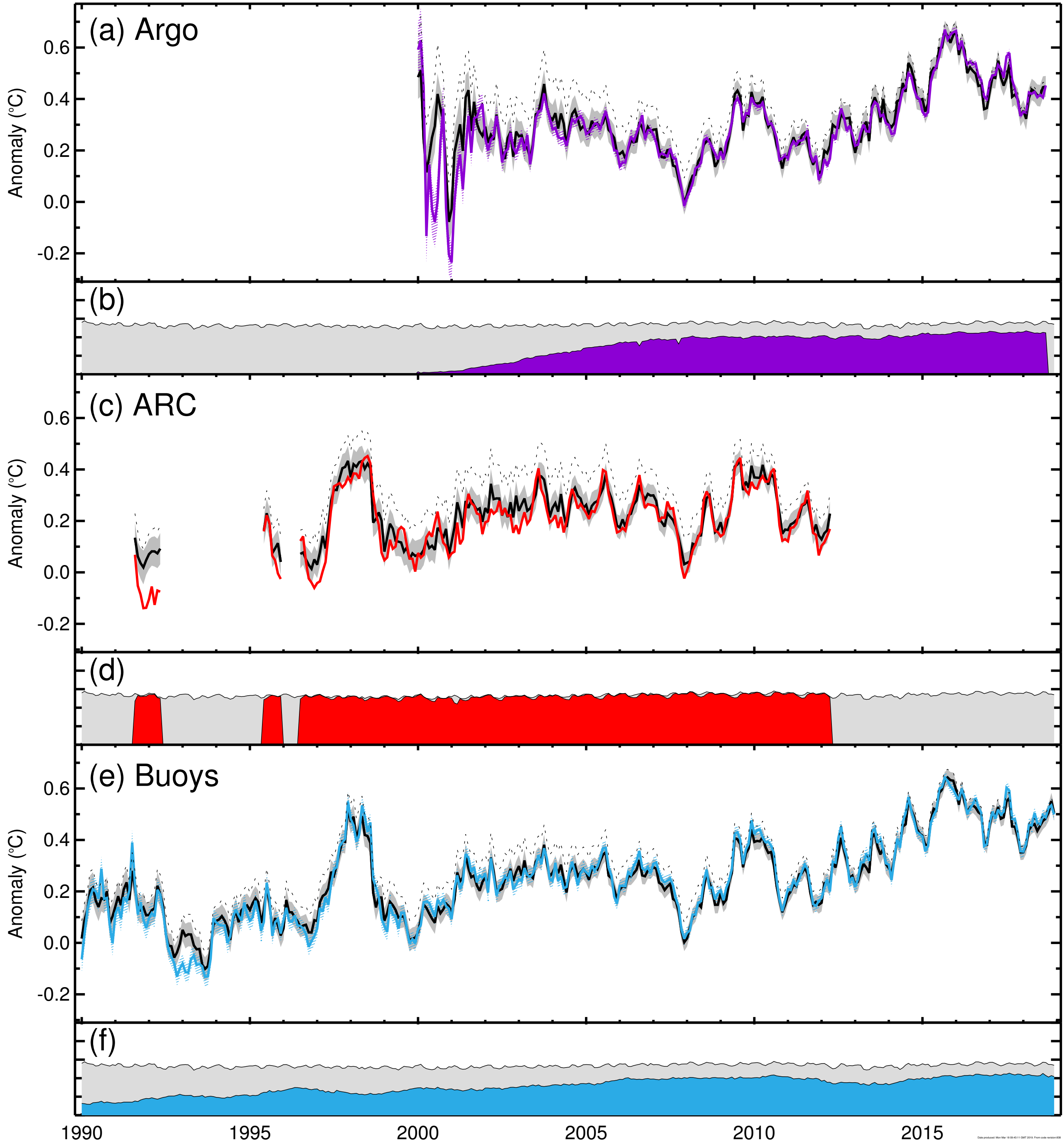

(a) Argo
0.6
0.4
0.2
0.0
-0.2
Anomaly (°C)
(b)
(c) ARC
Anomaly (°C)
(d)
(e) Buoys
Anomaly (°C)
(f)
1990
1995
2000
2005
2010
2015

**Figure 12.**

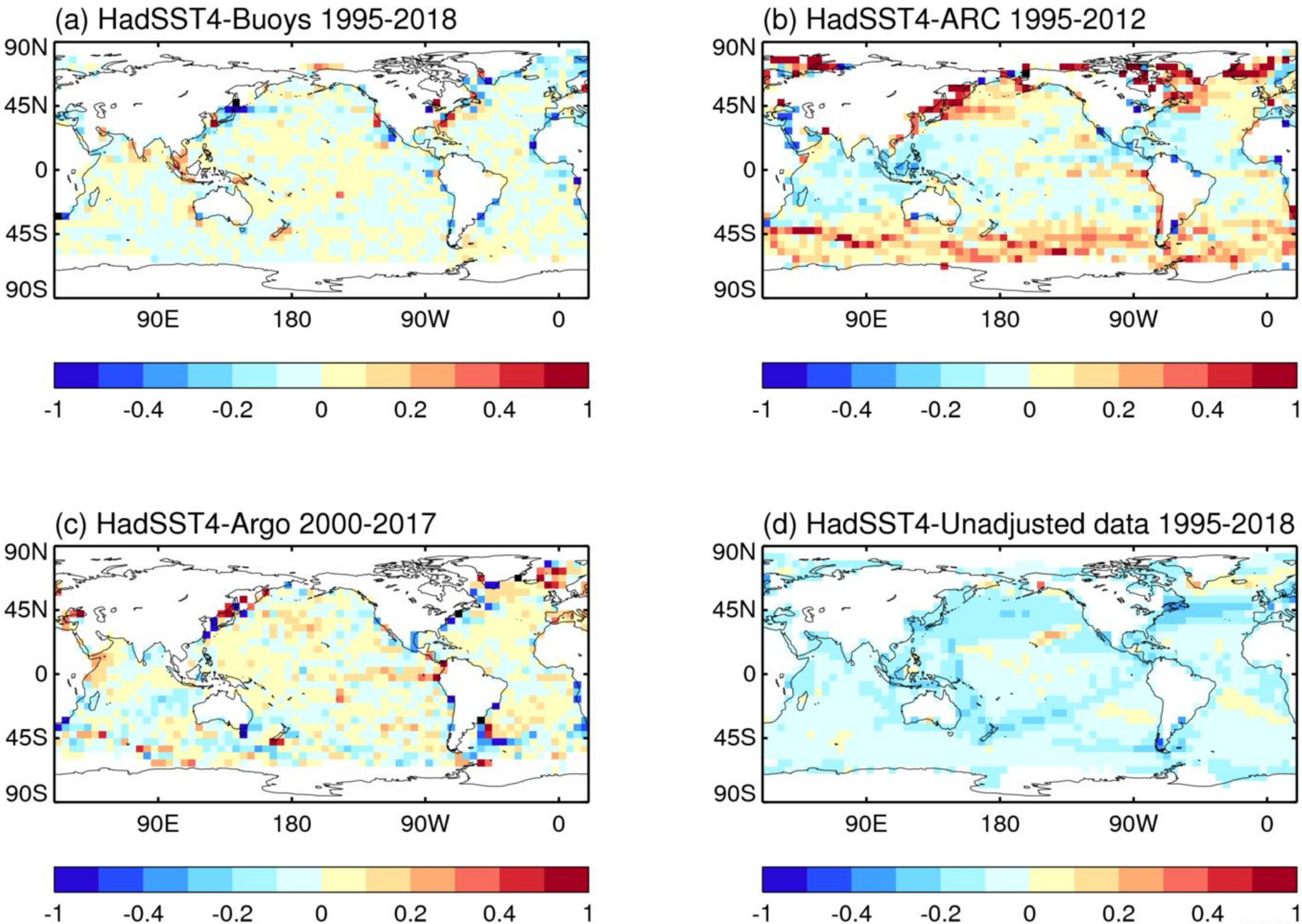

(a) HadSST4-Buoys 1995-2018
(b) HadSST4-ARC 1995-2012
(c) HadSST4-Argo 2000-2017
(d) HadSST4-Unadjusted data 1995-2018
90N
45N
0
45S
90S
90E
180
90W
0
-1
-0.4
-0.2
0
0.2
0.4
1

**Figure 13.**

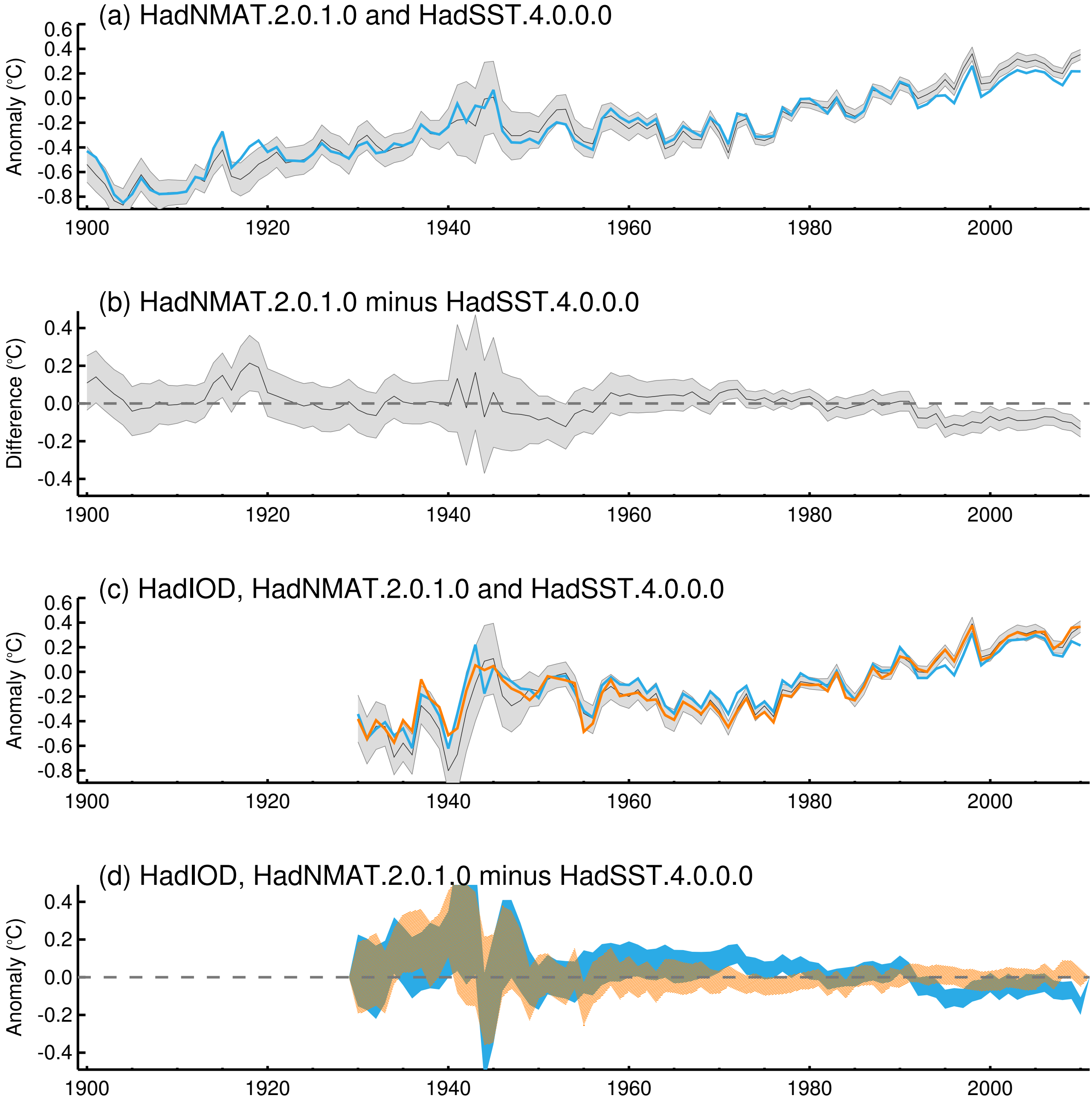
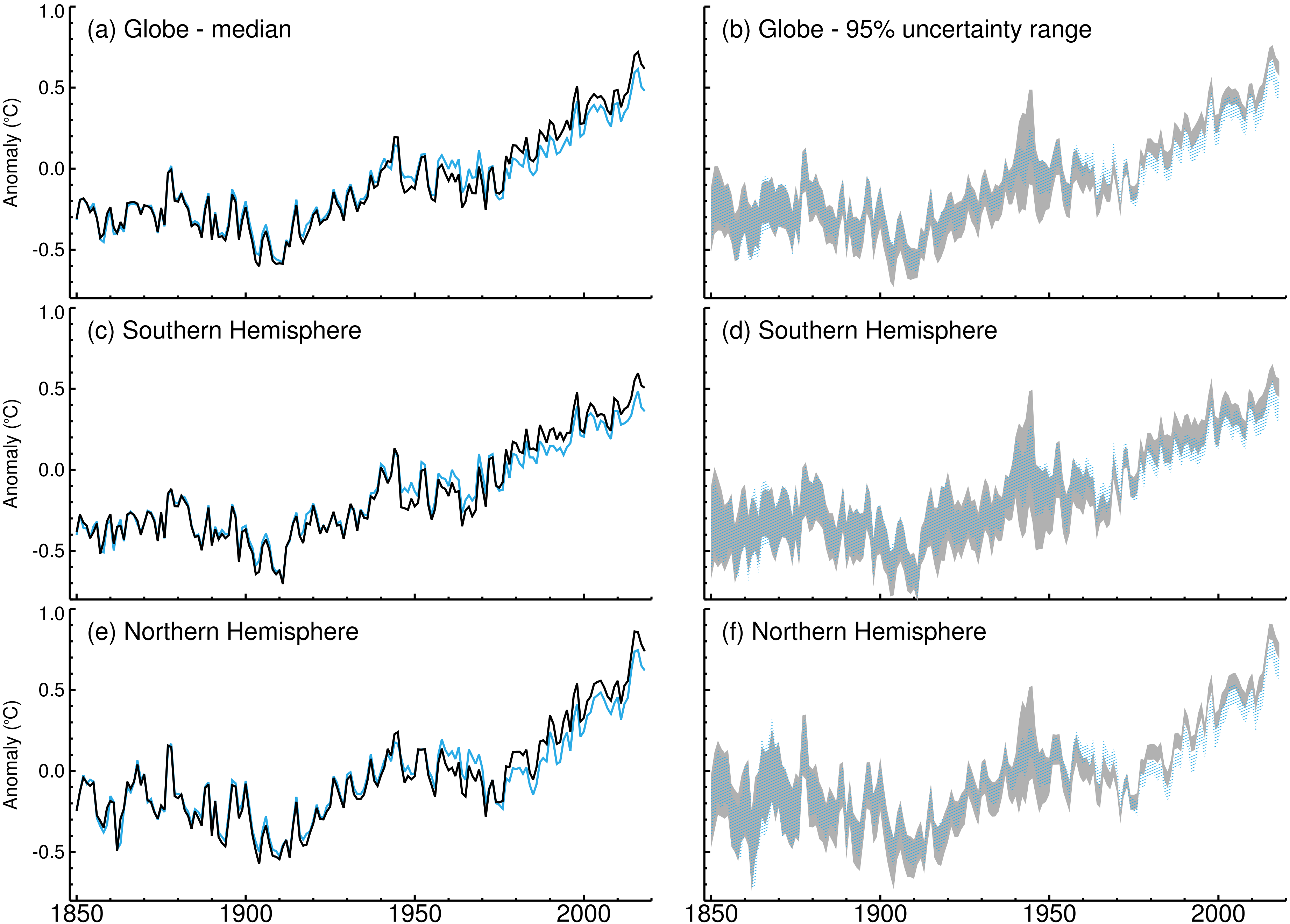
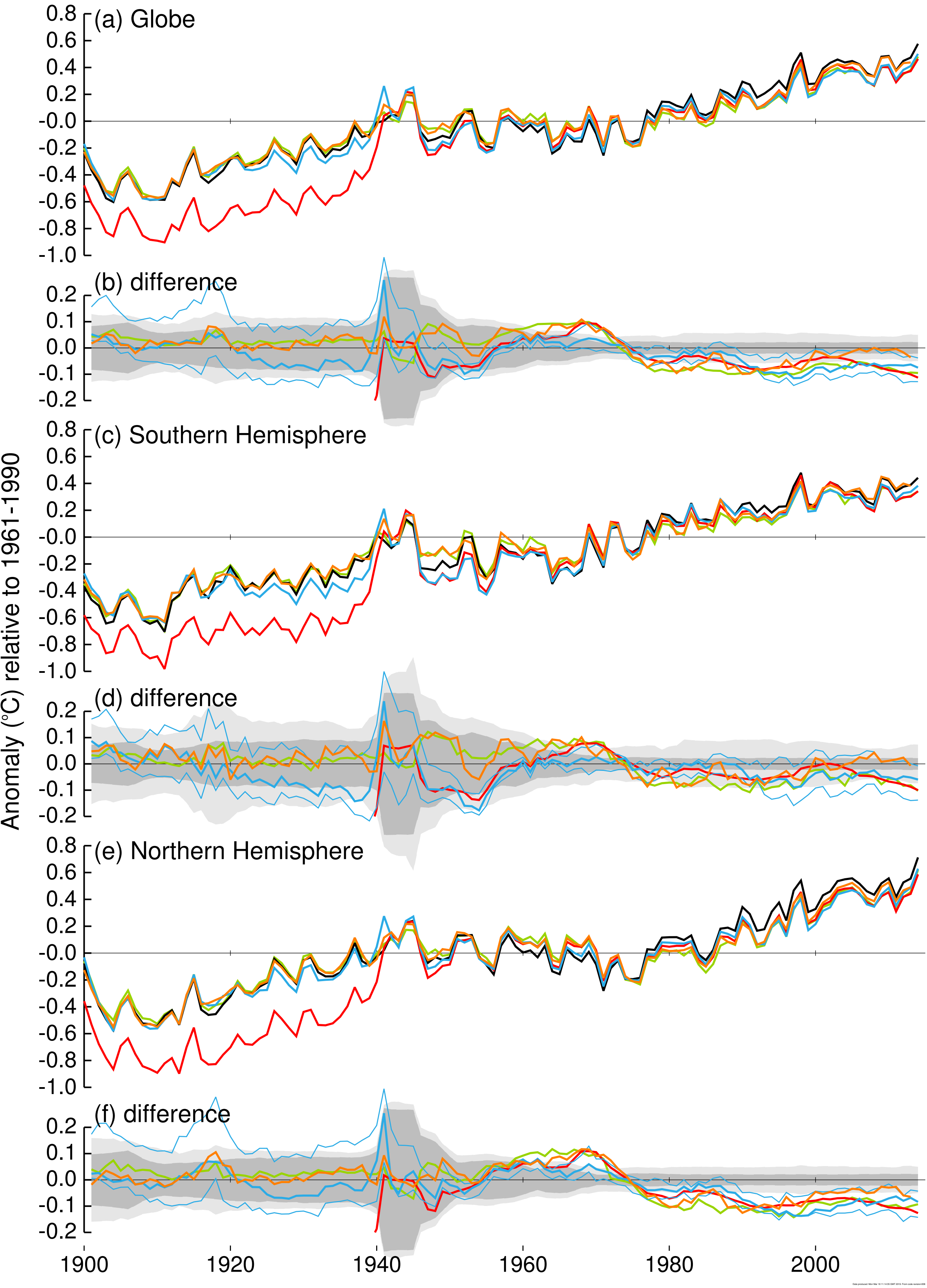

(a) HadNMAT.2.0.1.0 and HadSST.4.0.0.0
Anomaly (°C)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
1900
1920
1940
1960
1980
2000
(b) HadNMAT.2.0.1.0 minus HadSST.4.0.0.0
Difference (°C)
0.4
0.2
0.0
-0.2
-0.4
1900
1920
1940
1960
1980
2000
(c) HadIOD, HadNMAT.2.0.1.0 and HadSST.4.0.0.0
Anomaly (°C)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
1900
1920
1940
1960
1980
2000
(d) HadIOD, HadNMAT.2.0.1.0 minus HadSST.4.0.0.0
Anomaly (°C)
0.4
0.2
0.0
-0.2
-0.4
1900
1920
1940
1960
1980
2000

**Figure 14.**



(a) Globe - median
(b) Globe - 95% uncertainty range
(c) Southern Hemisphere
(d) Southern Hemisphere
(e) Northern Hemisphere
(f) Northern Hemisphere
Anomaly (°C)
Anomaly (°C)
Anomaly (°C)
1.0
0.5
0.0
-0.5
1850
1900
1950
2000

**Figure 15.**



(a) Globe
(b) difference
(c) Southern Hemisphere
(d) difference
(e) Northern Hemisphere
(f) difference
Anomaly (°C) relative to 1961-1990
0.8
0.6
0.4
0.2
-0.0
-0.2
-0.4
-0.6
-0.8
-1.0
0.2
0.1
0.0
-0.1
-0.2
1900
1920
1940
1960
1980
2000

**Figure 16.**

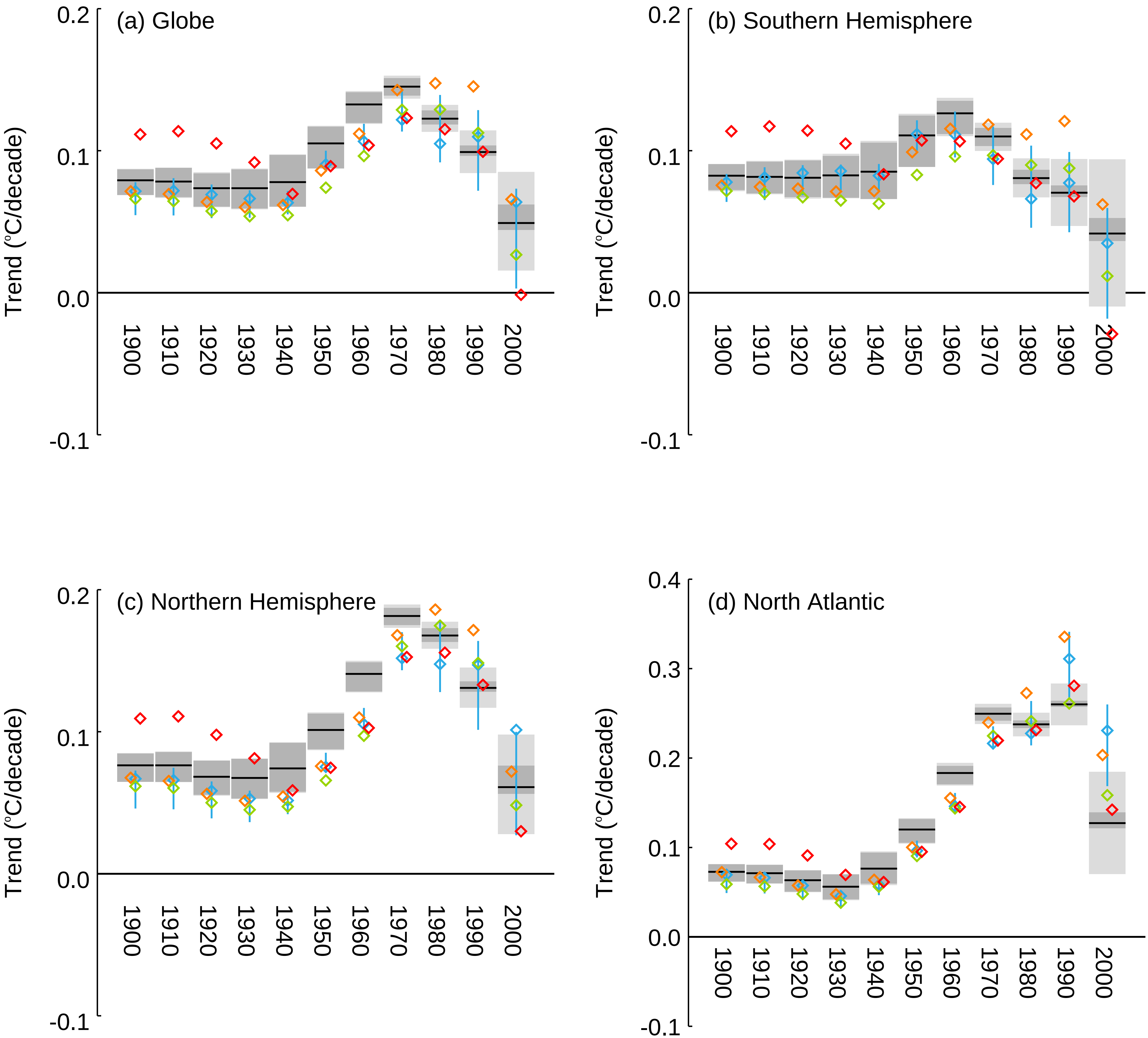

(a) Globe
(b) Southern Hemisphere
(c) Northern Hemisphere
(d) North Atlantic
Trend (°C/decade)
0.4
0.3
0.2
0.1
0.0
-0.1
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000

**Figure 17.**

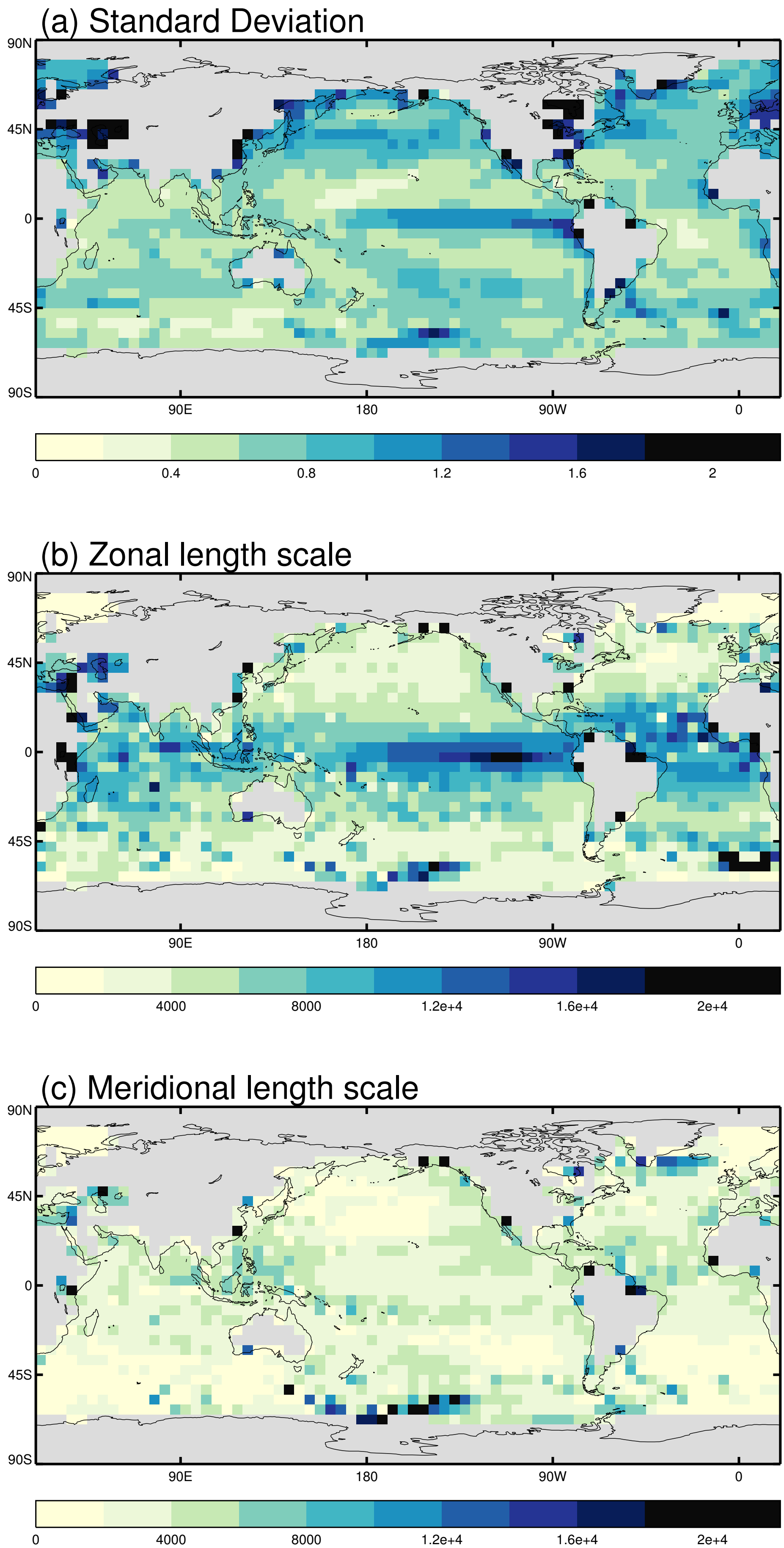

(a) Standard Deviation
90N
45N
0
45S
90S
90E
180
90W
0
0
0.4
0.8
1.2
1.6
2
(b) Zonal length scale
90N
45N
0
45S
90S
90E
180
90W
0
0
4000
8000
1.2e+4
1.6e+4
2e+4
(c) Meridional length scale
90N
45N
0
45S
90S
90E
180
90W
0
0
4000
8000
1.2e+4
1.6e+4
2e+4

**Figure 18.**

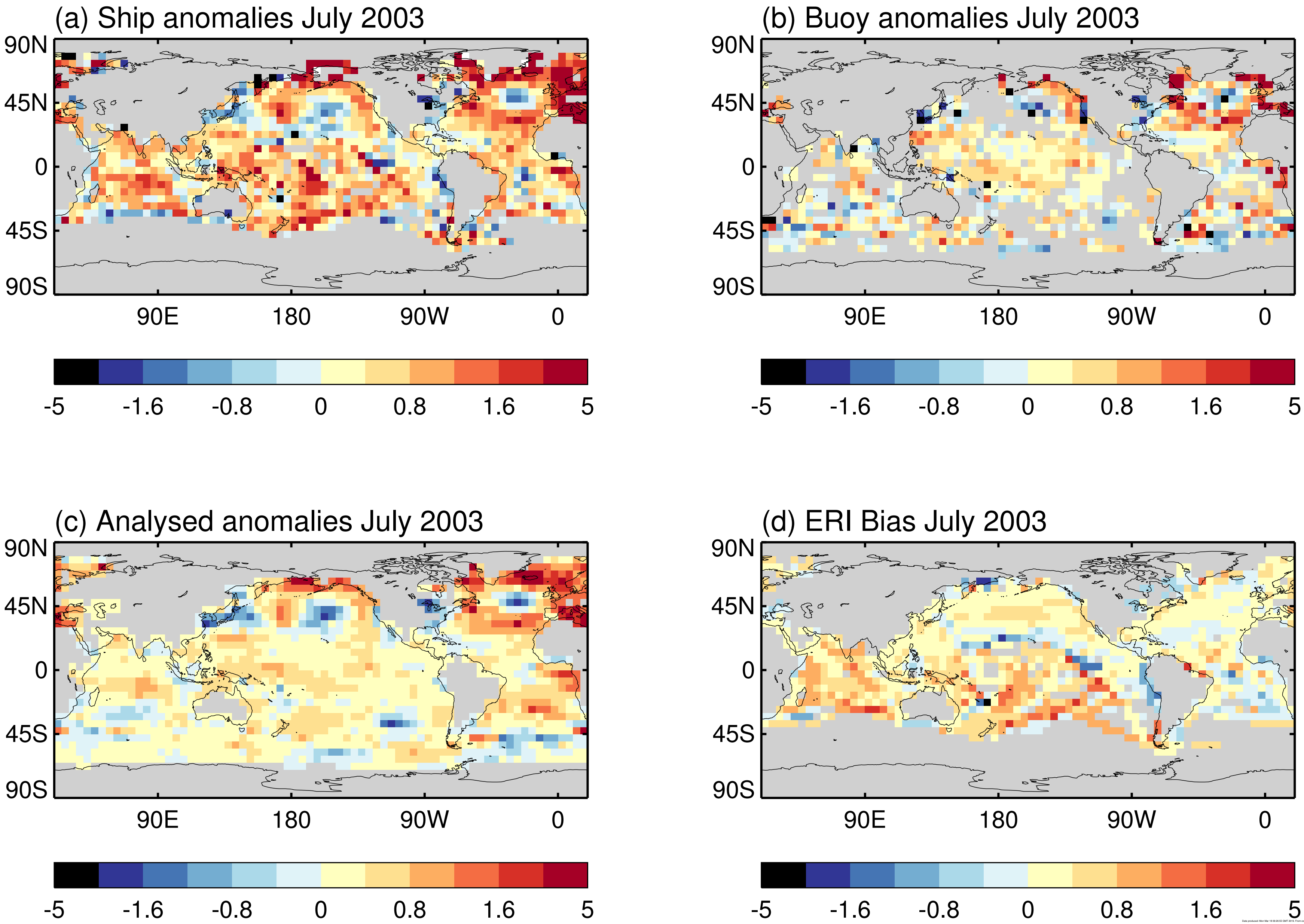

(a) Ship anomalies July 2003
(b) Buoy anomalies July 2003
(c) Analysed anomalies July 2003
(d) ERI Bias July 2003
90N
45N
0
45S
90S
90E
180
90W
0
-5
-1.6
-0.8
0
0.8
1.6
5

**Figure 19.**

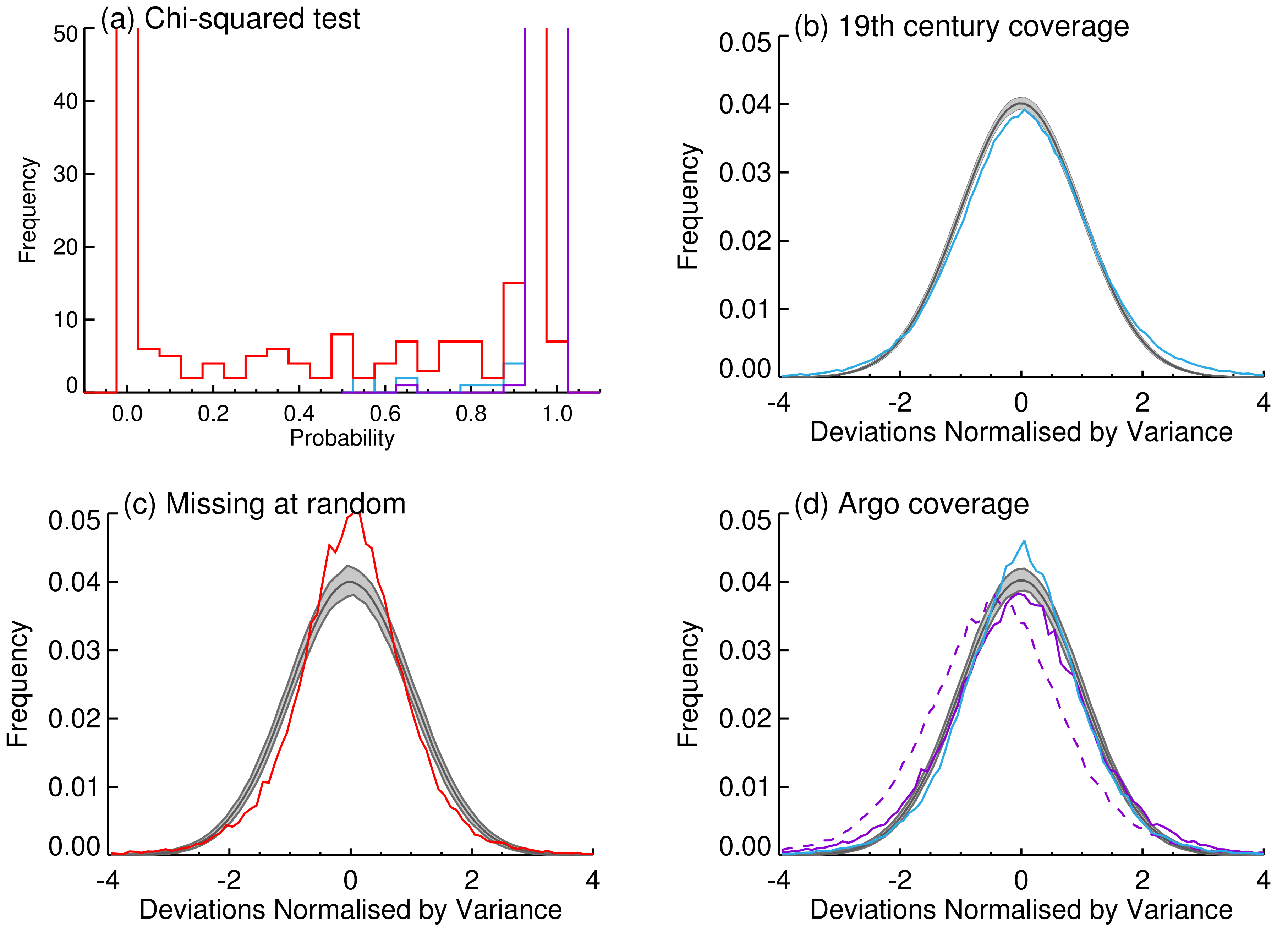
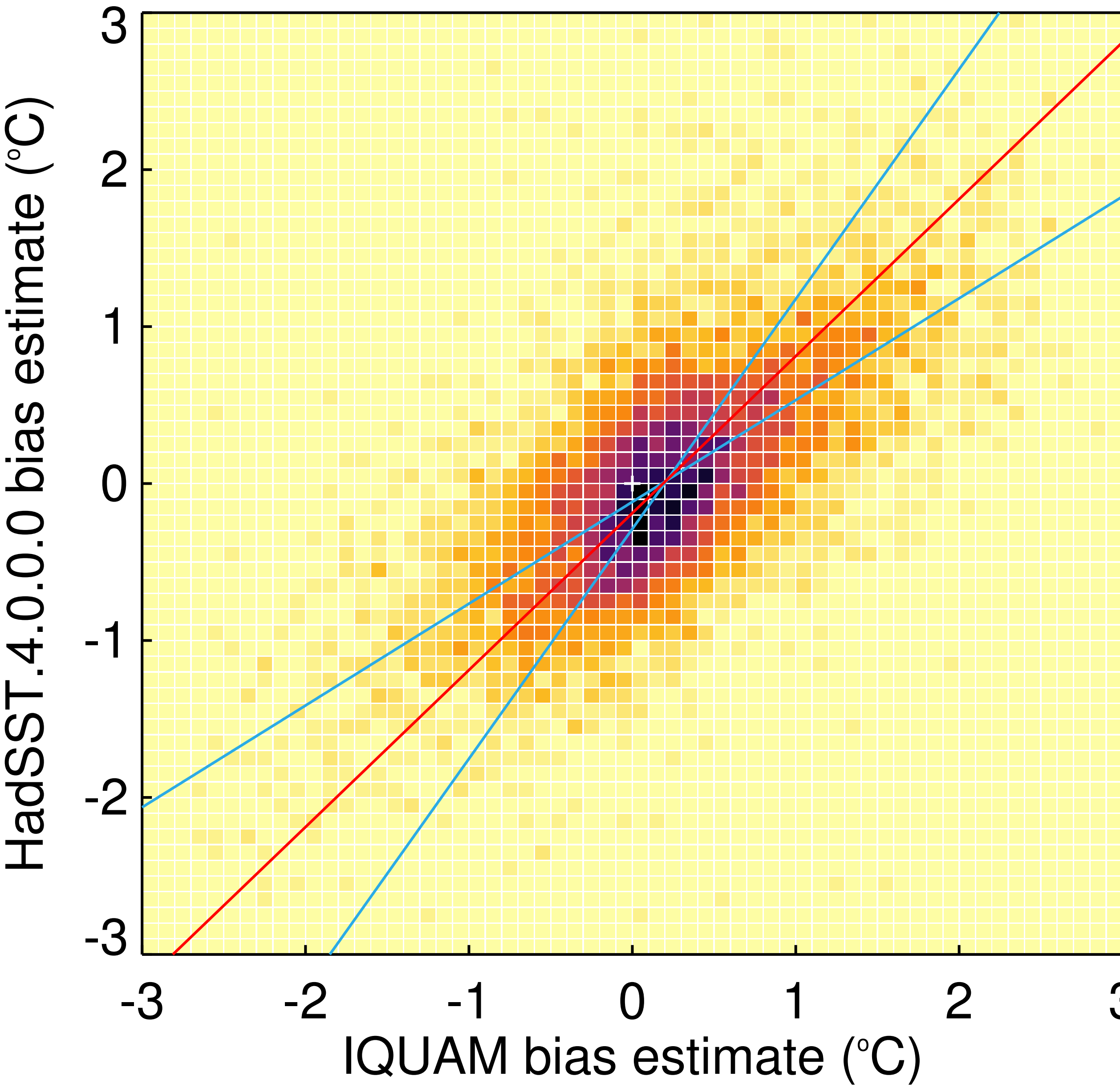
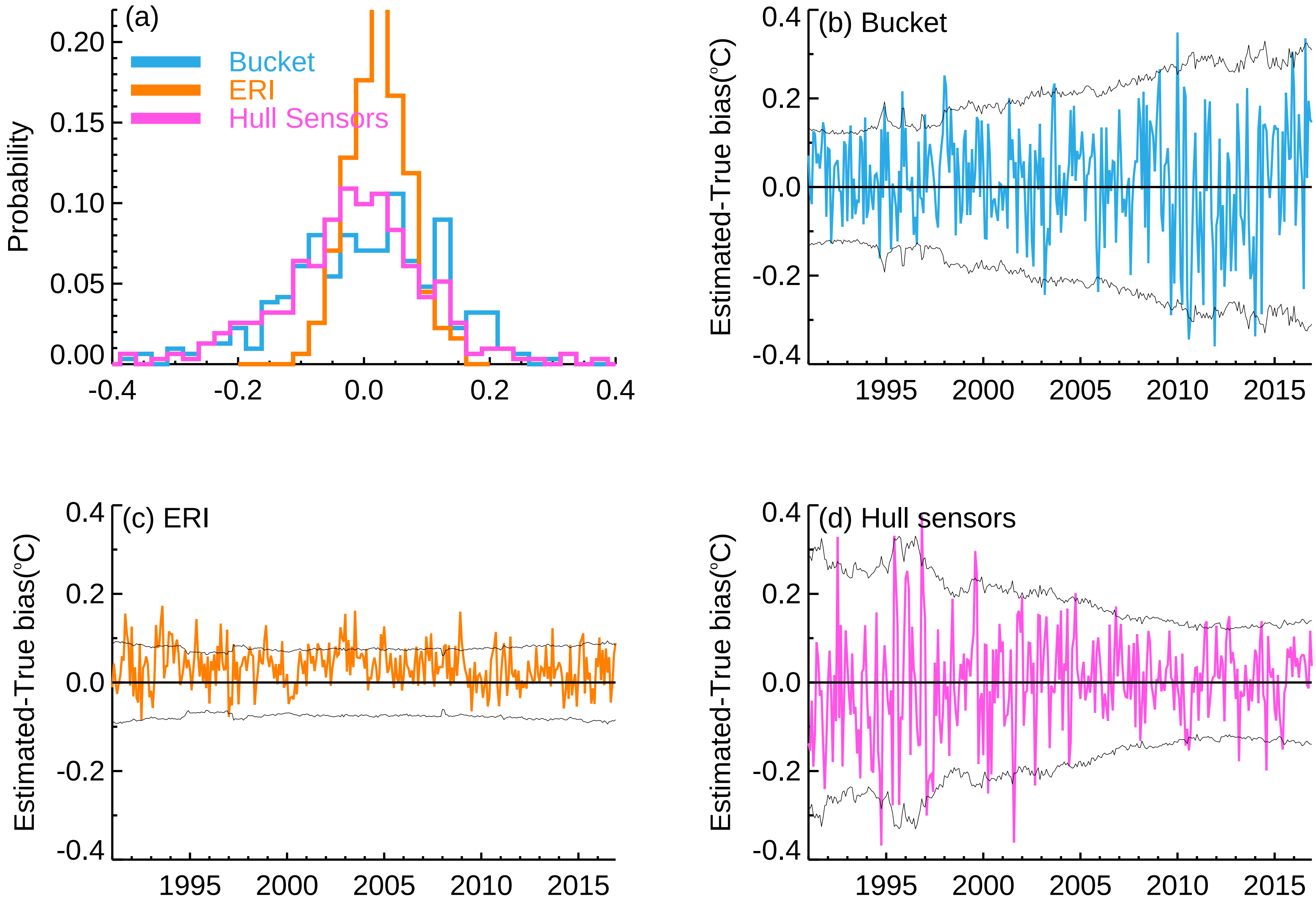

(a) Chi-squared test
Frequency
Probability
(b) 19th century coverage
Frequency
Deviations Normalised by Variance
(c) Missing at random
Frequency
Deviations Normalised by Variance
(d) Argo coverage
Frequency
Deviations Normalised by Variance

**Figure 20.**



3
2
1
0
-1
-2
-3
HadSST.4.0.0.0 bias estimate (°C)
-3
-2
-1
0
1
2
3
IQUAM bias estimate (°C)

**Figure 21.**



(a)
Bucket
ERI
Hull Sensors
Probability
0.20
0.15
0.10
0.05
0.00
-0.4
-0.2
0.0
0.2
0.4
(b) Bucket
Estimated-True bias(°C)
0.4
0.2
0.0
-0.2
-0.4
1995
2000
2005
2010
2015
(c) ERI
Estimated-True bias(°C)
0.4
0.2
0.0
-0.2
-0.4
1995
2000
2005
2010
2015
(d) Hull sensors
Estimated-True bias(°C)
0.4
0.2
0.0
-0.2
-0.4
1995
2000
2005
2010
2015

**Figure 22.**

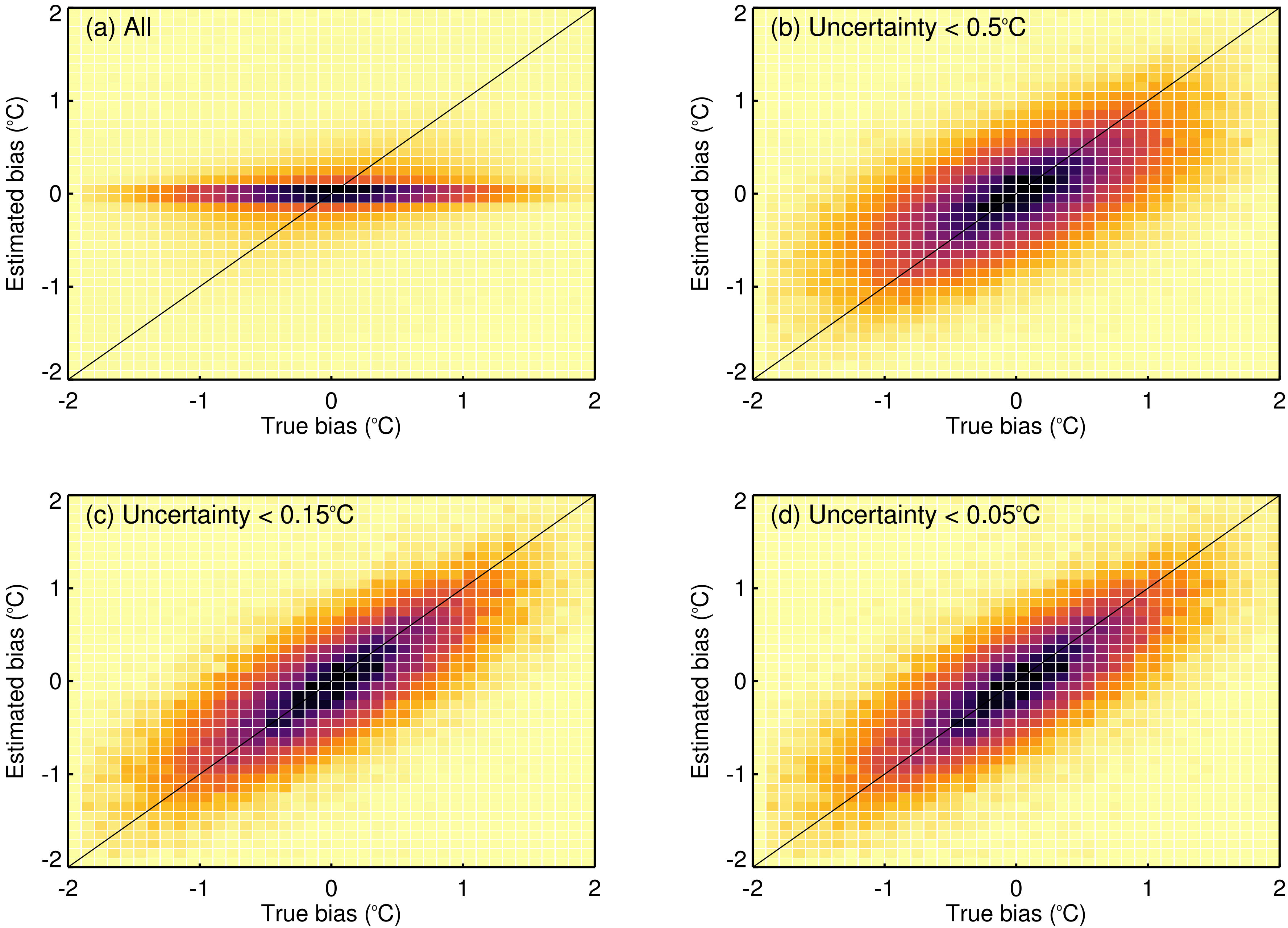
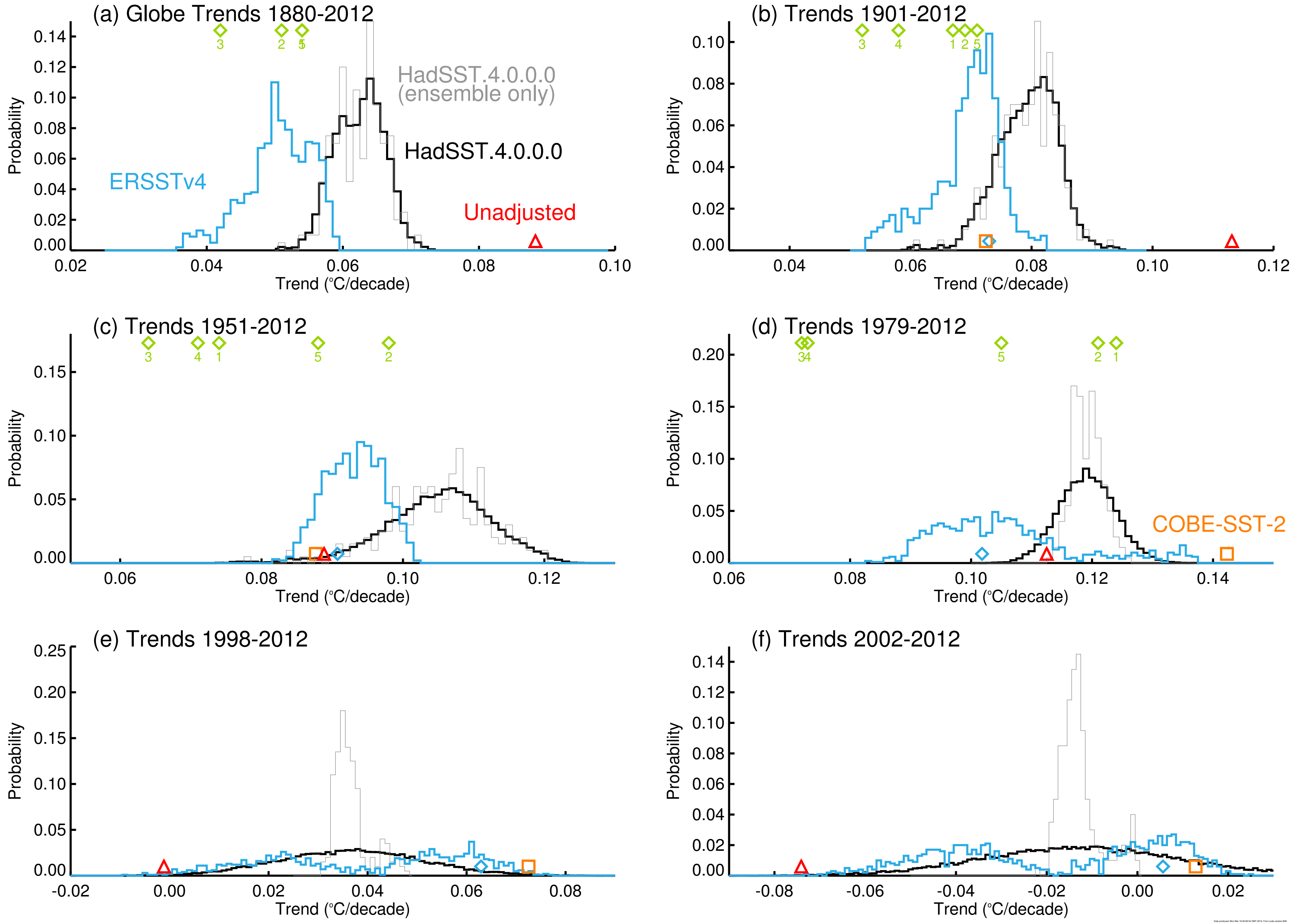

(a) All
(b) Uncertainty < 0.5°C
(c) Uncertainty < 0.15°C
(d) Uncertainty < 0.05°C
Estimated bias (°C)
True bias (°C)
-2
-1
0
1
2

**Figure 23.**


(a) Globe Trends 1880-2012
HadSST.4.0.0.0
(ensemble only)
HadSST.4.0.0.0
ERSSTv4
Unadjusted
Probability
Trend (°C/decade)
(b) Trends 1901-2012
Probability
Trend (°C/decade)
(c) Trends 1951-2012
Probability
Trend (°C/decade)
(d) Trends 1979-2012
COBE-SST-2
Probability
Trend (°C/decade)
(e) Trends 1998-2012
Probability
Trend (°C/decade)
(f) Trends 2002-2012
Probability
Trend (°C/decade)