SUPPLEMENT

A CALL FOR NEW APPROACHES TO QUANTIFYING BIASES IN OBSERVATIONS OF SEA SURFACE TEMPERATURE

Elizabeth C. Kent, John J. Kennedy, Thomas M. Smith, Shoji Hirahara, Boyin Huang, Alexey Kaplan, David E. Parker, Christopher P. Atkinson, David I. Berry, Giulia Carella, Yoshikazu Fukuda, Masayoshi Ishii, Philip D. Jones, Finn Lindgren, Christopher J. Merchant, Simone Morak-Bozzo, Nick A. Rayner, Victor Venema, Souichiro Yasui, and Huai-Min Zhang

This document is a supplement to "A Call for New Approaches to Quantifying Biases in Observations of Sea Surface Temperature," by Elizabeth C. Kent, John J. Kennedy, Thomas M. Smith, Shoji Hirahara, Boyin Huang, Alexey Kaplan, David E. Parker, Christopher P. Atkinson, David I. Berry, Giulia Carella, Yoshikazu Fukuda, Masayoshi Ishii, Philip D. Jones, Finn Lindgren, Christopher J. Merchant, Simone Morak-Bozzo, Nick A. Rayner, Victor Venema, Souichiro Yasui, and Huai-Min Zhang (*Bull. Amer. Meteor. Soc.,* **98,** 1601–1616) • ©2017 American Meteorological Society • *Corresponding author:* Elizabeth C. Kent, eck@noc.ac.uk • DOI:10.1175/BAMS-D-15-00251.2

Sections:

S1: SEA SURFACE TEMPERATURE (SST), NIGHT MARINE AIR TEMPERATURE (NMAT), AND THE GLOBAL SURFACE TEMPERATURE RECORD. The global surface temperature record is constructed from observations of air temperature over land or ice and sea surface temperature (SST) over the oceans (e.g., Morice et al. 2012; Karl et al. 2015). For consistency with the land observations, marine air temperature (MAT) would seem the obvious choice in preference to SST. One reason that SST has traditionally been used is due to the large thermal inertia of the ocean surface layer, which dampens the temporal variability of SST compared to that of the air temperature. This means that for the same number of observations, SST averages are more consistent than for MAT. Also, observations of MAT from ships are affected by the influence of daytime heating of the ship and sensor environment (Glahn 1933; Berry et al. 2004), so daytime observations are normally excluded and gridded analyses of night marine air temperature (NMAT) are constructed (Rayner et al. 2003; Kent et al. 2013). Because air temperature observations are more variable than SST, and only about half of the available MAT observations can be used, SST has been considered a better choice for constructing a global surface temperature record from a sampling perspective. In recent times, the coverage of NMAT observations has declined (Berry and Kent 2017) due to the rapid decline in voluntary observing ships (VOSs), whereas SST observation numbers have increased greatly, mainly due to the contribution from drifting buoys.

The adjustments required to standardize NMAT measurements to a reference level, often 10 m, are fairly well known (Kent et al. 2013) if the height of the measurement is available. Information on measurement height was introduced into World Meteorological Organization (WMO) Publication No. 47 (Publ. 47) in 1968 (Kent et al. 2007) and from that time it has been possible to make adjustments from a known height for those observations that can be linked to metadata in Publ. 47. For observations with no height information a default value based on monthly gridded known height values can be used. The earliest observations are believed to have been taken at about 6 m above sea level (Rayner et al. 2003) and are therefore typically biased warm compared to the 10-m reference level, as the sea surface is typically warmer than the air above it. NMAT measurement heights typically increase over time, as ships have increased in size. In the Hadley Centre and National Oceanography Centre Night Marine Air Temperature Data Set, version 2 (HadNMAT2), gridded NMAT analysis (Kent et al. 2013), a reduction in height during World War II (WWII) was assumed in order to account for a probable reduction in size of ships at this time due to wartime losses (Kent et al. 2013), adding to the uncertainty during this already uncertain period. Release 2.5 of the International Comprehensive Ocean–Atmosphere Data Set (ICOADS) issued in July 2009 (Woodruff et al. 2011; see section S3) did not contain ship call signs after 2007 due to concerns of ship operators. Thus, air temperature measurement height has not been linked to observations since that time, and the height adjustment has become more uncertain (Kent et al. 2013) and recent changes in typical heights have not been properly accounted for. The most recent release of ICOADS (release 3.0; Freeman et al. 2017) has reinstated some call signs, which permits linking with Publ. 47 metadata post-2007. However, encryption and the masking of call signs are making the task of constructing a well-documented climate record much more complicated.

Global-average height adjustments in HadNMAT2 vary from a reduction of slightly over 0.06°C in the 1850s to an average increase of nearly 0.17°C by 2010, thereby increasing the global increase over this period by about 0.23°C. As with SST bias adjustments, the NMAT height adjustment requires information on environmental conditions, in this case the stability of the lower atmosphere that defines the temperature gradient within the surface layer of the atmosphere. In HadNMAT2 stability is estimated from climatological monthly joint distributions of air–sea temperature difference and wind speed combined with climatological SST and humidity from the National Oceanography Centre (NOC) Surface Flux and Meteorological Dataset, version 2 (Berry and Kent 2009, 2011). Uncertainty in the height adjustment is derived by combining the estimated uncertainty in the measurement height with that due to variability in atmospheric stability.

Further adjustments are required to account for nonstandard observing practices, for example, during WWII when thermometers were probably read under cover at night to avoid using lights on deck (Rayner et al. 2003). SST is also very uncertain during the WWII period due to similar nonstandard practices in response to the war.

Ideally, all-hours MAT rather than NMAT would be used, but presently only MAT observations from 1970 onward have been adjusted for heating biases (Berry et al. 2004; Berry and Kent 2011)

Taken together the expectation is that although large-scale, long-term biases in NMAT are probably better understood and easier to adjust than those in SST, increased uncertainty on regional scales due

to higher variability and lower sampling for NMAT means that SST is the preferred variable for the marine component of global surface temperature. NMAT, after adjustment, is believed to be more stable than SST at large scales. The approach taken therefore is to construct the marine portion of the global surface temperature from SST, bias adjusted using either large-scale-adjusted NMAT (Huang et al. 2015) or physically based models that are more weakly linked to NMAT (Kennedy et al. 2011b; Hirahara et al. 2014).

S2: UNCERTAINTY AND BIAS IN MEASUREMENTS OF SST. *The definition of SST.* The traditional target of in situ measurements, "bulk" SST, is taken to be the temperature of the top several meters of the ocean, assuming that the top several meters are well mixed. Indeed, it is common in oceanography to refer to a near-surface mixed layer.

This assumption, while generally valid for nighttime conditions or when surface winds are present, does not hold during the daytime under calm conditions when a stratified warm surface layer (diurnal warming layer) develops (Kawai and Wada 2007). Even though such layers start forming at the ocean surface, manifest their largest warming there, and often remain confined close to the surface, they may extend down to depths of several meters under favorable conditions (i.e., calm and sunny). This phenomenon is apparent in some of the mooring temperature time series (e.g., Kennedy 2014). Since diurnal warming layers form during the daytime and are destroyed at night by vertical mixing due to convection (when surface cooling causes the surface water to become denser than the underlying layers), in situ SST measurements in such cases can be significantly affected by the time of day as well.

Satellite sensors measure SST in a very thin surface layer. Microwave measurements are sensitive to temperature variations in the upper millimeters of the ocean. Infrared measurements sense radiation emitted by the upper micrometers of the ocean. In these very near-surface layers, evaporation causes a "cool skin" effect that must be accounted for in addition to near-surface diurnal warming (Minnett and Corlett 2012).

The best way to define SST is unclear. In a traditional idealistic paradigm, there would be a surface mixed layer of a constant temperature and slowly (on a scale of a few days or longer) changing depth; the temperature of this layer would be called SST. This definition breaks down when diurnal warming layers are present. To address this, the satellite SST community defined a foundation SST (SST_f) as the temperature at the first time of the day when the heat gain from solar radiation exceeds the heat loss at the sea surface (Donlon et al. 2007). However, estimation of SST_f requires measurement of the surface temperature profile and the surface heat fluxes, or model estimates of these parameters. In practice, the minimum requirement should be to have a record of the depth at which individual temperature measurements were made (denoted, e.g., SST_{20cm} or SST_{2m}) or of the depth used as a reference for adjustment (Merchant et al. 2012). Bias in SST due to any differences in measurement depth will be present in observations made using any measurement method, even if the sensor works perfectly.

Estimation of diurnal warming. Physical modeling of near-surface diurnal stratification and the ocean surface skin effect has proven useful in reconciling matched satellite and drifting buoy SSTs. If the model is used to adjust the satellite SST to the nominal depth of the drifting buoy (e.g., 20 cm), then the mean and scatter of the differences between the satellite and drifter measurements are reduced (Embury et al. 2012). The models can also be useful in adjusting observations to a reference local time of day, which may be important if the local time of observation changes systematically over time, for example, because of drifts in satellite orbits or changes to ship observing schedules. A range of physical models has been explored for near-surface modeling, based on turbulence closure (e.g., Janssen 2012; Kantha and Clayson 2004). Fast, parameterized semiphysical models (e.g., Takaya et al. 2010; Gentemann et al. 2009) and empirical models (e.g., Filipiak et al. 2012; Gentemann et al. 2003) are also available. In practice all of the models embed some parameter tuning. The physical models are generally driven by wind stress, total nonsolar and solar heat flux forcing, and sometimes also wave state. Fields from numerical weather reanalyses are usually used for the forcing, interpolated to the time steps required by the models (generally subhourly). Statistical models have also been developed (e.g., Morak Bozzo et al. 2016). While the utility of models in connection to satellite–buoy comparisons is established, there is still scope to use models to understand differences between SSTs at different depths and local times of day more generally, and to explore the relationship of measurements to alternative definitions of "SST."

Consideration of biases versus random errors. Another clarification is in order with regard to what we call biases, when discussing different measurement

methods. In the realm of all errors made in a certain measurement system under changing conditions, the separation of the total error into systematic and random components is, in practice, not as clean as one might expect.

Suppose we have obtained measurement T_o of the true value T with error u,

$$T_o = T + u. \tag{ES1}$$

If we could view u as a random value that has a nonzero mathematical expectation $\langle u \rangle = b$, then we would call b a bias in our measurement and write $u = b + \varepsilon$, where ε is another random variable, with expected value zero, so that

$$T_o = T + b + \varepsilon, \quad \langle \varepsilon \rangle = 0 \tag{ES2}$$

and $T_o - b$ is now an unbiased measurement of T. However, both b and ε depend on the observational methods and weather and thus on space and time. For complicated datasets, we usually realize that external conditions affect the measurement process, and that it is unreasonable to expect u to have the same distribution (and thus the same b) if conditions are different. Thus, splitting the entire dataset into subsamples S_C corresponding to different measurement conditions C (location, time of day, wind, humidity, air temperature, details of the measuring instrument, etc.), we could make the bias b a function of C,

$$b(C) = \langle u \text{ in } S_C \rangle. \tag{ES3}$$

Therefore, for conditions C, the value $T_o - b(C)$ gives an unbiased measurement of T.

For the sake of argument, consider the case when $b(C)$ takes opposite signs for different values of C, and its weighted average over different values of C is approximately zero. In this case $\langle T_o - T \rangle$ averaged over the entire dataset is approximately zero. Nevertheless, we should not be calling T_o an unbiased measurement of T because we know that there is a C-dependent bias $b(C)$ that we can estimate (or verify) by splitting the entire dataset into subsamples S_C according to C.

In other words, separating the error into systematic and truly random components generally depends on our understanding of the dependence of this error on some observed (or modeled) factors. Therefore, any mechanism that affects the error in the measurement process is a potential mechanism for modeling a systematic error component. Successfully modeling such processes, and their complicated spatial and temporal correlations, will reduce the error variance attributed to unexplained random effects. All factors known to influence errors of in situ SST measurements may be used in bias modeling.

The relationship of SST bias to the definition of SST. The definition of SST bias has evolved over time, alongside changes to the way that SST gridded analyses have been constructed and with improvements to modeling and quantification of uncertainty. For example, Bottomley et al. (1990) considered bias to be differences from the mix of observations during the climatological period (1951–80). The same definition (but relative to 1961–90) was used implicitly by Smith and Reynolds (2003), Rayner et al. (2003), Ishii et al. (2005), and Rayner et al. (2006). As sampling from drifting buoys increased, and differences between SST observations from ships and drifting buoys became apparent (Emery et al. 2001), it was recognized that SST observations in the climatological period also contained biases and required adjustment.

Hirahara et al. (2014) considered SST measurements from drifting buoys to be, on average, unbiased, and Kennedy et al. (2011b) found this to be the case. However, because analyses are referenced to a climatological period where drifters were not present, the offset between drifters and other observation types, and its uncertainty, need to be considered (Kennedy et al. 2011b; Huang et al. 2015). This approach leads to the estimated uncertainty in analyses of SST being smallest during the climatological period and increased uncertainty both before and after. This has the counterintuitive effect of producing relatively large uncertainties for the modern period, where observations are most accurate and sampling most complete.

The next generation of SST analyses will more explicitly consider differences in SST observations due to differential sampling of ambient conditions (e.g., with depth or time of day) and with biases due to different methods of observation. Considering drifting buoy SST observations to be an unbiased standard is expected to give a more realistic picture of uncertainty in SST analyses and to allow a cleaner separation of uncertainty into random and correlated components.

S3: THE INTERNATIONAL COMPREHENSIVE OCEAN–ATMOSPHERE DATA SET (ICOADS). ICOADS is now used as the source database for all major historical SST analyses. Current analyses are based on ICOADS, release 2.5 (Woodruff et al. 2011), but there is now release 3.0.0 (Freeman et al. 2017). So what information does ICOADS contain that is important for the analysis of bias in SST?

Quantifying bias in SST ideally requires extensive information on the methods, instruments, and protocols used along with more general information about the platform and the ambient conditions.

The availability of this information in ICOADS varies dramatically with the source of the information. Ship logbooks often contained much of this information, but deficiencies in early digitization methods and data management means that what remains in ICOADS can be just a small fraction of the information recorded on the ship. Particularly compromised are observations derived from atlas collections or punched cards. Ship identifiers are often missing, and observational metadata have degraded locations and some contain only a subset of the originally recorded environmental variables. In contrast, more recent observations, or older observations that were recently digitized, are more likely to contain platform identifiers, metadata, and the full range of observed variables. An exception is the masking of many ship call signs since 2007 (Woodruff et al. 2011) and now only partially reinstated (Freeman et al. 2017). Table ES1 attempts to summarize the information that may be available in ICOADS fields that can be used to constrain or inform SST bias adjustment.

First and foremost, it is important to identify reports that have been taken on the same ship or platform. This makes it easier to identify platform-specific biases that might not be related to the observation method. Examples of such biases might be a miscalibrated thermometer, errors in coding, or very poor observing practices. Identifying the platform is easy when the ICOADS identification (ID) field contains meaningful information; however, for many observations this is not the case.

Some progress has been made toward identifying groups of observations likely to have been taken on the same ship when ID information is missing or incomplete in ICOADS using probabilistic ship tracking (Carella et al. 2017). Implementation of tracking highlighted some problematic features of the way ICOADS has been assembled through the merging of a wide range of archives, often containing multiple reports derived from the same original observations. ICOADS

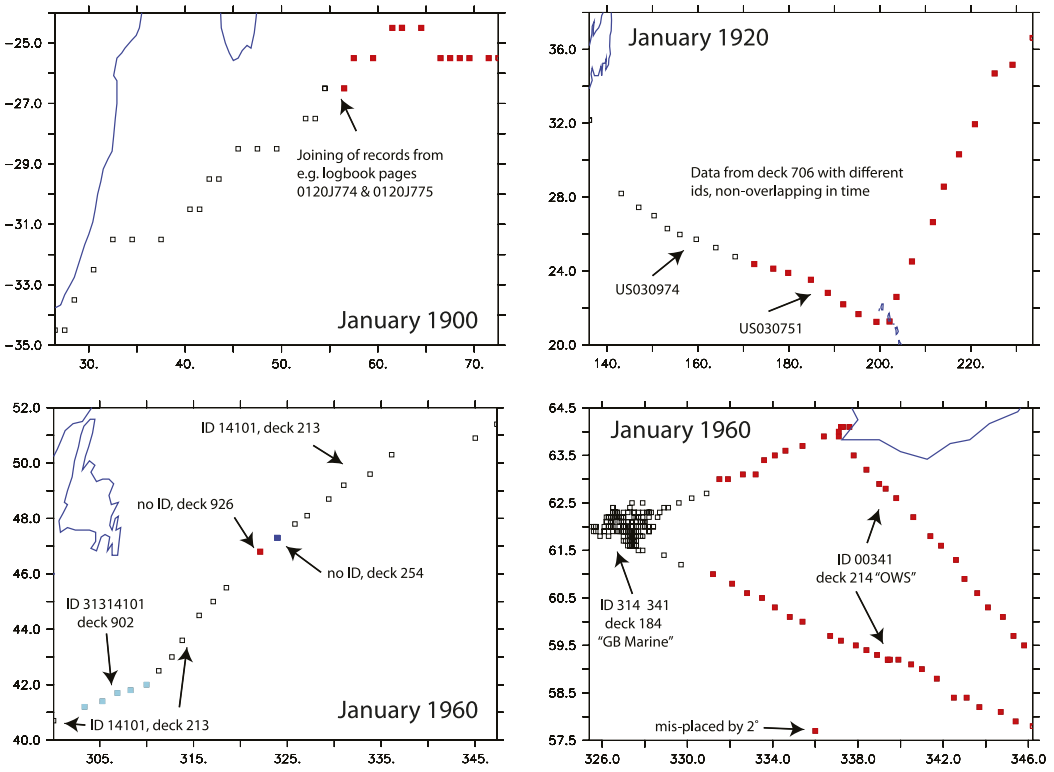

Fig. ES1. Examples of the ship tracking method of Carella et al. (2017) applied to ICOADS, release 2.5. (top left) In Jan 1900 a longer record is formed by associating observations with a consecutive ID, perhaps indicating a different logbook or logbook page from the same ship. (top right) In Jan 1920 a similar joining is effected, but with IDs that are related but more dissimilar. (bottom left) In Jan 1960 observations from a single ship are split between four different ICOADS decks, in two cases there are similarities in ID (cf. 14101, 31314101) but two further observations have no ID and each comes from a different deck. (bottom right) The final example (also from Jan 1960) shows that even data from OWSs are not immune to being split between different decks.

identifies potential duplicates with a complex fuzzy matching procedure and selects a best version, sometimes generating composite reports (Slutz et al. 1985). This means that reports from a single ship may appear in ICOADS with differing IDs, or none, with different parameters available and differences in metadata availability and content. Figure ES1 shows examples of ship tracking from ICOADS following Carella et al. (2017) illustrating the fragmentation of voyages that is often seen. The examples show that the development of rules to allow the joining of voyage fragments would improve the ability to associate groups of observations likely to have been taken from the same ship.

SST measurement method information is available for many reports in ICOADS through an indicator flag [SST indicator (SI); Table ES1]. Even where observational metadata are present, the information is likely to be incomplete, for example, typically ICOADS observational metadata might indicate use of a bucket, but no information on the type or size of bucket, or the sampling protocol. Missing or incomplete metadata can be estimated from contemporaneous literature, such as instructions to observers or other descriptions of measurements. Sometimes the original logbooks may give information that has not been transcribed to the digital record. For observations after 1955, information is available about ships and instruments in Publ. 47 (Kent et al. 2007). After about 1970, the metadata in Publ. 47 can sometimes be linked to individual observations using call signs [Kent et al. 2010; SI from metadata (SIM); Table ES1].

A systematic attempt to assign an observing method to every ICOADS ship-derived SST record, along with the uncertainty in that assignment, was made by Kennedy et al. (2011b). They assumed that observations with missing SI prior to 1941 were made using buckets. After that date reports with a call-sign ID but missing SI were linked to metadata in Publ. 47 where possible.

Table ES1. Metadata found in ICOADS, release 2.5, that can be used in the estimation of SST biases and how they bear on that problem.

Type of information	ICOADS field(s)	Information available	Application to SST bias adjustment
Observation source	DCK and SID	Deck (DCK) and source identifier (SID) codes give information on the origin of each report. The term *deck* originates from the "punched card decks" from which ICOADS was originally constructed.	DCK and SID provide links to additional information about the observation. Strong indicators of the completeness and quality of reports.
Type of observing platform	PT, OP	Broad categories of platform types (PT), including, ship, moored buoy, drifting buoy, fixed ocean platform, coastal installation, and oceanographic profiler. The level of detail varies by data source (DCK/SID) as does the mapping of different types of observation to PT codes. Observing platform (OP) provides additional information for observations from international logbook exchange.	Differentiation between observations from the major different types of platform is a first-order requirement for quantification of SST bias. ICOADS has some missing PT values.
Individual platform identifier	ID	The type of platform identifier (ID) information again varies with DCK/SID. Data derived from punched cards typically have no ID information, observations from the GTS may have a call sign, and observations digitized from ships logbooks may contain a full or truncated ship name.	Linking of observations to an individual platform important for uncertainty estimation. Sometimes the ID provides links to additional platform and observational metadata. Estimating biases for individual ships.
Country indicators	C1, C2, COR	Information on country that operates the ship or for more recent data the country that recruited the ship to its observing program. C1 (C2) is the primary (secondary) country information assigned by ICOADS. Country of registry (COR) is derived from Publ. 47.	There are strong country-specific preferences for different methods of observing SST, so in the absence of other information country can be used to assign measurement methods to observations. The reliability of observations is also known to vary with country.
Report metadata	OS, OPM	The observation source (OS) code gives information on whether the report was derived from a logbook, from the GTS, and the type of reports the ship makes. OPM is derived from Publ. 47 and gives information on the class of observing ship.	Indicator of report quality and availability of metadata.

The next step was to use the ICOADS country ID (C1; Table ES1) or deck (DCK) linking to the proportion of vessels listed from that country at that time in Publ. 47 to give a probabilistic measurement method assignment, for example, 70% likelihood of the measurement being made using a bucket, 30% likelihood of a hull sensor. This provides useful information because different countries operating observing fleets have preferences for different observation types. For example, the United States was an early adopter of engine room intake (ERI) technology, whereas U.K. ships predominantly used buckets. In this way an estimate of the basic measurement method can be obtained, albeit with substantial uncertainty, for much of ICOADS.

A different approach to measurement method metadata assignment was taken by Hirahara et al. (2014) for the Centennial Observation-Based Estimates of SST, version 2 (COBE-SST2) dataset. They estimated the bias for different measurement methods using observations that contained metadata. First, the relative proportions of insulated and uninsulated buckets were estimated by finding the ratio that minimized the difference between NMAT anomalies and anomalies based on adjusted SST measurements identified as being from buckets. Next, the relative proportions of ERI and bucket measurements were inferred for measurements with no metadata assuming the ERI bias is a constant offset.

The breadth of information available in ICOADS means it is possible to make an attempt to understand SST bias in the historical record. However, this comes at a considerable cost of complexity and consequently understanding, and interpreting the available information is a time-consuming task. This is despite the decades of work by ICOADS to present information in a consistent way, requiring detailed and traceable mapping of different information to common fields. Even with its level of complexity and a progressive revision and extension of data formats, some available information is not contained in the main ICOADS record. A strength of ICOADS is that any information not mapped into the main record is retained as a supplement to that record in a deck-specific format.

ICOADS is an invaluable asset, and decades ago it opened up climate observations for use without restriction in a way that has been seen as a model for others, such as the International Surface Temperature

Table ES1. Continued.

Type of information	ICOADS field(s)	Information available	Application to SST bias adjustment
Platform metadata	KOV, LOV	Kind of vessel (KOV) and length of vessel (LOV).	May be indicators of SST measurement depth for ERI and hull sensor measurements.
SST measurement method	SI, SIM	Flag that may indicate whether the report was made with a bucket, ERI, hull sensor, or electronic sensor. The SI is derived from the report or logbook. SIM from Publ. 47 metadata.	This is the first level of measurement method information giving basic information about the measurement method, but even this is often missing.
SST measurement depth	DOS	Depth of SST (DOS) measurement for ERI and hull sensors. Information available via linking to Publ. 47 metadata using ship call signs in the ID field. Planned changes to transmission and distribution data formats are likely to make this information more readily available in the future.	Depth of measurement should be an important parameter for understanding SST bias. However, expected variations with depth are often not detectable against a background of other errors.
Environmental information	AT, W, D, DPT, N, WH, WW, W1, W2	Observations of ambient conditions including air temperature (AT), wind speed (W) and direction (D), dewpoint (DPT), cloud cover (N) and sometimes more detailed cloud information, wave parameters [e.g., wave height (WH)], and weather codes (WW, W1, W2).	Knowing the environmental conditions is particularly important for bucket observations. Helps in understanding likely diurnal or depth variations in temperature. Some observations may have information on precipitation, ice conditions, or radiation that would affect both SST and SST bias.
Platform-specific environmental information	VS, DS, RWS, RWD	Ship speed (VS) and speed course (DS) in rather coarse categories. Relative wind speed (RWS) and relative wind direction (RWD) are available for a subset of ships.	Ship speed and direction can be used in platform tracking and quality control algorithms to verify spatiotemporal integrity. Relative wind speed and direction constrain maximum airflow around buckets

Initiative (ISTI) database (Rennie et al. 2014). However, the longevity of ICOADS means that modernization is now required so it can continue to best serve the international marine climate community. ISTI considers four levels of data that could be adapted for use with marine data.

- *Level 0* consists of images. For marine data this might contain digital images of logbooks, punched cards, or metadata.
- *Level 1* is data in the native format. For marine data this would be a wide range of formats representing data derived from many sources, for example, digitized from images, extracted from the Global Telecommunications System (GTS), from international exchange, extracted from national archives, or from delayed mode data centers. Each source will have its own format or formats, and contain observations and metadata in a variety of units or codings. All information needs to be decoded and the decoding documented. The input decks for ICOADS in their original formats (or as close to the original format as remain available) are level 1 data.
- *Level 2* is a merged dataset in common format, for example, ICOADS. ICOADS at this stage contains only records thought to be unique and also basic quality control (QC) flags and some observational metadata from external sources. Because the identification of multiple reports derived from the same original observation (duplicates) is imprecise, it would be valuable to have an intermediate product for specialist users that contained all available reports to allow for a diversity of approaches to duplicate elimination, as is desirable with all aspects of dataset development.
- *Level 3* enhances level 2 with value-added information, such as additional QC flags, statistically derived metadata, uncertainty estimates, and estimates of bias adjustments. The ICOADS Value-Added Database (IVAD; JCOMM 2015) has produced prototypes for this kind of information but is presently unfunded.

S4: BIAS ADJUSTMENT METHODS (I) FOLLAND AND PARKER (1995). *The Folland and Parker (1995) model and its implementation in the Hadley Centre SST Dataset, version 3 (HadSST3).* Folland and Parker (1995, hereafter FP95) developed two classes of models for the change in water temperature in the buckets used to measure SST. In one class, the bucket was uninsulated, being made of canvas; in the other class, the bucket was taken to be wooden, providing partial insulation. The models assume that the outside of the bucket remains wet and is therefore cooled by evaporation. Both types of models included equations for sensible and latent heat loss from vertical cylindrical surfaces, with circular bases and tops, along with gain of heat from solar radiation and gain or loss from longwave radiation exchange with the environment. The wooden bucket models also included equations for heat conduction through the wooden sides and base of the bucket, while assuming an open upper surface. The canvas bucket formulation produced results in good agreement with wind tunnel experiments on a Met Office canvas bucket by Ashford (1948). The FP95 bucket models require detailed information on the bucket construction and materials, the environmental conditions, and protocols (e.g., how long the measurement took to make, the extent to which the measurement was sheltered from the wind and sun). These details are not well known historically. Metadata describing the characteristics of the buckets and on how the measurements were taken were compiled from old instructions to observers, going back to Maury (1858). These revealed, for example, that buckets of diverse sizes were used, with diverse guidelines on operating procedures, such as the location on deck for reading the thermometer and the equilibration time needed for the reading to become steady after it was inserted. So, a family of canvas bucket models was developed, representing buckets of different sizes and measurement locations shaded from, or exposed, to direct sunlight. Table ES2 summarizes the parameters required to run the FP95 bucket models, and the values used in their ensemble.

The imprecisely known equilibration times were taken into account by assuming a fixed exposure time during hauling followed by a variable equilibration time within the documented range. FP95 showed that observations made prior to WWII had an excess annual-cycle variance compared to more recent measurements. This excess annual-cycle variance was attributed to seasonally varying biases in the bucket measurements, and a total exposure time (hauling plus equilibration) for canvas buckets was estimated from the model results to minimize this excess variance, ranging from 2.3 to 5 min. It was not possible to estimate the exposure time for wooden buckets (FP95). A fixed total exposure time of 4 min was therefore used for the wooden bucket models, following Maury (1858).

The environmental variables provided to the models included wind speed, relative humidity, air temperature, and solar radiation—both direct and diffuse. Many of these data were not available on an observation-by-observation basis, so climatological

averages were used, with the attendant increased uncertainty. Farmer et al. (1989) perturbed the environmental input parameters to a similar model and found uncertainties between 8% and 25% in resulting estimates of bucket temperature change, but that the procedure used to estimate exposure times showed lower sensitivity. Wind speed frequency distributions were used to enable nonlinearity in heat transfer to be incorporated via weighted contributions for each Beaufort force. However, this made little difference to the overall result. Wind speeds needed to be relative to the bucket, so the ship's speed and sheltering by the infrastructure were taken into account, the former with guidance from published information (Table ES2).

The published instructions to observers implied a gradual transition from mainly wooden buckets in the 1850s to entirely canvas by about 1920. The initial proportion of wooden buckets was estimated by maximizing the agreement between anomalies (relative to modern climatology) of SST and NMAT between the late 1850s and 1920 in regions where the latter had not been previously adjusted using SST. Rayner et al. (2006) refined this by using an upgraded dataset based on ICOADS, deck-height corrections to air temperature (Rayner et al. 2003), and Monte Carlo sampling of the FP95 bucket models to estimate uncertainties.

Challenges for the implementation of the FP95 model. FP95 showed that their canvas bucket model could replicate the wind-tunnel measurements of Ashford (1948) with reasonable accuracy. Thus, the main problem with the bucket models is not the physics, which is well-known, but the lack of both operational and environmental metadata. FP95 used climatological statistics for the environment, adjusted to compensate for the estimated influence of the ship on the wind and air temperature, entailing substantial uncertainty. FP95 estimated the corrections relative to the available SST observations for the period 1951–80. During this period the observations are made using a mix of methods (FP95; Matthews and Matthews 2013), but biases and other uncertainties are typically smaller than those before WWII (Kennedy et al. 2014). Improved estimates of SST are now available from drifting buoys and satellites, so adjustments can now be better referenced to the

Table ES2. Parameters considered by FP95 in their wooden and canvas bucket models. Information derived from FP95 (largely their Tables 1a and 1b), except exposure times, are from the Hadley Centre document "extra_aug93_run_specs.pdf."

Parameter	Range of values	
	Wooden	Canvas
Diameter	25.0 cm	8.0 and 16.3 cm
Height/depth of water	20.0 cm	12.0–14.0 cm
Thickness of bucket wall	1.0 and 1.6 cm	—
Time taken to haul bucket to deck	1 min	As wooden
Time taken to make measurement on deck after hauling	3 min	1.3–4.0 min
Equivalent mass of thermometer	35 g of water, inserted after 1 min	As wooden
Percentage of top or base assumed fully insulated	25% of top (base and sides are assumed partially insulated)	100% of base 0% of top
Leakage rate	None	0.5–1.0 cm min^{-1}
Solar radiation	Climatological all-hours 5° monthly average	As wooden
Shading of solar	50%–100%	As wooden
Ship speed	4 and 7 m s^{-1}	As wooden
Relative wind speed	Calculated from climatological Beaufort distribution of 5° monthly wind speed and assumed ship speed	As wooden
Sheltering of relative wind speed	Separately for ship speed and wind speed and for hauling and on-deck periods: sheltering ranges from 0% to 60%	As wooden, but ranging from 0% to 75%
Air–sea temperature difference	Climatological, 5° monthly	As wooden
Relative humidity	Climatological, 5° monthly	As wooden

modern observing system. It is possible to use information on environmental forcing from reanalyses to explore variability in the expected bias adjustments, but relying on reanalyses to produce the adjustments themselves is problematic.

The HadSST3 ensemble. The HadSST3 dataset (Kennedy et al. 2011a,b) is presented as an ensemble where each member has a different set of bias adjustments. The ensemble members were generated by randomly selecting plausible values for many of the

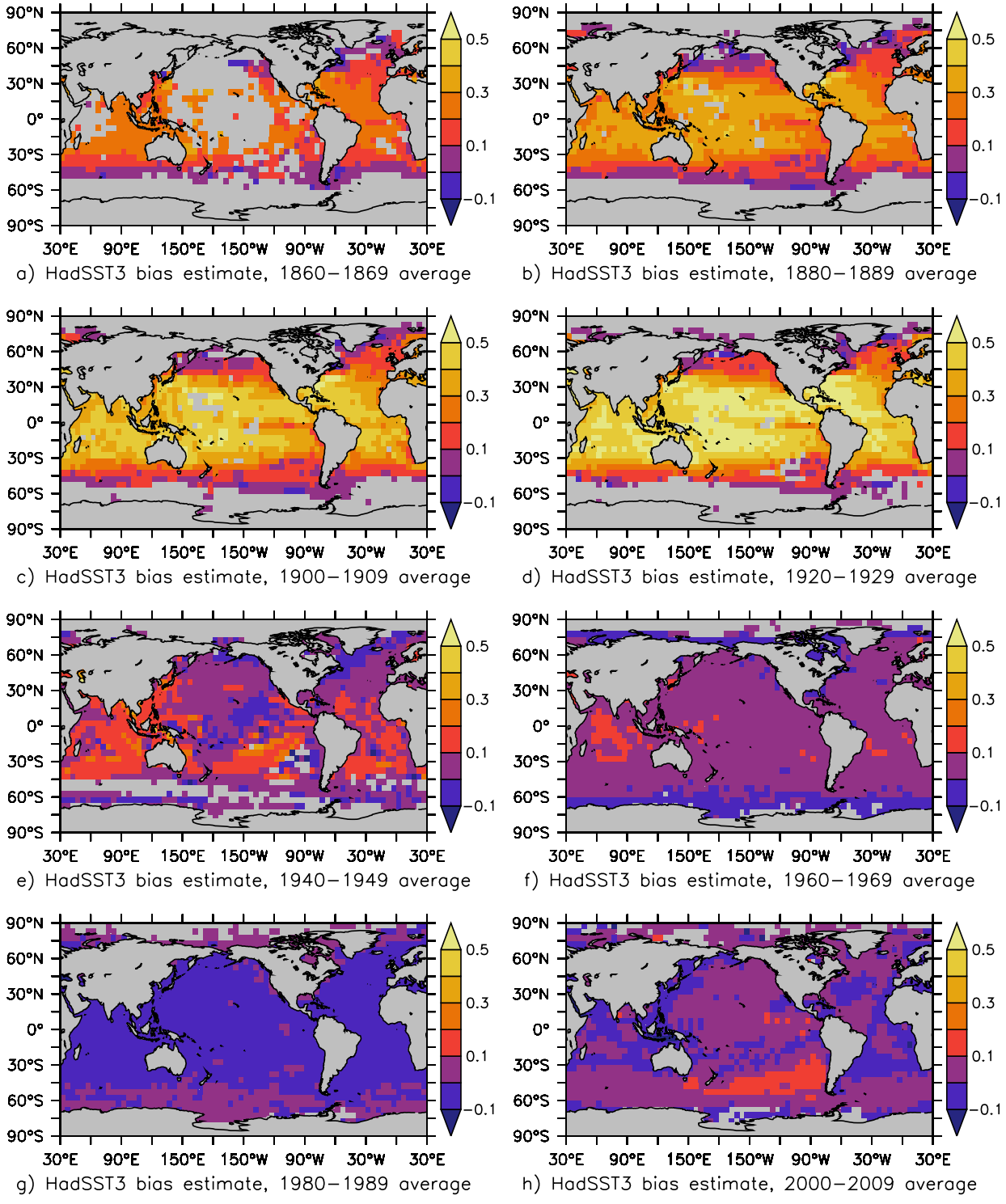

Fig. ES2. Median HadSST3 bias adjustment estimate, selected decadal averages (°C), masked for common coverage across HadSST3, COBE-SST2, and ERSSTv4.

parameters specified in the bias adjustment method. These parameters are described in Table ES3 and more detail is given in Kennedy et al. (2011b) and Rayner et al. (2006). Other choices—the effect of applying a ship–buoy adjustment to ships or buoys and the effect of using only ships—were also tested and shown to make little difference to the results. The adjustments applied to buoys, ERI, and bucket measurements brought these three separate elements of the observing system into better agreement and reduced the variability in global NMAT–SST differences. There is no preferred ensemble member, each is considered to be interchangeable. The combined measurement and sampling uncertainty is estimated separately. The HadSST3 ensemble covers the period 1850 to the present (October 2016 at time of writing).

Figure ES2 shows decadal average maps of bias adjustment estimates from HadSST3. Prior to WWII (Figs. ES2a–d) the bias adjustments are derived from the FP95 model and gradually increase over time (Fig. ES3a). During and post-WWII (Figs. ES2e–h), the adjustments are smaller and show an imprint of the increasing number of ERI measurements and the transition to a buoy-dominated SST observing system. Figure ES3 shows the ensemble of 100 global-average estimates of the bias adjustment for HadSST3. The global-mean bias adjustment increases from the start of the record to just before WWII, a result of the modeled transition from partly insulated wooden to uninsulated canvas buckets and an increase in ship speed (FP95; Kennedy et al. 2011b). All ensemble members show an increase over this period. The rapid change during WWII is very clear, and all ensemble members show a return toward pre–World War II values after the war. The annual cycle of the bias estimates is smaller post-WWII (cf. the black monthly lines with the 12-month-filtered red lines) and the 12-month-filtered ensemble spread is larger than in the prewar period.

S5: BIAS ADJUSTMENT METHODS (2) HIRAHARA ET AL. (2014). In addition to statistical bias adjustments, it is possible to use a combination of statistical and physical information to estimate SST bias. For example, Hirahara et al. (2014) combine the

Table ES3. summary of choices and perturbations for HadSST3.

Parameter	Range
NMAT dataset used to estimate fractions of canvas and wooden buckets in the nineteenth century	Two different NMAT datasets were used.
Bucket correction fields	These were generated using the method described in Rayner et al. (2006).
Fraction of fast ships (7 m s^{-1}; cf. slow ships, 4 m s^{-1})	Perturbed within the ranges specified in Rayner et al. (2006).
Bucket biases	For each of the 12 calendar months a Gaussian perturbation with standard deviation equal to 14% of the best estimate was applied to the canvas bucket fields. An independent Gaussian perturbation of 20% was applied to the wooden bucket fields.
NMAT–SST difference	Random samples were drawn from the NMAT and SST tropical averages used to estimate the fractions of wooden and canvas buckets in the nineteenth century.
ERI bias	The ERI bias was assumed to have a mean of 0.2 K and a range drawn at random from a uniform distribution between 0 and 0.2 K. Temporal variation was assumed to be strongly autocorrelated (Kennedy et al. 2011a).
ERI bias in the North Atlantic	Samples of ERI bias were drawn at random from the distributions and time periods given in Kent and Kaplan (2006).
Unknown measurements	Measurements for which no measurement method could be found were assigned to be either bucket or ERI measurements. A time-varying fraction of unknown observations were assigned as bucket observations, with strong autocorrelation. The remaining fraction was set to ERI. The same value is used at all places.
ERI recorded as bucket	30% ± 10% of observations identified as bucket observations were reassigned as ERI. One value per realization was applied at all times and places after 1940.
Canvas-to-rubber bucket transition	Linear switchover. Start point (all canvas) chosen randomly between 1954 and 1957. End point (all rubber) chosen randomly between 1970 and 1980.
Ship–buoy difference	Values were randomly generated for different regions using estimated mean and standard errors of collocated ship–buoy differences.

FP95 bias estimate with an estimated ERI bias using a statistical method that ensures consistency between different elements of the global fleet. There are two principal parameters that had to be determined statistically: the fraction of observations with an unknown measurement method that were ERI measurements and the fraction of bucket measurements that were made using insulated buckets. In cases where the type of SST measurement is clearly identified by metadata, a bias is assigned as appropriate for an uninsulated bucket, an insulated bucket, or an ERI measurement. Values of the two parameters were estimated such that there was consistency 1) between the different kinds of identified observations, 2) between the identified observations and the observations whose measurement methods were unknown, and 3) (before 1970) between SST and NMAT. Certain fixed points were also specified based on external information. Since metadata are poor for much of the historical record, the statistical estimate has a strong influence on the bias. Uncertainty in the biases was determined statistically from the fits. The resulting fields are shown in Fig. ES4 and in the global time series in Fig. 1 in the main paper. Prior to WWII the COBE-SST2 bias estimates have similar spatial patterns to the HadSST3 fields (cf. Figs. ES2, ES4), which is expected because both are based on the FP95 method.

S6: BIAS ADJUSTMENT METHODS (3) SMITH AND REYNOLDS (2002), HUANG ET AL. (2015), AND LIU ET AL. (2015). *The Smith and Reynolds (2002) model and its implementation in ERSST.* Smith and Reynolds (2002, hereafter SR02) developed a statistical bias estimate based on large-scale differences between SST and NMAT measured from ships. First maps of SST–NMAT smoothed differences are computed for a period when sampling is relatively dense (they used 1968–97). Earlier monthly SST–NMAT values are fit to these maps to compute the estimated bias for each month. The original SR02 estimate of the historical SST bias was obtained by fitting a straight line through the SST–NMAT filtered time series from the beginning of the record up to 1941, at which point the SST bias was assumed to be zero.

Since 2002 several updates were made to the statistical bias estimate. The first update was to gradually reduce the monthly bias to zero over 1938–41. That was done after it was discovered that new data in ICOADS, release 2.0 (Woodruff et al. 1998), substantially altered the bias composition, making the shift from bucket to ship intake temperatures more gradual over those years (C. K. Folland 2003, personal communication; Smith and Reynolds 2004).

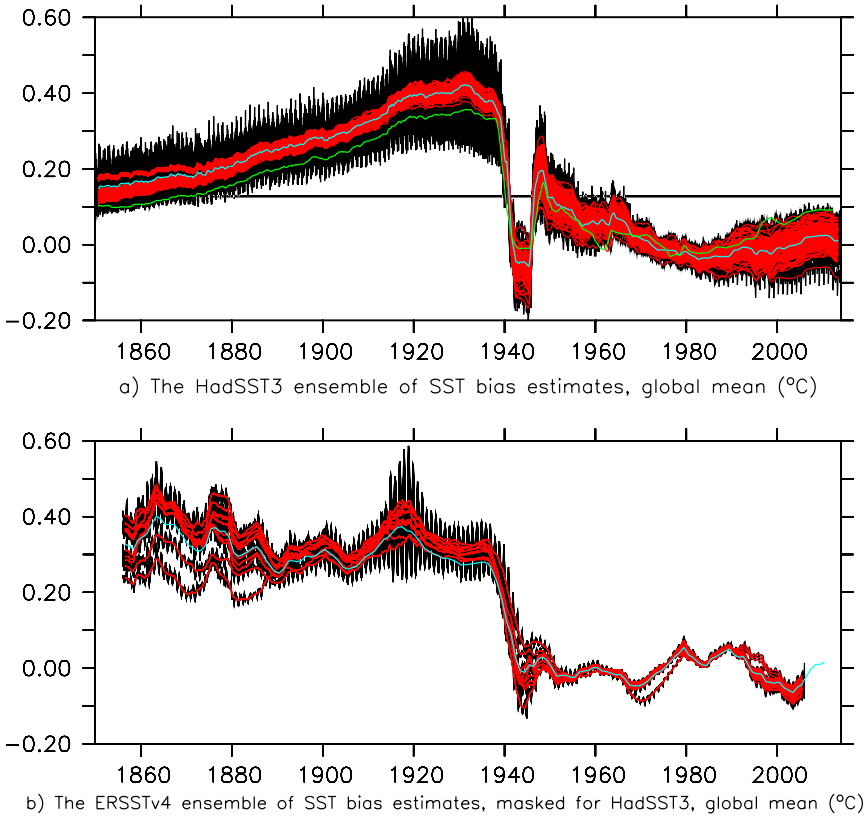

Fig. ES3. Global-mean estimates of SST bias adjustments from HadSST3 and ERSSTv4 ensembles. (a) A 100-member HadSST3 ensemble: monthly mean (black), 12-month running mean (red), median of ensemble (light blue), and 12-month running mean (green). (b) As in (a), but for three realizations of the 12-member ERSSTv4 ensemble, interpolated to the HadSST3 grid and masked for presence of HadSST3. The three realizations cover the estimated uncertainty in the ship–buoy mean difference (0.08°, 0.12°, and 0.16°C). Note that the three different estimates of the ship–buoy differences affect only the period from the 1970s onward.

Huang et al. (2015) reevaluated and updated the SR02 statistical method. They used climate model output to confirm that differences between SST and MAT are relatively stable on large scales. They also evaluated the influence of different time filtering on the bias estimate and chose a multiyear filter that gives more temporal variation than the linear filter used by SR02. In addition, while the original bias adjustment ended in 1942, the new version continues into the modern period, although the modern

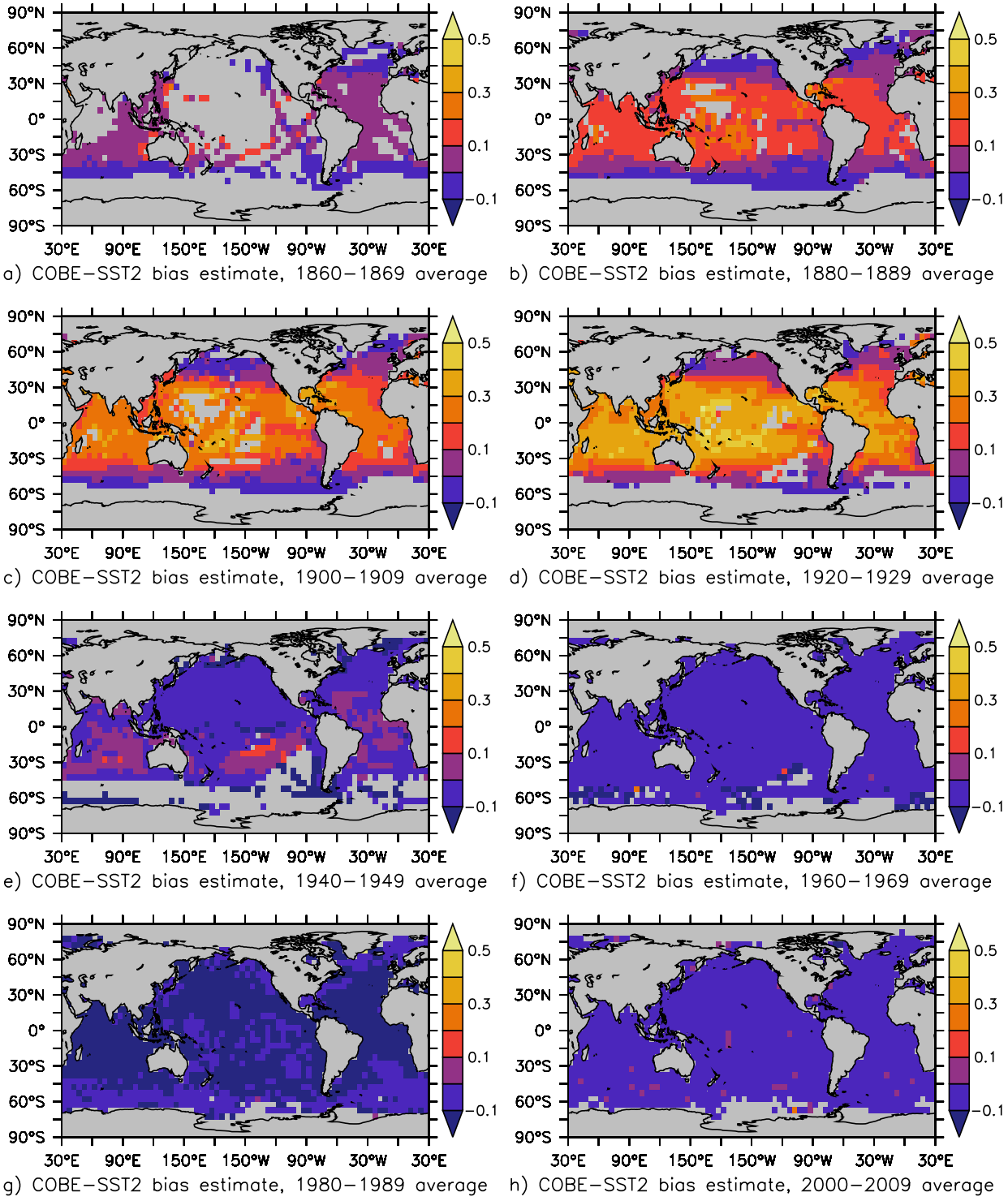

Fig. ES4. Median COBE-SST2 bias adjustment estimate, with selected decadal averages (°C), masked for common coverage.

bias estimates are much smaller than the pre-1942 estimates. An additional change in the Huang et al. (2015) bias estimate is an explicit accounting for the ship–buoy bias, due to warming of ship ERI temperatures (Thompson et al. 2008). To minimize that bias, they adjust the modern buoy temperatures to more closely match the historical ship temperatures. In Huang et al. (2015), ERI temperatures are considered to be the standard because of their relatively long record and prevalence during the climatology period. However, it should be noted that the mean ERI bias changes over time (Kent and Kaplan 2006).

Challenges for the implementation of the SR02 model. The implementation of the FP95 physical model requires the specification of which ships use buckets, what kind of bucket, and the mean wind speed across the deck. The SR02 statistical method has the advantage of not needing that information. However, the statistical method has a number of problems that cause uncertainty in its estimate. Although NMAT variations are representative of SST variations at the largest scales, the relationship can be weaker for some local regions. The computed spatial patterns of SST–NMAT are critical for the estimate, and assuming that the patterns are well known and invariant over time they also introduce uncertainty. In addition, although historical NMAT sampling methods are more stable than SST methods, there have been some changes over time that can influence bias estimates (Rayner et al. 2003; Kent et al. 2013).

Another problem with statistical bias estimates is that it is difficult to separate bucket biases from others, such as ship ERI biases. The spatial patterns used to compute statistical bias may filter out non-bucket bias if data used to compute the patterns are dominated by bucket SSTs. However, the estimates computed may reflect other SST biases to some extent.

Possible improvements to statistical methods could be obtained by using spatial patterns that more cleanly reflect separate physical processes. For example, bucket bias patterns could be computed using only SSTs identified as bucket observations. Bias time series could then be computed separately for each process using the available metadata or statistical estimates of observation type. Challenges for developing such improvements include developing measurement-type estimates for all data and developing spatial patterns for each SST sampling type. In particular, it is not clear that there are enough data to estimate spatial patterns for ERI biases or that the patterns would necessarily be stable over time.

The ERSST v4 ensemble. In the Extended Reconstructed SST, version 4 (ERSST v4; Huang et al. 2015), and earlier versions (Smith et al. 2008), the estimated ship SST bias depends on, and may be sensitive to, the following four internal parameters: 1) the acceptable range for SST–NMAT differences to be included in the analysis, 2) the pattern of the SST–NMAT climatology, and 3) the NMAT dataset used. The range [minimum, maximum] of SST–NMAT acts as a quality-control criterion: if the difference of SST–NMAT is outside the specified range, then the SST–NMAT pair will be rejected from the bias assessment. Six options are selected to include the potential impact from the SST–NMAT range (Table ES4). The SST–NMAT patterns may be constructed for different regions: i) global (as used in SR02), ii) three latitudinal belts, iii) three ocean basins, and iv) the 45° longitude × 25° latitude–running region. Two versions of NMAT datasets [HadNMAT2 and Met Office Historical Marine

Table ES4. The ensemble of 3 × 12 experiments used to assess the uncertainty of ship SST bias adjustment. Each experiment is made by perturbing one parameter option (third column) while other parameters use the operational ERSSTv4 options (second column). The ensemble of 12 different approaches to ship bias adjustment combines six different estimates of the SST–NMAT range with four different SST–NMAT climatological patterns and two different versions of the NMAT dataset. This 12-member ensemble is combined with three different estimates of the ship–buoy SST difference to give 36 ensemble members in total (Fig. ES3).

Parameter	Options in operational ERSSTv4	Options in bias uncertainty ensemble
SST–NMAT range (°C)	[–2.0, 4.5]	[–2.0, 4.5]; [–2.0, 6.5]; [–2.0, 8.5]; [–2.0, 10.5]; [–1.0, 4.5]; [–3.0, 4.5]
SST–NMAT climatological pattern	Global	Global; three latitudinal belts: 90°–30°S, 30°S–30°N, and 30°–60°N; three ocean basins: Indian, Pacific, and Atlantic Oceans (including Arctic Ocean); 45° longitude × 25° latitude–running region
NMAT	HadNMAT2	HadNMAT2;Met Office MOHMAT43N
Ship–buoy difference (°C)	0.12	0.08, 0.12, 0.16

Air Temperature (MOHMAT43N; Kent et al. 2013; Parker et al. 1995)] are used. The ERSSTv4 ensemble is completed by varying the fourth internal parameter. 4) This internal parameter is the mean difference between ship observations (after bias adjustment) and buoys. In contrast to the larger ensemble used by Huang et al. (2016), this ensemble includes only perturbations to parameters that affect the biases, and the ensemble members span what is thought to be the plausible range rather than being randomly

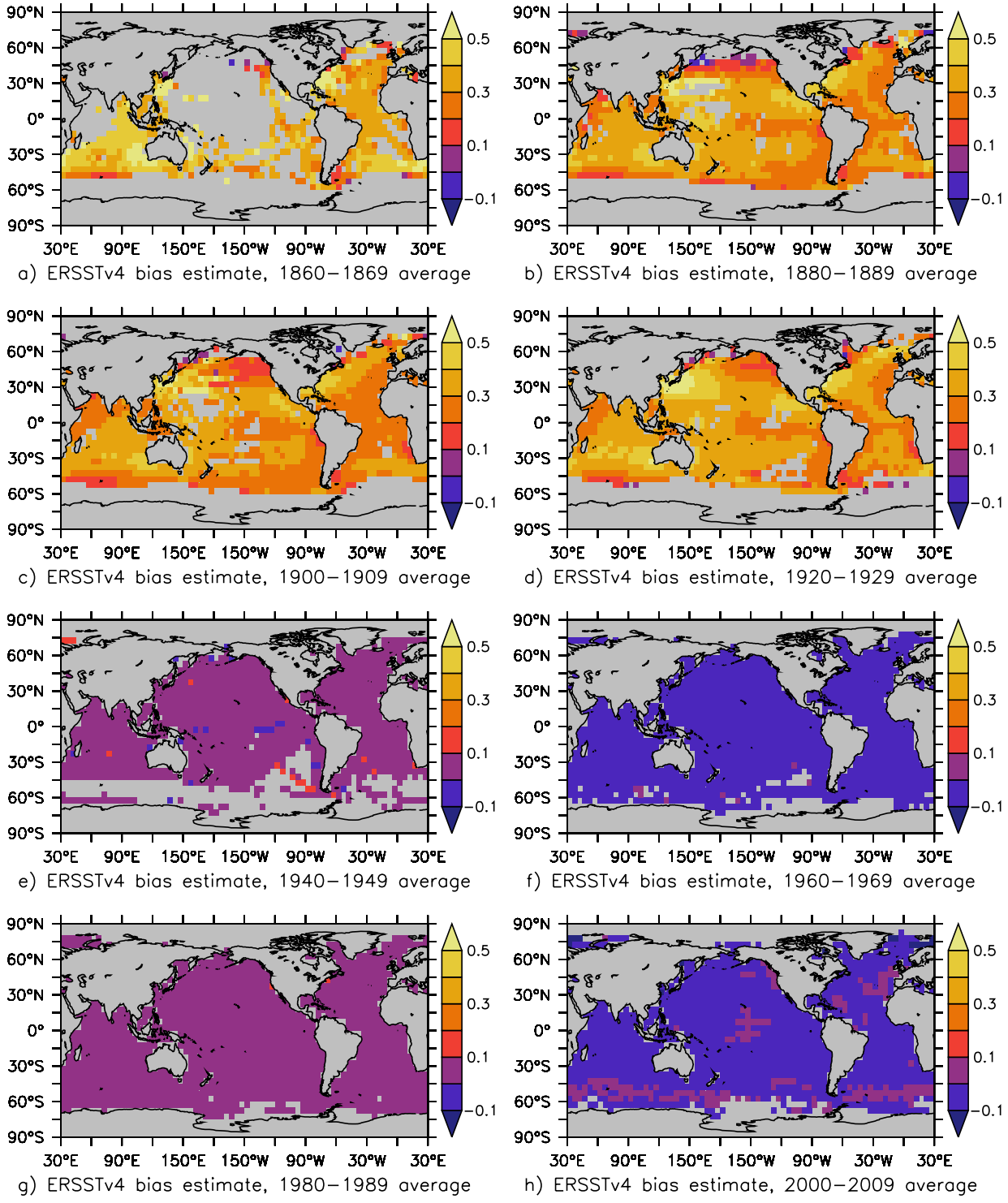

Fig. ES5. As in Fig. ES2, but for ERSSTv4, interpolated to the HadSST3 5° area grid and masked for common coverage.

drawn. Time filtering is set to annual (Huang et al. 2015), which gives the widest range of variability. The range rather than the standard deviation is used as the measure of uncertainty, as the ensemble members are not normally distributed (Fig. ES3), and the ensemble size is small. Prior to 1886 the bias estimates deviate from those used by Huang et al. (2015, 2016), who used the bias estimates for 1886 prior to that date. In this ensemble the sparse and uncertain NMAT data prior to 1886 are used directly to show the impact of its uncertainty on the bias adjustments. The ensemble covers the period 1856–2005 and the standard bias estimate covers 1856–2009.

Figure ES5 shows decadal-average maps of bias adjustment estimates from ERSSTv4, masked for availability of HadSST3 to aid comparison with Fig. ES2. Figure ES3b shows the ensemble of 36 global-average estimates of the ERSSTv4 bias adjustment. The ERSSTv4 bias estimates show no clear trend over the pre-WWII period and the ensemble spread is relatively large. The reduction in SST bias estimate over the WWII period is clear, and the bias estimates are fairly small, with small ensemble spread after WWII.

S7: COMPARISON OF BIAS ADJUSTMENT METHODS FROM HADSST3 AND ERSSTV4. Both the HadSST3 and ERSSTv4 datasets present uncertainty in the bias adjustments applied as an ensemble, and in this section those ensembles are compared. COBE-SST2 does not present bias estimates as an ensemble and is therefore not considered here. HadSST3 is available on a monthly 5° grid with missing values in unsampled grid boxes, whereas ERSSTv4 is presented on a monthly 2° grid with gap filling. To allow for a fair comparison, the ERSSTv4 values were interpolated onto the HadSST3 5° grid and masked for the presence of HadSST3. It obviously would be preferable to compare datasets that were more similar—for example, using only gap-filled datasets—but that is not presently possible. Figure ES3 shows the global-mean bias estimates for the HadSST3 and ERSSTv4 ensembles. Both estimates indicate that the reported SST is too cold in the global mean prior to WWII. Between about 1890 and WWII they agree that the global-mean estimate is in the range 0.3°–0.4°C but disagree earlier than this period. This is probably because HadSST3 has the documented transition from partially insulated wooden to uninsulated canvas buckets hardwired into the adjustment procedure, whereas ERSSTv4 does not. The difference in the global-mean adjustments is outside the sum of the range of the ensembles prior to about 1880 (Fig. 4 of the main text). The sudden decrease in SST bias adjustment in WWII is shown in both datasets. The behavior of the post-WWII bias adjustment estimates is different: HadSST3 indicates that uninsulated buckets became common again, but the ERSSTv4 bias adjustment estimates remain closer to zero (Fig. ES3). However, the difference remains just inside the joint ensemble range (see the main text). The difference in bias adjustments exceeds the range for a sustained period from the late 1970s to the early 1990s. This spans the start of the availability of in situ observations of SST from buoys and shows that more research is required to fully understand the transition from a ship-only to a buoy-dominated observing system. The global-average joint ensemble spread is at a minimum during this period and the ERSSTv4 ensemble range is smaller than the 2σ uncertainty from HadNMAT2 (not shown), suggesting that ERSSTv4 bias uncertainty is underestimated in this period.

Figure ES6 compares zonal-average bias adjustment estimates from HadSST3 (Fig. ES6a) and ERSSTv4 (Fig. ES6c) and their uncertainties (Figs. ES6b,d). HadSST3 shows a stronger latitudinal variation throughout the entire period than ERSSTv4. HadSST3 estimates of bias uncertainty in the high latitudes are small, particularly pre-WWII, and likely to be underestimated. HadSST3 bias adjustment estimates show hemispheric differences post-WWII, reflecting differences in the spatial distributions of SST measurements from ERI and buckets. Such large-scale spatial variations are not seen in the bias adjustment estimates from ERSSTv4.

S8: VALIDATION. *Improved independent datasets for validation.* As described in the main paper, there are very few independent data sources for the validation of the bias adjustments. The identification of additional data sources that can be used to assess SST bias adjustments is therefore extremely important. These data sources can have a variety of characteristics that are useful.

ACCURATE LONG-TERM OR LARGE-SCALE MEASUREMENTS OF SST. Currently, there is limited availability of long time series of known quality at fixed locations. Some moored buoys in the tropical Pacific have records from the early 1990s, with reasonable metadata, that can be used to evaluate long-term stability of a large and more heterogeneous network (Merchant et al. 2012). Other long series need to be identified and access to the data secured. Coastal stations are another possible source of validation data (e.g., Hanawa et al. 2000). The suitability of individual records would need to be given careful consideration because coastal variability can be large and secular changes relative

to open-ocean temperatures cannot be ruled out a priori.

Near-surface Argo (Argo 2000) and CTD temperature measurements are the closest thing to widespread reference that we have, but their coverage was limited until around 2005, when the Argo program reached its target deployment coverage. Argo floats, which can be configured to make detailed, repeated measurements near the surface, also could be of use in understanding near-surface temperature structure and its variability, which would help to understand differences in SST at different depths.

Most satellite-derived records of SST are obtained using empirical retrieval equations tuned to drifting buoys. While this is easy to do (the calibration of the satellites and the physics of the atmospheric radiation are handled implicitly), the lack of independence limits the use of such satellite SSTs for critical assessment of in situ SST analyses and their biases. Merchant et al. (2012) attempted physics-based retrieval of SST from infrared sensors on satellites, developing, from physical modeling of atmospheric radiative transfer (Merchant and Embury 2014), retrieval algorithms that are independent of any tuning to the in situ SST observing system. Results based on dual-view, well-calibrated sensors [along-track scanning radiometers (ATSRs)] have been found to have useful levels of accuracy and temporal stability for interrogating biases in the in situ record (Kennedy et al. 2012). To extend such approaches to cover the 1980s will involve deriving comparable physics-based SSTs from the Advanced Very High Resolution Radiometer (AVHRR), whose calibration systems were not designed with climate requirements in mind. This proves to be challenging and is the subject of ongoing efforts (Merchant et al. 2014), including development of new methods in collaboration with the metrology community (see www.fiduceo.eu).

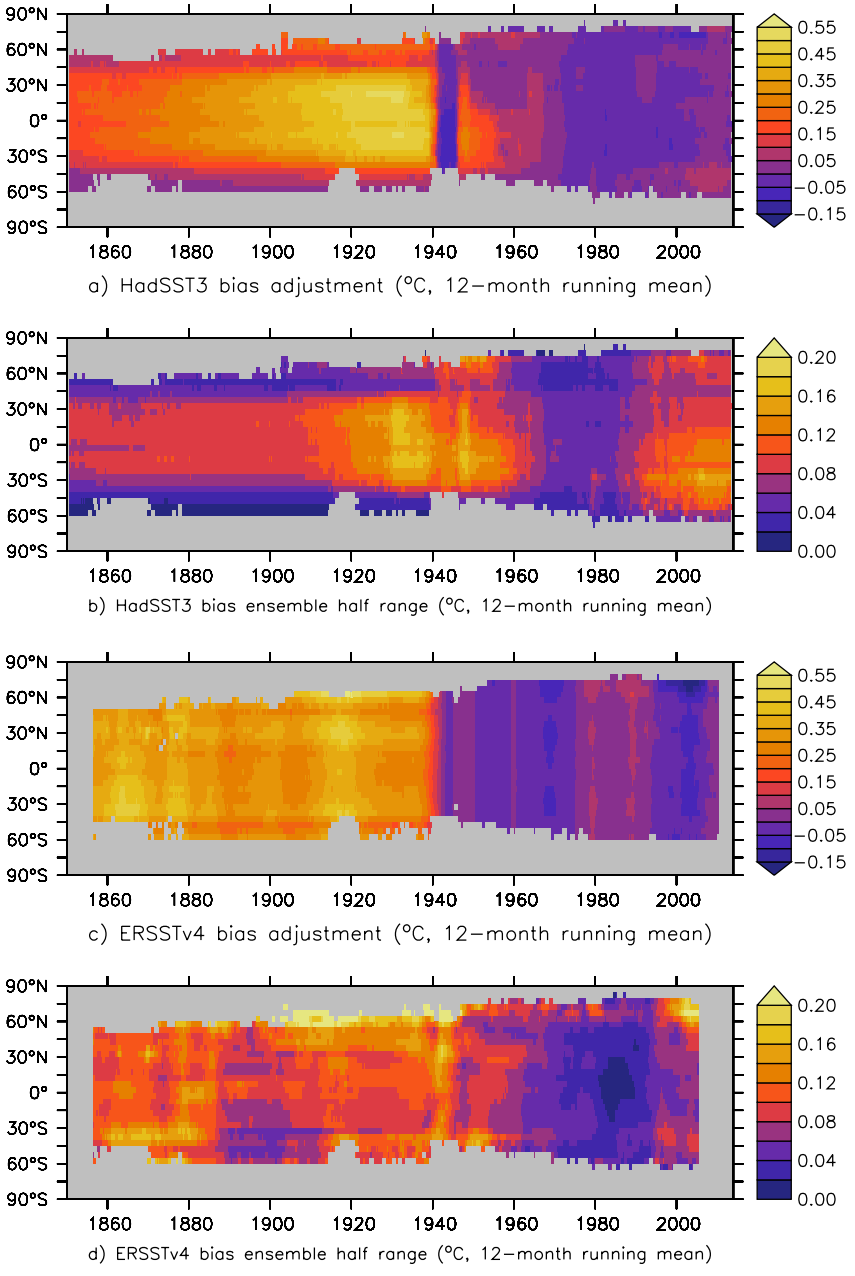

Fig. ES6. Bias adjustment estimates from HadSST3 and ERSSTv4 and their ensemble ranges. (a) HadSST3 median SST bias adjustment estimate, zonal mean, 12-month running-mean filter. (b) Half the full ensemble range for HadSST3 bias adjustment estimates, zonal mean, 12-month running-mean filter. (c) As in (a), but for ERSSTv4 masked on the HadSST3 grid. (d) As in (b), but for ERSSTv4 ensemble masked on the HadSST3 grid.

Global or large-scale estimates of related parameters. Marine air temperature is an obvious point of comparison (Kent et al. 2013; Huang et al. 2016) for assessing SST biases. It is thought that biases in NMAT are relatively tractable for much of the historical record. There are, however, periods such as WWII when biases in NMAT are large and difficult to assess. These are also periods when little is known about SST biases. Daytime MAT (DMAT) is affected by solar heating of the ship, which is a difficult, but not impossible, problem to solve (Berry et al. 2004).

There is a complication: biases in SST are thought to depend at least in part on MAT, which means that detailed physically based bias adjustments to SST measurements made using buckets cannot be independent of MAT. Careful partitioning of the areas used to define SST and NMAT relationships have been used in the past (FP95), but observation-by-observation adjustments will make that much harder. A more holistic approach would consider SST and MAT and their biases together. Development of both NMAT and DMAT should continue alongside SST. MAT is of interest in its own right.

Stations on land at coastal or island locations often have long records of air temperature. Measurements from these stations, although they are also affected by station moves and instrumentation changes, are unlikely to have experienced such a change that coincides with a change in measurement methods aboard ships. Care is needed to ensure that the comparison is appropriate, taking into account the very different properties of the two measurands as well as known and suspected inhomogeneities. Identifying these stations and obtaining data and metadata are a first step toward this.

Comparisons with data known to have some uncertainty that can reasonably be expected to be independent of, or smaller than, the biases in SST. The ocean weather ship (OWS) network started during WWII and declined during the 1970s and 1980s with the last station (OWS Mike at 66°N, 2°E) being maintained until December 2009. OWSs were manned by trained observers and reported atmospheric and oceanographic observations, including ocean and atmospheric profiles and observations near the surface. Observations were made using standard protocols and it is likely that information on the instruments used is still available. Despite their potential as a high-quality time series and as comparison or validation data, OWS data have never been fully collated as a climate resource.

Ship-based oceanographic measurements are another possibility for validation. They are likely to be independent from observations from the general run of ship observations and generally of higher quality. Some means of measuring temperature profiles [e.g., expendable bathythermographs (XBTs) and mechanical bathythermographs (MBTs)] are known to have biases of their own. XBTs are known to have a depth-dependent bias, which would be negligible near the surface, but there is also evidence of bias affecting temperatures at all depths. When comparing these with surface measurements, we benefit from ongoing efforts to improve our understanding of those biases and their uncertainties (Cheng et al. 2016). Such data, along with estimated biases and uncertainties, are already available in datasets such as the Hadley Center Integrated Ocean Database (HadIOD; Atkinson et al. 2014).

Other candidate data sources for SST bias adjustment validation include drifting buoys and the extratropical buoy networks operated by meteorological agencies in support of weather forecasting. The European Space Agency Climate Change Initiative (ESA CCI) for SST is producing AVHRR-based datasets (Merchant et al. 2014) independently of the in situ observing system; these are expected to have somewhat lower quality than the high-stability ATSR-based SSTs (Merchant et al. 2012) but, crucially, will extend both prior to and after the ATSR period (which spanned 1991–2012).

Future requirements for long-term stable SST measurements. To avoid these difficulties continuing into the future, we need long-term stable measurements. In recent years, the Argo network has started to provide this. Together with the tropical moored arrays (McPhaden et al. 1998; Bourlès et al. 2008; McPhaden et al. 2009), the drifting buoy network, and high-quality satellite retrievals such as those provided by the ATSR Reprocessing for Climate project (Merchant et al. 2012), the period since the turn of the century has seen a brief golden age of SST measurement.

However, the recent reduction in maintenance of the Tropical Atmosphere Ocean (TAO) tropical moored buoy array (Legler and Hill 2014) together with a large drop in the number of drifting buoys has left large areas of the global ocean, including parts of the tropical Pacific, without regular, reliable SST measurements. Reduced coverage of reliable observations has led to difficulties in estimating SST reliably in these regions (Huang et al. 2013). This highlights the difficulties and importance of maintaining a global integrated network of high-quality measurements.

As for the space observing system, it was intended that Sea and Land Surface Temperature Radiometers (SLSTRs; Coppo et al. 2010) would overlap in time

with and then supersede ATSRs, providing continuity of well-calibrated, dual-view radiometer SSTs between 1991 and at least circa 2030. The *Sentinel-3A* platform carries the first SLSTR and was launched in February 2016, so the gap in operational dual-view data exceeds 4 years. It is scientifically desirable (although challenging) to attempt to bridge this gap with SST-capable single-view sensors integrated into a stable SST record. This would entail traceably reconciling the calibrations of the last ATSR and first SLSTR through these intermediaries while maintaining independence from the in situ network.

SST, along with pressure, is presently much better observed in situ than other marine essential climate variables (ECVs; GCOS 2010) such as MAT (Berry and Kent 2017), near-surface humidity, wind speed and direction, and cloud cover, which are all needed for physically based bias estimates for SST. Satellite observing systems can estimate SST, winds, cloud, and surface radiation ECVs to useful accuracy (GCOS 2011) but not MAT or humidity. One of the challenges of understanding historical SST biases is to disentangle the actual relationship between environmental parameters and SST from the bias relationship with those same environmental parameters. High-quality modern observations can allow us to understand the real relationships (e.g., Morak-Bozzo et al. 2016). These relationships become more difficult to understand when the environmental information (MAT, humidity, winds, clouds) is more sparsely observed than SST itself, which is the current situation. This did not use to be the case. The other ECVs were observed on ships in the WMO VOS scheme, but VOS numbers have reduced dramatically in recent years compared to the 1970s and 1980s (Berry and Kent 2017).

Data management and availability of validation data sources. To validate bias estimates using the abovementioned sources of data, it goes without saying that we need access to these data. This could involve tracking down datasets from individual studies, obtaining the data from the authors, establishing terms and conditions on use, processing the data into a usable format either by reformatting digitally or digitizing the data, systematically recording useful metadata, and permanently archiving the resulting data. To do this for diverse sources such as harbor series, ocean weather ships, lighthouses, moored buoy collections, drifters, research vessels, and research moorings will be a challenge but would provide a valuable resource likely to have applications far beyond SST bias validation.

Even where extensive digital archives exist, they are not always definitive, complete, or curated for use in climate studies. Metadata for buoys are not always easily accessible and, in some cases, not accessible at all. Quality monitoring of drifting buoys occurs operationally, so that the data can be ingested into real-time SST analyses but that information is dispersed across a number of centers. Land data holdings are very fragmented, different variables are dispersed among different data centers, and different versions of data may be held regionally (Thorne et al. 2017).

Validation using derived SST fields as model forcing. Folland (2005) used an atmosphere-only climate model that was run with prescribed SSTs both with and without bias adjustments applied to data prior to 1941. Land air temperatures in those runs conducted with adjusted SSTs more closely matched observed land air temperatures than did the runs with unadjusted SST. This suggests that the assessment of bias adjustments can be informed by appropriately careful use of model runs. Atmospheric and oceanic reanalyses also use SST datasets, so some information might already exist about the effects of using different SST datasets in these contexts. Exploratory work is needed to assess the feasibility of detecting small changes in SST bias in this way.

With climate models, weather models, and reanalyses, it ought to be possible to simulate "bucket" measurements directly by including the physical FP95-style bucket model in the simulation either at run time or offline. A model would simulate many of the necessary environmental parameters and allow bucket measurements to be assimilated or simulated directly.

Validation using measures of internal consistency. Models of biases—physical models of biases most obviously—make predictions of observable phenomena. For example, FP95 predicted that buckets would have a cold bias that was particularly large over the western boundary currents in winter. They developed a number of diagnostics for identifying bucket biases, including enhanced variability at near-annual frequencies in certain latitude bands and geographically coherent enhancements in the seasonal cycle. With more detailed physical modeling, it should be possible to make more detailed predictions about the behavior of smaller subsets of data, which would help to assess the validity of the models.

Statistical estimates of biases can also be assessed using internal consistency. By withholding various subsets of data and predicting their behavior from the remainder, it should be possible to estimate the reliability of the estimates.

Comparison of SST datasets (and their uncertainties), bias adjusted in different ways. In the main paper we compare two SST datasets that apply different methods of bias adjustments and consider the result in the light of the uncertainty of the estimated biases. This illustrates the importance of generating multiple versions of SST datasets with different approaches to bias adjustment and bias uncertainty estimation, preferably with several research groups contributing. Although comparing the adjusted SST fields from different analyses is insightful, cleaner comparisons—where the effects of different choices can be isolated—are more illuminating. Ideally each step of the process can be examined, from input data choice, quality control, bias adjustment, and all the assumptions and choices involved in generating SST fields from the observations. This will require the coordination of several groups internationally but would represent important progress toward fully understanding and improving the bias adjustment of SST.

S9: EXAMPLE METHOD FOR STATISTICAL ESTIMATION OF BIASES WITH A SIMPLE ERROR MODEL. The method used to construct Fig. 5 in the main paper is described in this section as an example. Figure 5a shows 1 month of SST anomalies (°C) relative to 1961–90 based on ICOADS data for ships, drifting, and moored buoys gridded following Rayner et al. (2006). Monthly mean fields of an SST anomaly on a 5° latitude × 5° longitude grid have been reconstructed using a nonstationary optimal interpolation of observations from drifting buoys only [based on a method following Karspeck et al. (2012); Fig. 5b]. The mean bias in ERI measurements from ships was then estimated by taking their difference from the field derived using drifting buoy data only, taking into account the expected error structure of ERI observations (Fig. 5c). This error structure was modeled, following Kennedy et al. (2011a), by assuming a global offset common to the full fleet of ships reporting ERI SST, a constant offset specific to each ship, and taking into account to which grid boxes each ship contributed observations. This simple implementation does not allow a ship-by-ship quantification of biases for all ships. It does however allow grid boxes that are likely to show correlated errors (because observations from the same ship have contributed data) to be identified, giving an improved estimate of biases, in this case for ERI data.

ACKNOWLEDGMENTS. We thank the three reviewers for their help with improving this paper. Funding support was provided by the following organizations: Natural Environment Research Council (Grants NE/J020788/1, NE/I030127/1, and NE/J02306X/1); the Office of Naval Research (Grant N00014-12-1-0911); Deutsche Forschungsgemeinschaft (Grant DFG VE 366/8); Ministry of Environment, Japan (ERDTF 2-1506); and BEIS/Defra (Grant GA01101).

REFERENCES

Argo, 2000: Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANOE, doi:10.17882/42182.

Ashford, O. M., 1948: A new bucket for measurement of sea surface temperature. *Quart. J. Roy. Meteor. Soc.*, **74**, 99–104, doi:10.1002/qj.49707431916.

Atkinson, C. P., N. A. Rayner, J. J. Kennedy, and S. A. Good, 2014: An integrated database of ocean temperature and salinity observations. *J. Geophys. Res. Oceans*, **119**, 7139–7163, doi:10.1002/2014JC010053.

Berry, D. I., and E. C. Kent, 2009: A new air–sea interaction gridded dataset from ICOADS with uncertainty estimates. *Bull. Amer. Meteor. Soc.*, **90**, 645–656, doi:10.1175/2008BAMS2639.1.

——, and ——, 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.

——, and ——, 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.

——, ——, and P. K. Taylor, 2004: An analytical model of heating errors in marine air temperatures from ships. *J. Atmos. Ocean. Technol.*, **21**, 1198–1215, doi:10.1175%2F1520-0426(2004)021%3C1198:AAMOHE%3E2.0.CO;2.

Bottomley, M., C. K. Folland, J. Hsiung, R. E. Newell, and D. E. Parker, 1990: *Global Ocean Surface Temperature Atlas (GOSTA)*. MIT and Meteorological Office, 20 pp. + plates.

Bourlès, B., and Coauthors, 2008: The PIRATA Program: History, accomplishments, and future directions. *Bull. Amer. Meteor. Soc.*, **89**, 1111–1125, doi:10.1175/2008BAMS2462.1.

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

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

Coppo, P., and Coauthors, 2010: SLSTR: A high accuracy dual scan temperature radiometer for sea and land surface monitoring from space. *J. Mod. Opt.*, 57, 1815–1830, doi:10.1080/09500340.2010.503010.

Donlon, C. N., and Coauthors, 2007: The Global Ocean Data Assimilation Experiment High-Resolution Sea Surface Temperature Pilot Project. *Bull. Amer. Meteor. Soc.*, **88**, 1197–1213, doi:10.1175/BAMS-88-8-1197.

Embury, O., C. J. Merchant, and G. K. Corlett, 2012: A reprocessing for climate of sea surface temperature from the Along-Track Scanning Radiometers: Initial validation, accounting for skin and diurnal variability. *Remote Sens. Environ.*, **116**, 62–78, doi:10.1016/j.rse.2011.02.028.

Emery, W., D. Baldwin, P. Schlüssel, and R. Reynolds, 2001: Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measurements. *J. Geophys. Res.*, **106**, 2387–2405, doi:10.1029/2000JC000246.

Farmer, G., T. M. L. Wigley, P. D. Jones, and M. Salmon, 1989: Documenting and explaining recent global-mean temperature changes. Final Report to the Natural Environment Research Council, Contract GR3/6565, University of East Anglia, 141 pp. [Available online from www.cru.uea.ac.uk/cru/pubs/pdf/Farmer-1989-NERC.pdf.]

Filipiak, M. J., C. J. Merchant, H. Kettle, and P. Le Borgne, 2012: An empirical model for the statistics of sea surface diurnal warming. *Ocean Sci.*, **8**, 197–209, doi:10.5194/os-8-197-2012.

Folland, C. K., 2005: Assessing bias corrections in historical sea surface temperature using a climate model. *Int. J. Climatol.*, **25**, 895–911, doi:10.1002/joc.1171.

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

Freeman, E., and Coauthors, 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.

GCOS, 2010: Implementation plan for the global observing system for climate in support of the UNFCCC (2010 update). World Meteorological Organization Tech. Doc. WMO/TD-1523, GCOS-138, 180 pp. [Available online at www.wmo.int/pages/prog/gcos/Publications/gcos-138.pdf.]

——, 2011: Systematic observation requirements for satellite-based data products for climate: 2011 update; Supplemental details to the satellite-based component of the "Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC (2010 update)." GCOS-154, WMO/IOC/UNEP/ICS, 127 pp.

Gentemann, C. L., C. J. Donlon, A. Stuart-Menteth, and F. J. Wentz, 2003: Diurnal signals in satellite sea surface temperature measurements. *Geophys. Res. Lett.*, **30**, 1140, doi:10.1029/2002GL016291.

——, P. J. Minnett, and B. Ward, 2009: Profiles of ocean surface heating (POSH): A new model of upper ocean diurnal warming. *J. Geophys. Res.*, **114**, C07017, doi:10.1029/2008JC004825.

Glahn, W., 1933: False measurements of air temperatures on ships. *Der Seewart*, **2**, 250–256.

Hanawa, K., S. Yasunaka, T. Manabe, and N. Iwasaka, 2000: Examination of correction to historical SST data using long-term coastal SST data taken around Japan. *J. Meteor. Soc. Japan*, **78**, 187–195, doi:10.2151/jmsj1965.78.2_187.

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

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

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

——, and Coauthors, 2016: Further exploring and quantifying uncertainties for Extended Reconstructed Sea Surface Temperature (ERSST) version 4 (v4). *J. Climate*, **29**, 3119–3142, doi:10.1175/JCLI-D-15-0430.1.

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

Janssen, P. A. E. M., 2012: Ocean wave effects on the daily cycle in SST. *J. Geophys. Res.*, **117**, C00J32, doi:10.1029/2012JC007943.

JCOMM, 2015: Proceedings of the Fourth JCOMM Workshop on Advances of Marine Climatology (CLIMAR-4) and of the First ICOADS Value-Added Database (IVAD-1) Workshop. JCOMM Tech. Rep. JCOMM-TR-079, 30 pp. [Available online at www.jcomm.info/index.php?option=com_oe&task=viewDocumentRecord&docID=15293.]

Kantha, L. H., and C. A. Clayson, 2004: On the effect of surface gravity waves on mixing in the oceanic mixed layer. *Ocean Modell.*, **6**, 101–124, doi:10.1016/S1463-5003(02)00062-8.

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

Karspeck, A. R., A. Kaplan, and S. R. Sain, 2012: Bayesian modelling and ensemble reconstruction of mid-scale spatial variability in North Atlantic sea-surface temperatures for 1850–2008. *Quart. J. Roy. Meteor. Soc.*, **138**, 234–248, doi:10.1002/qj.900.

Kawai, Y., and A. Wada, 2007: Diurnal sea surface temperature variation and its impact on the atmosphere and ocean: A review. *J. Oceanogr.*, **63**, 721–744, doi:10.1007/s10872-007-0063-0.

Kennedy, J. J., 2014: A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Rev. Geophys.*, **52**, 1–32, doi:10.1002/2013RG000434.

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

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

——, R. Smith, and N. Rayner, 2012: Using AATSR data to assess the quality of in situ sea surface temperature observations for climate studies. *Remote Sens. Environ.*, **116**, 79–92, doi:10.1016/j.rse.2010.11.021.

Kent, E. C., and A. Kaplan, 2006: Toward estimating climatic trends in SST. Part III: Systematic biases. *J. Atmos. Oceanic Technol.*, **23**, 487–500, doi:10.1175/JTECH1845.1.

——, S. D. Woodruff, and D. I. Berry, 2007: Metadata from WMO Publication No. 47 and an assessment of voluntary observing ships observation heights in ICOADS. *J. Atmos. Oceanic Technol.*, **24**, 214–234, doi:10.1175/JTECH1949.1.

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

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

Legler, D. M., and K. Hill, 2014: Tropical Pacific observing for the next decade. *Eos, Trans. Amer. Geophys. Union*, **95**, 196, doi:10.1002/2014EO230006.

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

Matthews, J. B. R., and J. B. Matthews, 2013: Comparing historical and modern methods of sea surface temperature measurement—Part 2: Field comparison in the central tropical Pacific. *Ocean Sci.*, **9**, 695–711, doi:10.5194/os-9-695-2013.

Maury, M. F., 1858: *Explanations and Sailing Directions to Accompany the Wind and Current Charts*. Vol. 1. W. A. Harris, 477 pp. [Available online at http://icoads.noaa.gov/reclaim/pdf/maury1858.pdf.]

McPhaden, M. J., and Coauthors, 1998: The Tropical Ocean-Global Atmosphere (TOGA) observing system: A decade of progress. *J. Geophys. Res.*, **103**, 14 169–14 240, doi:10.1029/97JC02906.

——, and Coauthors, 2009: RAMA: The Research Moored Array for African–Asian–Australian Monsoon Analysis and Prediction. *Bull. Amer. Meteor. Soc.*, **90**, 459–480, doi:10.1175/2008BAMS2608.1.

Merchant, C. J., and O. Embury, 2014: Simulation and inversion of satellite thermal measurements. *Optical Radiometry for Ocean Climate Measurements*, G. Zibordi, C. J. Donlon, and A. C. Parr, Eds., Experimental Methods in the Physical Sciences, Vol. 47, Academic Press, 489–526, doi:10.1016/B978-0-12-417011-7.00015-5.

——, and Coauthors, 2012: A 20 year independent record of sea surface temperature for climate from Along-Track Scanning Radiometers. *J. Geophys. Res.*, **117**, C12013, doi:10.1029/2012JC008400.

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

Minnett, P. J., and G. K. Corlett, 2012: A pathway to generating Climate Data Records of sea-surface temperature from satellite measurements. *Deep-Sea Res. II*, **77–80**, 44–51, doi:10.1016/j.dsr2.2012.04.003.

Morak-Bozzo, S., C. J. Merchant, E. C. Kent, D. I. Berry, and G. Carella, 2016: Climatological diurnal variability in sea surface temperature characterized from drifting buoy data. *Geosci. Data J.*, **3**, 20–28, doi:10.1002/gdj3.35.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, doi:10.1029/2011JD017187.

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

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface

temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, doi:10.1029/2002JD002670.

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

Rennie, J., and Coauthors, 2014: The international surface temperature initiative global land surface databank: Monthly temperature data release description and methods. *Geosci. Data J.*, **1**, 75–102, doi:10.1002/gdj3.8.

Slutz, R. J., S. J. Lubker, J. D. Hiscox, S. D. Woodruff, R. L. Jenne, D. H. Joseph, P. M. Steurer, and J. D. Elms, 1985: Comprehensive Ocean-Atmosphere Data Set, release 1. NOAA/Environmental Research Laboratories/Climate Research Program, 268 pp. [NTIS PB86-1 05723.]

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

——, and ——, 2003: Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997). *J. Climate*, **16**, 1495–1510, doi:10.1175/1520-0442-16.10.1495.

——, and ——, 2004: Improved extended reconstruction of SST (1854–1997). *J. Climate*, **17**, 2466–2477, doi:10.1175/1520-0442(2004)017<2466:IEROS>2.0.CO;2.

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

Takaya, Y., J.-R. Bidlot, A. C. M. Beljaars, and P. A. E. M. Janssen, 2010: Refinements to a prognostic scheme of skin sea surface temperature. *J. Geophys. Res.*, **115**, C06009, doi:10.1029/2009JC005985.

Thompson, D. W. J., J. J. Kennedy, J. M. Wallace, and P. D. Jones, 2008: A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature*, **453**, 646–649, doi:10.1038/nature06982.

Thorne, P. W., and Coauthors, 2017: Toward an integrated set of surface meteorological holdings for climate science and applications. *Bull. Amer. Meteor. Soc.*, doi:10.1175/BAMS-D-16-0165.1, in press.

Woodruff, S. D., H. F. Diaz, J. D. Elms, and S. J. Worley, 1998: COADS Release 2 data and metadata enhancements for improvements of marine surface flux fields. *Phys. Chem. Earth*, **23**, 517–526, doi:10.1016/S0079-1946(98)00064-0.

——, and Coauthors, 2011: ICOADS Release 2.5 and data characteristics. *Int. J. Climatol.*, **31**, 951–967, doi:10.1002/joc.2103.