AMERICAN METEOROLOGICAL SOCIETY

Bulletin of the American Meteorological Society

EARLY ONLINE RELEASE

The DOI for this manuscript is doi: 10.1175/BAMS-D-15-00251.1

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

Elizabeth C. Kent, David I. Berry and Giulia Carella: National Oceanography Centre, UK

John J. Kennedy, David E. Parker, Christopher P. Atkinson and Nick A. Rayner: Met Office Hadley Centre, Exeter, UK

Thomas M. Smith: NOAA/NESDIS/STAR, USA

Shoji Hirahara: Global Environment and Marine Department, Japan Meteorological Agency, Tokyo, Japan and ECMWF

Boyin Huang and Huai-Min Zhang: NOAA's National Centers for Environmental Information, Asheville, NC, USA

Alexey Kaplan: LDEO of Columbia University, USA

Yoshikazu Fukuda: Japan Meteorological Agency, Japan

Masayoshi Ishii, Climate Research Division, Meteorological Research Institute, Tsukuba, Ibaraki, Japan

Philip D. Jones: University of East Anglia, Climatic Research Unit, School of Environmental Sciences, UK and Center of Excellence for Climate Change Research, Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia

Finn Lindgren: University of Edinburgh, UK

Christopher J. Merchant and Simone Morak-Bozzo: University of Reading, UK

Victor Venema: University of Bonn, Germany

Souichiro Yasui: Global Environment and Marine Department, Japan Meteorological Agency, Tokyo, Japan

Corresponding author: Elizabeth C. Kent, National Oceanography Centre, Southampton, SO14 3ZH, UK. eck@noc.ac.uk

Capsule Summary

Global surface-temperature is a fundamental measure of climate change. We discuss bias estimation for sea-surface temperature and recommend the improvements to data, observational metadata, and uncertainty modeling needed to make progress.

Abstract

Global surface-temperature changes are a fundamental expression of climate change. Recent, much-debated, variations in the observed rate of surface-temperature change have highlighted the importance of uncertainty in adjustments applied to sea-surface temperature (SST) measurements. These adjustments are applied to compensate for systematic biases and changes in observing protocol. Better quantification of the adjustments and their uncertainties would increase confidence in estimated surface-temperature change and provide higher-quality gridded SST fields for use in many applications.

Bias adjustments have been based either on physical models of the observing processes or on the assumption of an unchanging relationship between SST and a reference data set such as night marine air temperature. These approaches produce similar estimates of SST bias on the largest space and timescales, but regional differences can exceed the estimated uncertainty. We describe challenges to improving our understanding of SST biases. Overcoming these will require clarification of past observational methods, improved modeling of biases associated with each observing method, and the development of statistical bias estimates that are less sensitive to the absence of metadata regarding the observing method.

New approaches are required that embed bias models, specific to each type of observation, within a robust statistical framework. Mobile platforms and rapid changes in observation type require biases to be assessed for individual historic and present-day platforms (i.e., ships or buoys) or groups of platforms. Lack of observational metadata and of high-quality

observations for validation and bias model development are likely to remain major challenges.

1. Background

The global surface temperature record is constructed by blending sea-surface temperature (SST) with air temperature over land and ice (see also section S1 of the supplemental material). Both SST and land-air temperature require adjustments to account for changes such as in depth or height of measurement, instrumentation, and siting. Improvement of estimated biases in historical measurements of SST will have a major effect on estimates of global surface temperature change and their uncertainty (Jones 2016).

The historical record of observations of the temperature of water at the "sea surface" is a disparate collection of measurements made using different methods from different measurement platforms. Most measurements come from platforms that move (mostly ships and drifting buoys) with relatively few providing time series at fixed locations (e.g., ocean weather ships, fixed platforms, coastal installations or moored buoys). Adjustment of near-surface air temperatures over land, often called homogenization, relies on comparisons of a candidate station with nearby stations to identify and correct unphysical changes (Trewin 2010). The continually evolving, and largely mobile, marine observing system means that such approaches cannot be easily applied to marine observations.

Folland *et al.* (1984) applied first-order SST bias adjustments, adding a constant value of 0.3°C to observations made before 1942, based on the difference between global night marine air temperature (NMAT) and SST. By the time of the Intergovernmental Panel on Climate Change (IPCC) First Assessment Report (Houghton *et al.* 1990), more complex models of SST bias had been developed (Jones *et al.* 1986, Bottomley *et al.* 1990) and presently several different estimates of SST bias exist. Figure 1 shows global mean SST anomalies for the

current, commonly-used, long-term gridded SST analyses: HadSST3 (Kennedy *et al.* 2011a, b); ERSSTv4 (Huang *et al.* 2015); and COBE-SST2 (Hirahara *et al.* 2014), along with their bias estimates and uncertainties.

SST observations and gridded datasets underpin many thousands of published research papers every year, including their use as boundary conditions for atmospheric reanalysis, so the benefits of improved SST bias estimation are wide-reaching. However, severe challenges arise because the observations we have are not from a dedicated climate observing system. Early observers were largely concerned with navigation and safety. Observations were collated to document climatology rather than climate change. Detailed information on the ships and the different methods of measurement, now known to be of immense value to assess changes, has been lost. Different measurement methods have different characteristic biases, and there are variations peculiar to individual platforms and installations. The characteristic biases also depend on environmental conditions such as wind speed, solar radiation and air-sea temperature contrasts, as does the real variability of ocean temperature, with further real variations due to the depth of measurement. Reconciling all of this to make consistent estimates of SST changes would be a challenge with good documentation. The patchy availability of observational and platform metadata, and sparse sampling in some regions and periods, makes it even harder.

The first-order bias adjustments required to account for changes in methods of SST observation over the past more than 150 years are known. We know that adjustments are required and the direction and approximate size of the change at very large scales. However, comparison of the different approaches used to estimate SST bias adjustments shows that differences remain that are hard to fully explain. Unexplained differences occur at smaller scales and in periods where measurement methods change quickly. This shows the need to

103 better understand the biases, improve adjustment methods and refine the uncertainty
104 estimates.

105 Our recommendations to improve the situation are in four areas. Firstly enhancement of the
106 source archive to provide more observations, more complete metadata and improve quality.
107 Second is a need to develop better models of SST bias, and to maintain a range of SST
108 products using different approaches to bias adjustment. Thirdly there is a need for accessible,
109 high-quality, consistent validation data sets to be assembled from existing archives and for
110 the availability of such data to be established as metrics for assessing the observing system.
111 Finally we would like to see more people working in this area and suggest how barriers to
112 getting started might be reduced.

113 **2. What is SST and how is it measured?**

114 *2.1 What is SST?*

115 The temperature of the water near the sea surface varies on all space and time scales. The
116 term SST has typically been used to describe the mean temperature of the upper few meters
117 of the ocean. Historically measurements taken at depths from the surface and down to about
118 20 m have all been assumed representative of the SST. Under well-mixed conditions this is a
119 good assumption. However, there are well-known variations of ocean temperature with
120 depth, especially at low wind speeds and sunny conditions (Kawai and Wada, 2007).
121 Developers of long-term datasets have taken a pragmatic approach, assuming either that
122 measurements represent well-mixed conditions, or that conditions were well-sampled and
123 therefore representative of the surface layer even if it was not well-mixed. When considering
124 biases, it is necessary to consider spatial differences in the depth dependence of temperature.
125 Further discussion on the definition of SST and its uncertainty can be found in Section S2 of
126 the supplemental material.

2.2 How is SST measured?

SST has been measured in different ways over the past 200 years. The observations record real variations in temperature but also contain an imprint of how they were measured. Both the real variations and the biases are affected by the ambient environmental conditions, making them hard to disentangle.

The earliest observations were probably made by sampling seawater in a bucket. Maury (1858) recommended wooden buckets which were likely used around this time. The type of bucket used evolved over time, with canvas buckets becoming predominant, later replaced by better-insulated rubber and plastic buckets. Figure 2a summarizes the different factors that can cause bias in observations of SST made using buckets.

For measurement, the bucket is thrown into the water to collect a sample. The exact depth of sampling is unknown, but is close to the surface, especially if the ship is moving fast. If the bucket is at a very different temperature from the water, or contained water from a past sample, then the time the bucket spends in the water to equilibrate is important. We do not know how much care the observers took in following instructions on sampling protocol in this regard, nor in others. Once a bucket leaves the sea, both the bucket and water sample exchange heat with the atmosphere in a way that is dependent on their volume, thermal properties and the environmental conditions. The temperature continues to change while the thermometer is read; the change is related to the length of time taken to get a stable reading, and whether the bucket is taken out of the wind and/or into the shade. The initial temperature and response time of the thermometer can also influence the reported temperature.

For ships with engines, the temperature of water pumped onboard to cool the engines can be used as an estimate of SST (Figure 2b). Sampling is usually deep as the inlet has to be below the surface whatever the loading of the ship. The ship may also mix the water, so the

effective depth of sampling is ambiguous even if the inlet depth is known. Typically, most details of the installation are unknown, so it is hard to determine how an observation might be affected by heat exchange between the inlet and the point of measurement. Historically, there is evidence for inaccurate thermometers and poor installation (Kent and Taylor 2006). An extensive analysis of engine-room intake (ERI) observations by James and Fox (1972) showed ERI SSTs, at that time, were particularly warm for large ships with thermometers more than 3 meters inboard from the inlet. Technological developments have likely resulted in thermometers placed nearer to the hull (possible with remote-reading automatic sensors) and further from the engine-room. The type of ERI thermometer was also important with precision thermometers and thermistors showing smaller offsets relative to bucket measurements than mercury or other types of thermometer. There is some evidence that ERI biases have reduced over time (Kent and Kaplan 2006), which could be explained by better thermometers or improved siting. Determining a ship-by-ship estimate of mean ERI bias would represent a significant advance, perhaps permitting more subtle variations due to greater measurement depths or ship speed to be explored.

Hull-mounted sensors (also shown in Figure 2b) are dedicated SST sensors. Kent *et al.* (1993) showed, for a small subset of ships, that hull sensors were more accurate (smaller bias and noise) than ERI, but good insulation is required (Beggs *et al.* 2012). A wider analysis of hull sensor accuracy in the field is long overdue.

Surface drifting buoys (Figure 2c) measure at shallow depths, nominally 10-20cm. Biases in drifter measurements might arise due to error in sensor calibration, temperature calibration "drift" while deployed, or bio-fouling on the sensor. Drifting buoys presently provide measurements of SST that are near-globally distributed and have better accuracy than from ships (Kennedy *et al.*, 2011c), since problems with early drifters were resolved (Bitterman and Hansen, 1993). Careful quality control is still required to identify spurious spikes in

reported position or SST measurements from when the buoy is out of the water (due to pre-deployment data transmission, beaching or human interference) and instrument failure or other causes of erroneous data (Lumpkin *et al.* 2012, Atkinson *et al.* 2013). Observations made available in delayed mode (e.g. by Integrated Science Data Management (ISDM) or the Atlantic Oceanographic and Meteorological Laboratory) typically have quality control flags appended, but checks of ICOADS have revealed additional problematic reports in both delayed mode (from ISDM) and real time data (Atkinson *et al.* 2013).

Moored buoys produce continuous measurements at fixed locations at a depth of about 1m or at several predetermined depths (Kennedy 2014), typically only near coasts or in tropical regions. The mechanisms causing their biases are similar to those for surface drifters but it is often possible to recover instrumentation from moored buoys for recalibration, improving their overall accuracy.

2.3 Availability of observations and ancillary information

SST observations were first made available in the 19th Century as charts to aid navigation (Rennell 1832; Maury 1858). Much later, national compilations of marine observations were used to generate gridded analyses of SST for scientific applications (e.g., Bunker 1976; Bottomley *et al.* 1990). The US national collection developed into a publicly available databank (Woodruff *et al.* 1987) which became the International Comprehensive Ocean-Atmosphere Data Set (ICOADS), currently on Release 3.0 (Freeman *et al.* 2016). ICOADS is the preferred source for constructing historical SST analyses, providing traceability of the data, simpler comparison among derived data products and access to newly digitized observations (e.g., Allan *et al.* 2011) and to observational metadata (Kent *et al.* 2007). Moreover, it enables a dialogue that can lead to improvements in ICOADS and in the many ICOADS-derived datasets (JCOMM 2015).

Quantifying SST bias ideally requires accurate location and time information, platform information, complete information of methods, instruments and protocols used, and of the ambient conditions (Figure 2). ICOADS contains some of the information required (described in Section S3 of the supplemental material), but its availability is patchy. We make recommendations that will enhance the amount of SST data and metadata available by digitization of data and metadata from ships logbooks (Recommendation 1), by reprocessing of the existing ICOADS archive (Recommendation 2) and by improved use of external sources of observational metadata (Recommendation 3).

3. Current approaches to SST Bias Estimation

3.1 Physics-based bias models

The factors affecting bucket SST measurements are well-known (Figure 2a) and have been discussed since the time of Maury (1858). The heat exchange experienced by a water sample in a bucket can be estimated with a physical model (Folland and Parker (1995), hereafter FP95). The bucket is represented by a partly-closed cylinder with appropriate thermal properties: uninsulated for canvas buckets, partly insulated for wooden buckets. More difficult is applying these models to historical measurements made using buckets of unknown dimensions and thermal properties in environmental conditions that are also not well-known. The approach of FP95 to this problem, as used in HadSST3 and COBE-SST2, is summarized in Section S4 of the supplemental material. Recommendation 4 addresses the need for simplified physical models of SST biases from buckets and better estimates of the thermodynamic forcing required.

Physical models for biases in ERI SSTs have not been developed as the detailed information required on individual installations (Matthews and Matthews 2013) is almost always

unavailable (Figure 2b). Similarly the estimation of bias in hull sensors has not yet been tackled with physically-based models.

Although drifter and moored-buoy SSTs are usually considered to be bias free, adjustments for their differences relative to ship-derived SSTs are typically made (Kennedy *et al.* 2011b, Hirahara *et al.* 2014, Huang *et al.* 2015). This choice has been shown to have little effect on long term trends (Kennedy *et al.* 2011b).

Physical models for the ocean cool-skin effect and for thermal stratification within the upper few meters of ocean (which can be significant during day-time if mixing is small) are used to relate satellite SSTs to SST at the depths representative of buoys (Merchant *et al.* 2012). The models are driven by weather-analysis fields, and have skill in reconciling satellite and sub-surface measurements (Embury *et al.* 2012). Such models could be used to inform comparisons of *in situ* measurements made at different depths.

3.2 Application of physics-based models

The two main barriers to the application of physical-correction models are uncertainty in the measurement method used and in the environmental conditions pertaining to individual observations. Section S3 of the supplemental material describes the information available in ICOADS to determine the type of platform and measurement method.

Kennedy *et al.* (2011b) brought together evidence from ICOADS, external sources of measurement metadata (such as that published by the WMO in "Publication No. 47", hereafter "Pub. 47", Kent *et al.* 2007), and other documentary information, to estimate measurement methods and their uncertainties (Figure 3). They weighted bias estimates for each method to produce estimated fields of the unbiased SST. Method weightings, and bias estimates, were varied within plausible ranges to produce an ensemble of SST fields spanning the likely uncertainty. In contrast, Hirahara *et al.* (2014) approached the problem by

estimating the proportions of different methods from differences in the data. They assumed a bias model for each type (insulated bucket, uninsulated bucket or engine intake) to adjust observations where the method was known. Proportions of observations with unknown method were then assigned to the different methods such that global SST averages from observations with unknown methods agreed with SST averages from known methods when combined with the method-dependent bias models. These approaches show broad agreement in inferred measurement methods (Figure 3b). Notable discrepancies include estimates of the rate of transition from uninsulated to insulated buckets (Kennedy, 2014).

Once the measurement method has been assigned, the bias adjustment can be calculated using the appropriate bias model. This is presently done simply: bucket bias adjustments are applied using the fields calculated by FP95 weighted by the proportions of observations thought to be made using wooden, canvas or rubber buckets (Kennedy *et al.* 2011b, Hirahara *et al.* 2014). The relative biases between ships and drifting buoys are fixed. Biases for ERI or hull sensors are fixed in the COBE-SST2 analysis, and vary within an estimated range in the HadSST3 analysis.

3.3 Large-scale statistical adjustments using air temperature

A statistical approach to bias adjustment of ship observations was developed by Smith and Reynolds (2002, hereafter SR02) based on large-scale differences between SST and NMAT measured from ships. The rationale is that biases in NMAT are more straightforward to adjust (Kent *et al.* 2013, supplemental material Section S1) and that the large-scale differences between SST and NMAT will not vary markedly over time (Huang *et al.* 2015). NMAT, rather than all-hours MAT, is used to avoid uncertainty due to daytime heating on ships. Details of the SR02 statistical bias model and its implementation by Huang *et al.* (2015) are described in the supplemental material Section S6.

This method does not need the detailed information required by physical models, but there are still uncertainties. Any residual biases in adjusted NMAT will influence the SST bias estimates (Rayner *et al.* 2003, Kent *et al.* 2013) and uncertainty in NMAT will propagate through to the SST estimates. Although NMAT variations are representative of SST variations on the largest scales (Huang *et al.* 2015), the relationship is likely to be locally weaker. The computed spatial patterns of SST-NMAT are critical for the estimate, and assuming that the patterns are well-known and invariant over time also introduces uncertainty. SR02 originally used the bias model only in the pre-World War 2 (WW2) period dominated by bucket measurements (Figure 3). Huang *et al.* (2015) extended the method throughout the record and generated an ensemble to explore uncertainty (described in supplemental material Section S6).

Recommendation 5 calls for the extension of statistical-based modeling of SST biases beyond large-scale adjustments based on NMAT.

4. Comparison and evaluation of estimates of SST bias

4.1 Comparison of bias estimates

The first test of the different bias adjustments is whether the estimates agree within their uncertainty ranges. Figure 4 compares the bias adjustments from HadSST3 and ERSSTv4. In these datasets the sensitivity of the bias estimates to assumptions and values chosen for internal parameters (parametric uncertainty, Kennedy 2014) has been quantified through making plausible perturbations to each of these choices to create an ensemble of bias estimates spanning the known uncertainty in the method (the supplemental material describes the calculation of the ensembles in Sections S4 and S6). Figure 4 illustrates the differences between the bias adjustment in the context of the range of the uncertainty ensembles and shows that, by this measure, we don't yet fully understand the biases and their uncertainties at

all times throughout the record. Maps showing average spatial variation of the biases averaged over 1890 to 1919 (Figures 4a, c) show differences that exceed the range of their combined uncertainty ensembles over large regions (Figure 4e). Even in the more recent period 1995 to 2004 (Figures 4b, d) there are regions where the difference exceeds the ensemble range (Figure 4f). Zonal mean (Figure 4g) and global average differences (Figure 4h) show that during these periods the large-scale biases are relatively well-understood, albeit with compensating bias differences with latitude giving global average agreement within uncertainty in the earlier period. Differences in the bias adjustments fall outside the ensemble range in two periods: at the start of the record (before about 1880), and around the 1980s. In the early period both SST and NMAT data are sparse so it is not surprising that our understanding is limited. The later period, from the late 1970s to the early 1990s is where the proportion of SST observations made by ERI is increasing (Figure 3), and the buoy observing system for SST is not yet well-established. Figure 4h suggests that the discrepancy is likely to arise from an underestimate in uncertainty during this period. However, improving our understanding of *in situ* SST bias during this period is necessary if the data are to be used with confidence to produce adjustments or validation for satellite-derived estimates of SST. The period around WW2 is known to be problematic (e.g., Thompson *et al.* 2008) as making observations became dangerous, especially at night when the use of lights could attract an attack. During WW2 a greater proportion of observations are made during daylight hours, engine intake measurements were preferred to buckets, and buckets may have been carried inside: all tending to give a warm bias. The WW2 period shows rapid variations in the difference between the bias estimates (Figures 4g and 4h), but also a large ensemble range, so by this metric these differences are understood, albeit very uncertain. Such comparisons can help to focus attention on periods and regions where differences are large (e.g., prior to about 1880 or in Tropical and high latitude regions prior to the mid-1990s), when uncertainties are

large (e.g., during WW2) or where the uncertainty may be underestimated (e.g., during the 1980s).

The comparison shows we are yet to fully reconcile the biases in all types of SST observations throughout the historical record. It also shows that improvements in uncertainty estimation must go hand-in-hand with improvements in bias estimates. Nevertheless, uncertainties in the bias adjustments are not thought to be large enough to alter the conclusion that global SSTs have increased over the historical record (Hartmann *et al.* 2013). However, confidence in regional adjustments is lower than for the global mean as the spatial patterns predicted by the different methods do not agree well (Figure 4 e-g, also Huang *et al.* 2015 and supplemental material Section S7). Uncertainty due to under-sampling can be large in some regions and periods (Kennedy 2014), particularly early in the record (Hirahara *et al.* 2014) and outside major shipping lanes prior to the extension of coverage provided by drifting buoys (Zhang *et al.* 2009).

Such comparisons of different estimates of the bias, or (less directly) data sets adjusted in different ways are a good first step toward understanding uncertainty in bias adjustments. A range of different approaches to bias estimation should be maintained and compared (Recommendation 6). However, more is learned by disagreement than by agreement, and in order to evaluate the estimated biases an independent reference is needed.

4.2 Evaluation by comparison with independent data

Comparisons with validation data should cover a range of diagnostics including mean bias and variance relative to validation data evaluated across a range of locations and throughout the annual and diurnal cycles. Attention should be paid to differences arising from the depths of the measurements.

In the modern period – from the mid-1990s – there are multiple sources of validation data for estimation of biases in SST observations from ships. Drifting and moored buoys take measurements of better accuracy and stability than is routinely obtained by shipboard measurements. Argo floats provide accurate data, but low sampling rates, and can be used for validation after about 2005. Some satellite data sets covering the 1990s to present are of the desired accuracy, and largely independent of the *in situ* record (Merchant *et al.* 2012, 2014) and therefore suited to validation or independent assessment of SST bias adjustments applied to ship observations. Validating over longer time scales is more difficult. Drifting buoys can be used back to the early 1990s before which there was no standardized design. Oceanographic measurements are available (Gouretski *et al.*, 2012), but are also affected by biases (Cheng *et al.* 2016) and seldom numerous. Ocean weather ships and underway observations from research vessels are potential sources of validation data. Although they may be affected by biases, there is a greater chance of obtaining a full set of high-quality marine meteorological variables and metadata. Work is ongoing to extend independent satellite SST records back to the early 1980s, but the achievable stability of observation is as yet unknown. Careful consideration must be given to the uncertainty inherent in all these data sources.

Extending validation to a wider range of comparison data sets would be valuable. Careful analysis is required if comparisons are made with different parameters (such as air temperature), with coastal observations (that might not be fully representative of open-ocean conditions) or with observations that may have their own biases. Records with consistent instrumentation over the several decades when the observing system was in flux could be valuable – perhaps records from harbor logs, lighthouses or atolls should be considered. Land-station air temperature data from other regions could also be used indirectly via experiments with climate models run with prescribed SSTs bias adjusted in different ways

(e.g., Folland 2005). An overview of potential validation data is given in Section S8 of the supplemental material. Recommendation 7 outlines the need for improved accessibility and management of existing potential sources of validation data. Recommendation 8 considers how the need for consistent and high quality observations can be built into observing-system adequacy requirements.

4.3 Evaluation using measures of internal consistency

The different types of bias can leave their own characteristic fingerprint on the SST record. For example, FP95 showed that there were signals in the data, related to the seasonal cycle, which could be explained by the characteristic biases in bucket measurements. In this case a measure of the effectiveness of the bucket bias adjustment would be the removal of spurious signals in the seasonal cycle of SST. In another example Kennedy *et al.* (2011b) showed that adjustments applied to ERI and bucket measurements improved agreement between these two subsets of data from the 1950s on.

Separating data into two datasets, one used for estimation and training and the other for validation, is a good general approach. This is widely used in assessing statistical techniques and might be applied to existing statistical methods of bias estimation (e.g., SR02). The method can also be applied more generally by setting aside a subset of data for validation, preferably a subset of known high quality that is not used in the estimation or correction of biases. Unfortunately, the data most suitable for validation also have great value for estimating biases. The price paid for having a data set with credible, validated, uncertainty estimates might be a slightly-higher overall uncertainty; the alternative is a lower overall uncertainty that was impossible to assess fairly. Research vessel data and Argo data, that are not yet widely used in historical SST data sets might be used to validate modern periods. Newly digitised data could be used for historical assessments. A degree of independence should also be maintained between the institutions producing bias adjustments and those

393 performing validation. This could be achieved if validation were carried out by an
394 organization independent of the dataset developers, or by using a standard set of widely
395 agreed criteria and comparisons.

396 To date, the evaluation of bias adjustments using measures of internal consistency has been
397 limited. The development of bias-adjustment methods to be applied to individual
398 observations or to data from individual ships would enable the extension of this type of
399 evaluation to other metrics including perhaps a consistent representation of diurnal variations
400 or a minimization of ship-to-ship differences.

401 **5. Priorities for the future**

402 *5.1 Improvements to data and metadata*

403 Fundamentally, there is scope for improvements to ICOADS. Although ICOADS is often
404 thought of as "raw" data, it is derived from a larger, more heterogeneous, underlying
405 databank from diverse sources. Further reprocessing of the databank could help to better
406 resolve duplicate observations, incomplete ship identifiers, scale conversions, missing
407 metadata, and positional errors amongst other basic problems (Recommendation 2). The
408 recent addition (Release 2.5.1 and later) of unique IDs (UID) to each report in ICOADS is
409 tremendously helpful. Tying quality control information and metadata studies back to the
410 ICOADS via the UID and sharing code and methods will improve traceability, promote
411 collaboration and help new researchers enter the field (Recommendation 9).

412 Much is to be gained from improvements to metadata (Recommendations 1-3). Ship tracking
413 – the association of individual reports into coherent voyages (Carella *et al.* 2015) – will
414 enable the better characterization of ship-by-ship biases and other errors. Bringing together
415 known sources of metadata into a single repository would be a step towards a more holistic
416 synthesis. A start has been made on inferring absent metadata (Kent *et al.* 2007, Kent *et al.*

2010, Kennedy *et al.* 2011b, Hirahara *et al.* 2014, Carella *et al.* 2015) and resolving conflicts that arise when different sources present inconsistent information, but more needs to be done.

A barrier to the use of recent marine data from ships is the decision by some countries to anonymize ship reports. The reasons often given are that the information has commercial value, or that there are concerns about security. Whatever the reason, it prevents the matching of ships to the relevant metadata in Pub. 47. We hope that a solution can be found to provide this information in a way consistent with the safety of the vessels, if not in real time, then after an appropriate delay.

There is also a need for existing sources of high-quality independent validation data to be collated. While such compilations exist for e.g., Argo and drifting buoy observations, complete authoritative archives of data and metadata do not exist for moored buoys, Ocean Weather Ships or Research Vessels. Land-based coastal observations are difficult to identify in global and regional archives and multi-variate records are often fragmented (Thorne *et al.* 2016). A consistent approach to the management of such high-quality observations, quality assured by experts in each data type, would be valuable for the validation of SST biases (Recommendation 7). The need for such consistent observations, and their appropriate management should be recognized in climate observing-system requirements (Recommendation 8)

5.2 Improvements to physically-based models of SST bias

Development of the physical models used to estimate bucket biases should continue. Models will be most valuable if independently tested in well-designed experiments under controlled laboratory conditions and at sea. Well-validated physical models will give improved estimates of the expected mean biases, their uncertainties, and allow the possibility of estimating biases for each observation individually. Careful experimental design is needed

before undertaking expensive and time-consuming measurements at sea. Simplified parameterizations of the bucket models are needed for application to a wider range of bucket designs including modern insulated buckets (Recommendation 4).

To drive physical models, we need to understand the inputs to those models and their uncertainties. Estimates of air temperature, humidity, cloud, and wind speed and direction are all needed and all are affected by biases comparable in magnitude to those affecting SST (Berry *et al.* 2004, Willett *et al.* 2008, Berry and Kent 2011, Eastman *et al.* 2011, Thomas *et al.* 2008). Reanalyses may prove a valuable tool for understanding the expected spatio-temporal variability of bucket-related SST biases and could reveal components of bias variability related to weather and longer-term effects (Recommendation 4). It might be expected that as understanding of these dependencies increases, the estimated random error of the measurements, which is partly an aggregation of many unresolved systematic processes, will decrease. Improved bias estimates will consequently need to go hand-in-hand with revisions to estimates of other components of the uncertainty.

Some other biases are not easily modeled. It may be impossible to derive meaningful physically-based estimates of bias for an individual ERI installation (Figure 2b) so these ship-specific biases may need to be characterized statistically.

5.3 Improved statistical approaches

SST biases are statistically and computationally challenging. There are several hundred million *in situ* observations in ICOADS. This data volume is modest by modern standards, but complexity arises because the data are from diverse sources representing reports from perhaps hundreds of thousands of individual ships and buoys, some uniquely identified, some not. The data are of varied quality. Metadata are sometimes incomplete or conflicting. Reference observations are few and not always of unimpeachable quality. Improved

statistical methods are required to advance and capitalize fully on the improvements in the basic data and modeling described above. Progress is likely to come from working more closely with statisticians, data scientists and computational experts to develop state-of-the-art analysis systems. It may also be possible to adapt methods developed for the homogenization of land station data (Venema *et al.* 2012).

It is possible to write a system of equations encapsulating a full statistical description of the problem of estimating spatially-complete unbiased fields, and their uncertainty, from sparse, noisy and biased measurements of SST. In practice, however, the terms in these equations are subject to the same effects causing uncertainty in the current approaches. For example, the form of the method-dependent bias model must still be specified. Solving even a simplified version at coarse resolution is presently computationally challenging. The goal is to include all we know about SST biases into a holistic, statistically rigorous, Bayesian analysis framework. The framework should embed method-dependent physically-based bias models within a full description of the correlation structure of the variability of SSTs and their biases (Recommendation 5).

Elements of such a holistic statistical approach are now being developed. The UK Met Office is developing methods to generate SST fields using estimates of the correlation structures of variability associated with both real changes in SST and biases. In this approach, individual ship biases and their uncertainties can be identified (Figure 5). This relatively simple implementation, described in more detail in Section S9 of the Supplemental Material, is able to identify biased measurements made by individual ships, and could reduce the obvious SST artifacts related to "ship tracks" often present in SST analyses.

Everything we have learnt from the existing approaches can feed into new statistical models. Every scrap of information about the structure of expected biases can be used to constrain and inform statistical analyses. Further constraints could also be applied, such as a large-scale

consistency with NMAT. The development of improved statistical models should proceed in tandem with efforts to better characterize the observations and their biases.

5.4 Maintaining research effort and extending the community

Huge progress has been made since the first estimates of SST bias were published in 1984. There are currently three families of SST datasets available that take different approaches to bias adjustment (HadSST/HadISST, ERSST and COBE). However all still use approaches that are essentially adaptations of methods originally developed decades ago. We now need to develop new approaches to bias adjustment that take advantage of recent advances in statistical methods and computing power (Recommendation 5) while maintaining a diversity of different methods (Recommendation 6). Diversity of methods helps quantify structural uncertainty: the spread between datasets arising from fundamental choices in analysis method and assumptions underlying them that are difficult and, in many cases, impossible, to capture by varying the parameters or modules within a single analysis system (Thorne *et al.* 2005).

Progress has been slower than we would like as the number of researchers active in the area is small and fresh perspectives would be welcome. There are many barriers to new researchers entering this area; presenting the data and metadata in accessible ways and providing a range of different types of documentation is essential to engage a wider community in assessment and validation (Recommendation 9).

Recommendations

Recommendation 1. Add more data and metadata to ICOADS

Additional observations of SST and associated variables such as air temperature, humidity, wind, cloud, pressure and weather information recovered from logbook digitization will help improve estimates of SST and SST bias. Every effort should be made to retain observational metadata and to keep multi-variate observations together.

Recommendation 2. Reprocess existing ICOADS records

Older ICOADS acquisitions are often lacking metadata and compromised by legacy deficiencies in data management and storage formats. A full reprocessing of ICOADS legacy data, alongside improvements to data formats, would improve SST bias adjustment through improved ship tracking, recovery of information on platform identity, better identification of mispositioned and duplicate reports, better quality control, and recovery of additional data and metadata from the existing reports. A critical review of all input ICOADS data sources should be carried out to ensure that ICOADS contains the best available data, metadata and quality information.

Recommendation 3. Improve information on observational methods

A comprehensive review of documentary sources will better constrain the uncertainty in methods and protocols for historical observations. ICOADS callsign recovery and reprocessing of WMO Pub. 47 metadata will help link observations to metadata for individual ships.

Recommendation 4. Improve physical models of SST bias

Simplified and validated physically-based models of SST bias are required along with better estimates of ambient conditions and understanding of how to use those estimates to drive the models.

Recommendation 5. Improve statistical models of SST bias

More holistic and powerful statistical approaches to the problem of estimating SST biases and their uncertainties are needed, especially to study presently unknown causes for inhomogeneities.

Recommendation 6. Maintain and extend the range of different estimates of SST bias

SST datasets and gridded analyses will continue to improve, but will never become identical. A wider range of bias estimates taking different approaches to adjustment will enable improved understanding of structural uncertainty. Carefully designed comparisons including all the developers of bias-adjusted SST analyses will improve understanding of biases and their uncertainties.

Recommendation 7. Expand data sources for validation and extend use of measures of internal consistency in validation

Resources for validating SST bias adjustments include SST from satellites and ocean reanalyses, as well as observed air temperatures, albeit with their own uncertainties. Collating, assembling and extending consistent datasets providing validation sources will enable more thorough validation of SST bias adjustments. Such sources include ocean weather ships, research vessels, moored buoys, land-based coastal stations and independent satellite SST records. A more imaginative approach is required to make best use of available validation data and to widen the use of measures of internal consistency in SST bias validation.

Recommendation 8. Ensure adequacy and continuity of the observing system

It is important that the challenges we have encountered in understanding the historical SST record do not persist into the future. Requirements for consistency, metadata, subsets of high-quality validation data, and appropriate curation for climate applications should be integrated into the metrics for assessing observing system adequacy and performance (e.g., GCOS 2010).

Recommendation 9. Improve openness and access to information

Despite the complexity of the problem, SST bias adjustment has only been tackled by a small number of small groups producing SST products. Many aspects of the problem are potentially of much wider interest to: physicists, metrologists, historians, computer scientists and statisticians amongst others. Providing modular software tools, improved access to data, metadata and historical documentation will help to widen the range of approaches to the important, complex and interesting problem of SST bias adjustment.

Acknowledgements

We thank the 3 reviewers for their help in improving this paper.

References

Allan R., P. Brohan, G. P. Compo, R. Stone, J. Luterbacher, and S. Brönnimann, 2011: The International Atmospheric Circulation Reconstructions over the Earth (ACRE) Initiative. *Bull. Amer. Meteor. Soc.*, **92**, 1421–1425. doi: 10.1175/2011BAMS3218.1.

Ashford, O. M., 1948: A new bucket for measurement of sea surface temperature. *Quart. J. Roy. Meteor. Soc*., ***14,*** 99-104, doi: 10.1002/qj.49707431916.

Atkinson, C. P., N. A. Rayner, J. Roberts-Jones, and R. O. Smith, 2013: Assessing the quality of sea surface temperature observations from drifting buoys and ships on a platform-by-platform basis, *J. Geophys. Res. Oceans*, **118**, 3507–3529, doi: 10.1002/jgrc.20257.

Beggs, H. M., R. Verein, G. Paltoglou, H. Kippo and M Underwood, 2012: Enhancing ship of opportunity sea surface temperature observations in the Australian region. *J. Operational Oceanography*, **5**(1), doi: 10.1080/1755876X.2012.11020132

Berry, D. I. and E. C. Kent, 2011: Air-Sea Fluxes from ICOADS: The Construction of a New Gridded Dataset with Uncertainty Estimates, *Int. J. Climatol.*, **31**(7), 987–1001, doi: 10.1002/joc.2059.

582 Berry, D. I., E. C. Kent and P. K. Taylor, 2004: An analytical model of heating errors in
583 marine air temperatures from ships, *J. Atmos. Ocean. Tech*, **21**(8), 1198 - 1215, doi:
584 10.1175%2F1520-0426(2004)021%3C1198:AAMOHE%3E2.0.CO;2.

585 Bitterman, D. S., and D. V. Hansen, 1993: Evaluation of sea surface temperature
586 measurements from drifting buoys. *J. Atmos. Ocean. Tech*, **10**(1), 88-96. doi:
587 10.1175/1520-0426(1993)010<0088:EOSSTM>2.0.CO;2.

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

590 Brooks, C. 1926: Observing water-surface temperatures at sea. *Monthly Weather Review*,
591 **54**(6), 241-253, doi: 10.1175/1520-0493(1926)54<241:OWTAS>2.0.CO;2.

592 Brooks, C., 1928: Reliability of different methods of taking sea-surface temperature
593 measurements. *Journal of the Washington Academy of Sciences*, 18, 525-545. doi:
594 10.1029/TR009i001p00076-1.

595 Bunker, A. F. 1976: Computations of Surface Energy Flux and Annual Air–Sea Interaction
596 Cycles of the North Atlantic Ocean. *Mon. Wea. Rev*., **104**, 1122–1140. doi: 10.1175/1520-
597 0493(1976)104<1122:COSEFA>2.0.CO;2.

598 Carella, G., E. C. Kent and D. I. Berry, 2015: A probabilistic approach to ship voyage
599 reconstruction in ICOADS, *Int. J. Climatol.*, early view, doi: 10.1002/joc.4492.

600 Cheng, L., J. Abraham, G. Goni, T. Boyer, S. Wijffels, R. Cowley, V. Gouretski, F.
601 Reseghetti, S. Kizu, S. Dong, F. Bringas, M. Goes, L. Houpert, J. Sprintall, and J. Zhu,
602 2016: XBT Science: Assessment of Instrumental Biases and Errors. *Bull. Amer. Meteor.*
603 *Soc.* **97**, 924–933, doi: 10.1175/BAMS-D-15-00031.1.

Eastman, R., S. G. Warren, and C. J. Hahn, 2011: Variations in cloud cover and cloud types over the ocean from surface observations, 1954-2008. *J. Climate*, **24**(22), 5914-5934. doi: 10.1175/2011JCLI3972.1.

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 Sensing of Environment*, **116**. 62-78. doi: 10.1016/j.rse.2011.02.028.

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.

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

Folland, C. K., D. E. Parker and F. E. Kates, 1984: Worldwide marine temperature fluctuations 1856–1981, *Nature* **310**, 670 - 673, doi: 10.1038/310670a0.

Freeman E., S. D. Woodruff, S. J. Worley, S. J. Lubker, E. C. Kent, W. E. Angel, D. I. Berry, P. Brohan, R. Eastman, L. Gates, W. Gloeden, Z. Ji, J. Lawrimore, N. A. Rayner, G. Rosenhagen and S. R. Smith, 2016. ICOADS Release 3.0: A Major Update to the Historical Marine Climate Record, *Int. J. Climatol.*, doi: 10.1002/joc.4775.

GCOS. 2010. Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC. GCOS-138, World Meteorological Organization: Geneva (online under: http://www.wmo.int/pages/prog/gcos/Publications/gcos-138.pdf, updated in 2010).

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

628 Hartmann, D. L., and Coauthors, 2013: Observations: atmosphere and surface. Climate
629 Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth
630 Assessment Report of the Intergovernmental Panel on Climate Change, T. F. Stocker et
631 al., Eds. Cambridge University Press.

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

634 Houghton, J. T., G. J. Jenkins and J. J. Ephraums (eds.). 1990. Report prepared for
635 Intergovernmental Panel on Climate Change by Working Group I, Cambridge University
636 Press, Cambridge, Great Britain, New York, NY, USA and Melbourne, Australia, 410 pp.

637 Huang, B., V. F. Banzon, E. Freeman, J. Lawrimore, W. Liu, T.C. Peterson, T.M. Smith,
638 P.W. Thorne, S.D. Woodruff, and H.-M. Zhang, 2015: Extended Reconstructed Sea
639 Surface Temperature Version 4 (ERSST.v4). Part I: Upgrades and Intercomparisons. *J.*
640 *Climate*, **28**, 911–930, doi: 10.1175/JCLI-D-14-00006.1.

641 James, R. W. and P. T. Fox, 1972: Comparative sea surface temperature measurements.
642 World Meteorological Organization Reports on Marine Science Affairs, Rep. 5, WMO
643 336, 27pp.

644 JCOMM, 2015: Proceedings of the Fourth JCOMM Workshop on Advances in Marine
645 Climatology (CLIMAR-4) and of the First ICOADS Value-Added Database (IVAD-1)
646 Workshop. JCOMM-TR-079, 30 pp.
647 http://www.jcomm.info/index.php?option=com_oe&task=viewDocumentRecord&docID=
648 15293.

649 Jones, P. D., 2016: The Reliability of Global and Hemispheric Surface Temperature Records,
650 *Adv. Atmos. Sci.*, **33**, 269-282, doi: 10.1007/s00376-015-5194-4.

651 Jones, P. D., T. M. L. Wigley and P. B. Wright, 1986: Global temperature variations between
652 1861 and 1984, *Nature*, 332, 430-434, doi: 10.1038/322430a0.

653 Kawai, Y. and A. Wada, 2007: Diurnal sea surface temperature variation and its impact on
654 the atmosphere and ocean: A review. *J. Oceanography*, **63**(5), 721-744, doi:
655 10.1007/s10872-007-0063-0.

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

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

662 Kennedy, J. J., N.A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011b: Reassessing
663 biases and other uncertainties in sea surface temperature observations measured *in situ*
664 since 1850: 2. Biases and homogenization, *J. Geophys. Res*., **116**, D14104, doi:
665 10.1029/2010JD015220.

666 Kennedy, J. J., R. Smith, and N. Rayner, 2011c: Using AATSR data to assess the quality of
667 in situ sea surface temperature observations for climate studies, *Remote Sens. Environ.*,
668 **116**, 79–92, doi:10.1016/j.rse.2010.11.021.

669 Kent, E. C., and A. Kaplan, 2006: Toward Estimating Climatic Trends in SST, Part 3:
670 Systematic Biases. *J. Atmos. Ocean. Tech.*, **23**(3), 487-500. doi: 10.1175/JTECH1845.1.

671 Kent, E. C., and P. K. Taylor, 2006: Toward Estimating Climatic Trends in SST, Part 1:
672 Methods of Measurement. *J. Atmos. Ocean. Tech.*, **23**(3), 464-475. doi:
673 10.1175/JTECH1843.1.

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

677 Kent, E. C., P. K. Taylor, B. S. Truscott and J. S. Hopkins, 1993: The accuracy of voluntary
678 observing ship's meteorological observations - Results of the VSOP-NA. *J. Atmos. Ocean.*
679 *Tech.*, **10**(4), 591-608, doi: 10.1175/1520-0426(1993)010<0591:TAOVOS>2.0.CO;2.

680 Kent, E. C., N. A. Rayner, D. I. Berry, M. Saunby, B. I. Moat, J. J. Kennedy, and D. E.
681 Parker, 2013: Global analysis of night marine air temperature and its uncertainty since
682 1880: The HadNMAT2 data set. *J. Geophys. Res. Atmos.*, **118**, 1281– 1298, doi:
683 10.1002/jgrd.50152.

684 Kent, E. C., S. D. Woodruff and D. I. Berry, 2007: Metadata from WMO Publication No. 47
685 and an Assessment of Voluntary Observing Ships Observation Heights in ICOADS, *J.*
686 *Atmos. Ocean. Tech.*, **24**(2), 214–234, doi: 10.1175/JTECH1949.1.

687 Lumpkin, R., N. Maximenko and M. Pazos, 2012: Evaluating where and why drifters die, *J.*
688 *Atmos. Ocean. Tech.*, **29**(2), 300–308, doi: 10.1175/JTECH-D-11-00100.1.

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

692 Maury, M. F., 1858: *Explanations and sailing directions to accompany the wind and current*
693 *charts.* Vol. 1. W. A. Harris, Washington DC.

694 Merchant, C. J., O. Embury, N. A. Rayner, D. I. Berry, G. K. Corlett, K. Lean, K. L. Veal, E.
695 C. Kent, D. T. Llewellyn-Jones, J. J. Remedios and R. Saunders, 2012: A twenty-year
696 independent record of sea surface temperature for climate from Along Track Scanning
697 Radiometers, *J. Geophys. Res.*, **117**, C12013, doi: 10.1029/2012JC008400.

698 Merchant, C. J., O. Embury, J. Roberts-Jones, E. Fiedler, C. E. Bulgin, G. K. Corlett, S.
699 Good, A. McLaren, N. A. Rayner, S. Morak-Bozzo and C. Donlon, 2014: Sea surface
700 temperature datasets for climate applications from Phase 1 of the European Space Agency
701 Climate Change Initiative (SST CCI). *Geoscience Data Journal*, **1**: 179–191. doi:
702 10.1002/gdj3.20.

703 Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E.
704 C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and
705 night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, D14,
706 4407, doi: 10.1029/2002JD002670.

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

711 Rennell, J. 1832: An investigation of the currents of the Atlantic Ocean and of those which
712 prevail between the Indian Ocean and the Atlantic, J. G. & F. Rivington for Lady Rodd,
713 London.

714 Reverdin, G. S. Morisset, H. Bellenger, J. Boutin, N. Martin, P. Blouch, J. Rolland, F.
715 Gaillard, P. Bouruet-Aubertot, and B. Ward, 2013: Near–Sea Surface Temperature
716 Stratification from SVP Drifters. *J. Atmos. Oceanic Technol.*, **30**, 1867–1883, doi:
717 10.1175/JTECH-D-12-00182.1

718 Roll, H. U. 1951: Water temperature measurements on deck and in the engine room. *Ann.*
719 *Meteor.*, **4,** 439-443.

720 Roll, H. U. 1951: The accuracy of measuring water temperature with the water scoop
721 thermometer, *Ann. Meteor.*, **4,** 480-482.

722 Smith, T. M., and R. W. Reynolds, 2002: Bias corrections for historic sea surface
723 temperatures based on marine air temperatures. *J. Climate*, **15**, 73-87. doi: 10.1175/1520-
724 0442(2002)015<0073:BCFHSS>2.0.CO;2.

725 Thomas, B. R., E. C. Kent, V. R. Swail and D. I. Berry, 2008: Trends in ship wind speeds
726 adjusted for observation method and height, *Int. J. Climatol.*, **28**(6), 747-763, doi:
727 10.1002/joc.1570.

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

731 Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears, 2005: Uncertainties in Climate
732 Trends: Lessons from Upper-Air Temperature Records. *Bull. Amer. Meteor. Soc.* **86**,
733 1437–1442, doi: 10.1175/BAMS-86-10-1437.

734 Thorne, P. W., R. Allan, L. Ashcroft, P. Brohan, R.J.H Dunn, M.J. Menne, P. Pearce, J.
735 Picas, K.M. Willett, S. Bronnimann, P. Canziani, J. Coll, R. Crouthamel, G. Compo, D.
736 Cuppett, M. Curley, C. Duffy, J. Guijarro, S. Jourdain, E. Kent, H. Kubota, T. Legg, J.
737 Matsumoto, C. Murphy, L. Qingxiang, N. Rayner, E. Rustemeier, L. Slivinski, V.
738 Slonosky, A. Squintu, B. Tinz, M.A. Valente, S. Walsh, X. Wang, N. Westcott, K. Wood,
739 S. Woodruff and S. Worley, 2016: Steps towards an integrated set of surface
740 meteorological holdings to meet the needs of 21st Century Climate Science and
741 applications, *Bull. Amer. Meteor. Soc.* **submitted.**

742 Trewin, B., 2010: Exposure, instrumentation, and observing practice effects on land
743 temperature measurements. *WIREs Clim Change*, 1, 490-506, doi: 10.1002/wcc.46.

744 Venema, V., O. Mestre, E. Aguilar, I. Auer, J.A. Guijarro, P. Domonkos, G. Vertacnik, T.
745 Szentimrey, *et al.*, 2012: Benchmarking homogenization algorithms for monthly data,
746 *Climate of the Past*, **8**, pp. 89-115, doi: 10.5194/cp-8-89-2012.

747 Willett, K. M., P. D. Jones, N. P. Gillett and P. W. Thorne, 2008: Recent changes in surface
748 humidity: Development of the HadCRUH dataset. *J. Climate*, **21**(20), 5364-5383. doi:
749 10.1175/2008JCLI2274.1.

750 Woodruff, S. D., R. J. Slutz, R. L. Jenne, and P. M. Steurer, 1987: A comprehensive ocean-
751 atmosphere data set. *Bull. Amer. Meteor. Soc.*, **68**, 1239-1250. doi: 10.1175/1520-
752 0477(1987)068<1239:ACOADS>2.0.CO;2.

753 Zhang, H. M., R.W. Reynolds, R. Lumpkin, R. Molinari, K. Arzayus, M. Johnson, T.M.
754 Smith, 2009: An integrated global observing system for sea surface temperature using
755 satellites and in situ data: Research to operations, *Bull. Amer. Meteor. Soc.,* 90, 31–38,
756 doi: 10.1175/2008BAMS2577.1.

757

758

Lost datasets – can you help?

Over the years there have been several studies either comparing SST measurements made by different methods or detailed wind-tunnel and ship-based assessments of temperature change from buckets. We have learnt a lot from the papers and reports describing these experiments, but much more could be done if we were able to track down the original measurements. We've tried, and failed, but still hope they are out there and someone knows where they are. And of course if you know the whereabouts of any similar measurements we'd be delighted to hear from you.

James and Fox – 1972: 16k log entries each containing at least 2 measurements of SST and ancillary data and metadata collected under the auspices of the WMO and analyzed at the U.S. Naval Oceanographic Office, Washington D.C.

Roll – 1951a,b: Wind tunnel measurements of the temperature change of a German SST bucket made at the Meteorological Office for NW Germany, Central Office, Hamburg. Also pairs of SST measurements made on the fisheries patrol vessel "Meerkatze" during 1950.

Ashford -1948: Wind tunnel measurements of temperature change of a range of SST buckets carried out in the Instruments Branch of the Meteorological Office, Air Ministry.

Brooks – 1926/1928: Paired measurements of SST made on the "Empress of Britain" and other ships in the 1920s. Analysis was at Clark University, Worcester, MA, and at least a subset of the data was filed with the Library, U. S. Weather Bureau, Washington, D. C.

We are also on the lookout for instructions given to observers, descriptions of how measurements were made, photographs, diagrams and other metadata, so again if you have anything that might be useful, please get in touch.

Figure Captions

Figure 1: Global average SST anomaly from HadSST3, ERSSTv4 and COBE-SST2. In each panel the shaded region is the approximate 95% uncertainty range and the grey areas are the other two data sets and their uncertainty ranges for comparison. Biases and anomalies have been set to average zero over the period 1961-1990.

1a Timeseries of global average SST anomalies from HadSST3 (yellow)

1b As 1a but from ERSST v4 (green)

1c As 1a but for COBE-SST2 (blue)

1d Estimated bias adjustments and their uncertainties from each dataset using the same colour scheme.

Figure 2: Illustrations of factors affecting SST measurements made using different methods.

a) Bucket measurements of SST are affected by ambient conditions (solar radiation, wind speed, temperature, humidity and air-sea temperature difference) that control the thermodynamic forcing. The construction of the bucket is important: different materials will insulate the water sample from the external thermal forcing to varying extents; the volume and water level affect the heat capacity; a lid may reduce heat exchange from the top. Observing protocol may prescribe how long the bucket should remain in the sea, whether the sample is to be stirred, whether the bucket should be shaded from the sun or sheltered from the wind, how it should be stored and how long an exposure time should be allowed for the thermometer to reach equilibrium. And of course important aspects of observing protocol may be either undefined or not followed by an observer;

b) Both engine intake and hull contact sensor measurements of SST are made at depths that may vary with ship loading. The ship may mix the water or draw down surface water and this may vary with ship speed. The temperature of the pumped water at the

784 measurement site will depend on the flow rate and the properties of any sea chest, the
785 distance inboard, the amount of insulation of the pipe and the temperature difference
786 between the water and the ship interior. The type of thermometer and its mounting
787 affects the measurement and bio-fouling may build up with certain types of installation.
788 How the thermometer is read is important. Remote reading permits thermometer
789 installation near the inlet which may not be easily accessible. The thermometers used
790 may have coarse gradations (particularly dial thermometers) and are subject to parallax
791 errors if inconveniently sited. Observations may have been relayed from the engine
792 room to the bridge, possibly incurring delay and communication errors. Hull sensor-
793 derived SST observations may be affected by the thickness and construction of the hull,
794 by the amount of insulation and the temperature contrast between the water temperature
795 and the internal temperature of the ship.;
796 c) Drifting buoys are expected to give the best quality SST observations overall, but there are
797 still several problems that may be encountered, including drift of the calibration over
798 time. Solar radiation on the drifter body may cause errors, either through direct heating
799 or through temperature effects on the electronics: the size of any effect will vary with
800 buoy design. The depth of measurement may vary: the drogue is designed to keep the
801 drifter sphere largely submerged, if lost the measurement will be closer to the surface
802 (Reverdin *et al.* 2013) and the buoy might not remain correctly oriented. Water may be
803 disturbed by motion of the buoy. Bio-fouling can be significant in some regions and has
804 the potential to affect the temperature measurement. Detailed quality control is required
805 to identify pre-deployment activation, beaching and degradation over time, especially at
806 the end of the drifter life.
807

Figure 3: a) Estimates of measurement method composition for ship data only from ICOADS Release 2.5 for the period January 1930 to January 2007 after Kennedy *et al.* (2011b). Darker shading represents measurement method obtained by the SST measurement method indicator in ICOADS (SI) or from a match to an entry via callsign to Pub. 47. Lighter shading represents measurement method obtained indirectly, either through country preference or inferred bucket for the earliest observations.

b) As 3a but also splitting the bucket observations indicating whether the observation was likely to be taken with an uninsulated (canvas) or insulated (rubber or plastic) bucket. The hatched area indicates the estimated uncertainty in that assignment. The white area represents ERI and measurements of unknown source. The dashed lines show the measurement method assignments following (Hirahara *et al.* 2014) partitioning between uninsulated buckets (lower portion), insulated buckets (center portion) and ERI (top portion).

Figure 4: Comparison of SST bias adjustments used in HadSST3 and ERSSTv4 (°C). Grey shaded areas in panels a-g are unsampled.

a) averaged bias adjustment from HadSST3, 1890- 1919;

b) averaged bias adjustment from HadSST3, 1995-2004;

c) as a) but for ERSSTv4; d) as b) but for ERSSTv4;

e) bias adjustment difference (HadSST3 - ERSSTv4), 1890- 1919, hatching indicates 5˚ areas where the difference exceeds half the sum of the full range of the ensemble estimates of bias uncertainty.

f) as e) but for 1995-2004

g) as e) but zonal mean smoothed with a 12-month running mean filter.

h) global mean bias adjustment difference (black) and full range of ensemble differences (grey)

Figure 5:

a) SST anomalies (˚C) relative to 1961-1990 for August 2014 based on ICOADS real time extension based on data for ships, drifting and moored buoys, quality controlled and gridded according to Rayner *et al.* (2006). Grey areas indicate regions with no observations.

b) SST anomalies for August 2014 after interpolation using a local optimal interpolation with varying length scales and successively assimilating buoy and ship measurements.

c) Estimated average biases in gridded engine room measurements assessed using the residual of the interpolation scheme from the previous panel. Details on the method used can be found in the Supplemental Material.

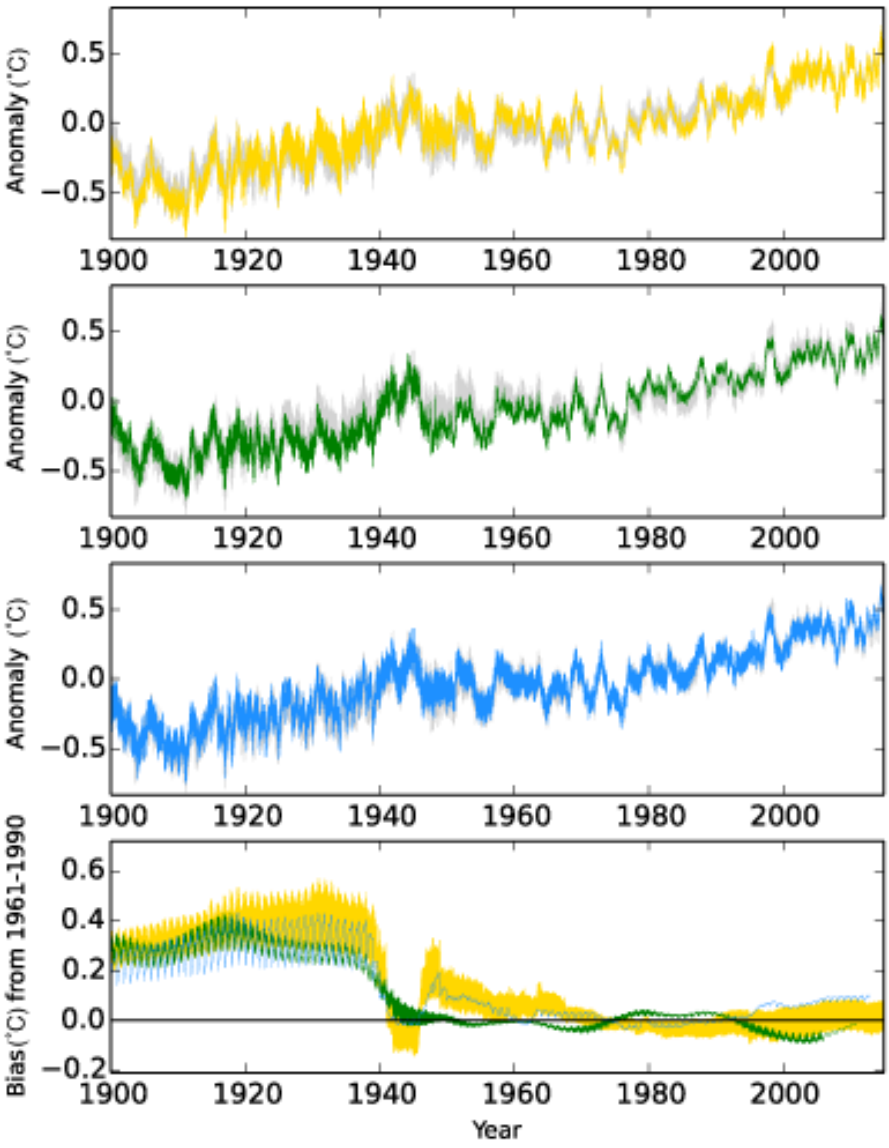


Figure 1: Global average SST anomaly from HadSST3, ERSSTv4 and COBE-SST2. In each panel the shaded region is the approximate 95% uncertainty range and the grey areas are the other two data sets and their uncertainty ranges for comparison. Biases and anomalies have been set to average zero over the period 1961-1990.

1a Timeseries of global average SST anomalies from HadSST3 (yellow)

1b As 1a but from ERSST v4 (green)

1c As 1a but for COBE-SST2 (blue)

1d Estimated bias adjustments and their uncertainties from each dataset using the same colour scheme.

a)

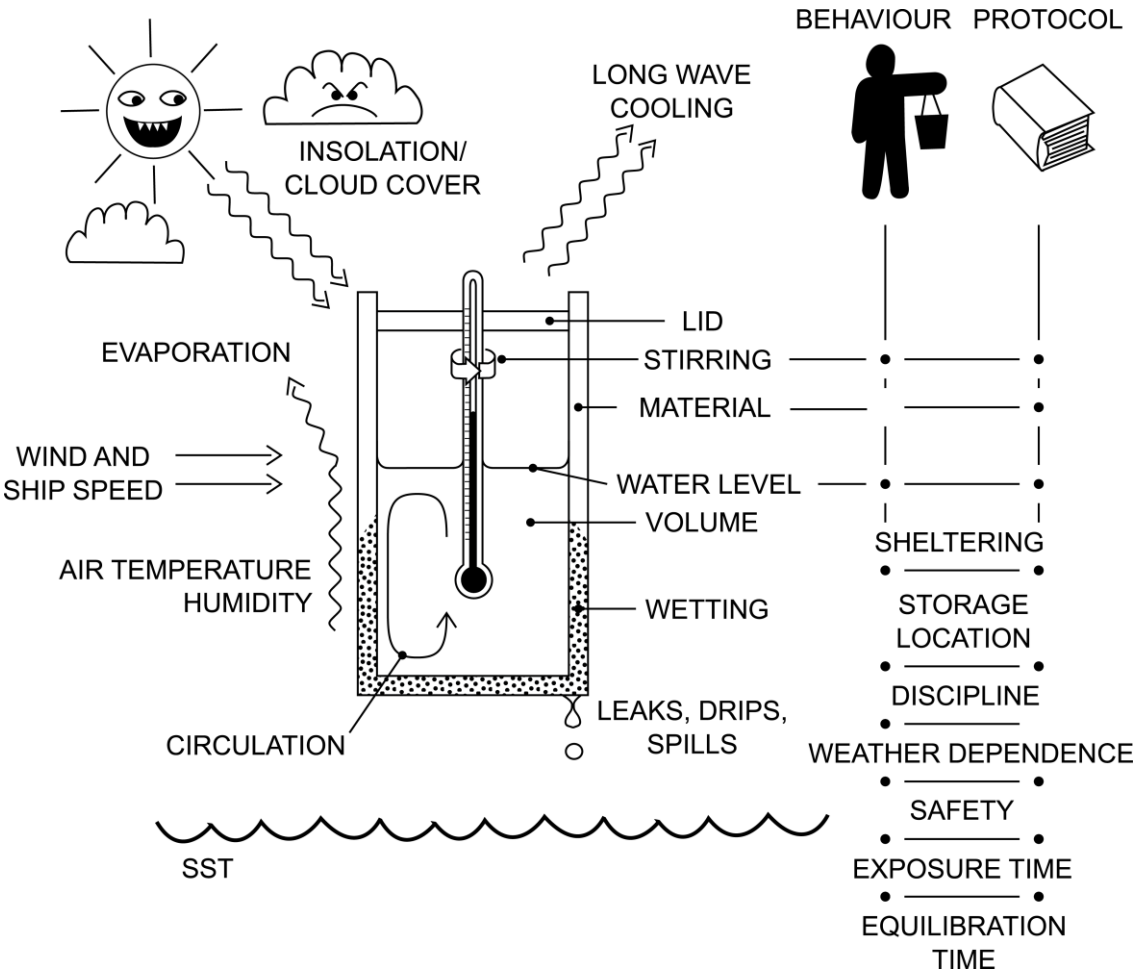
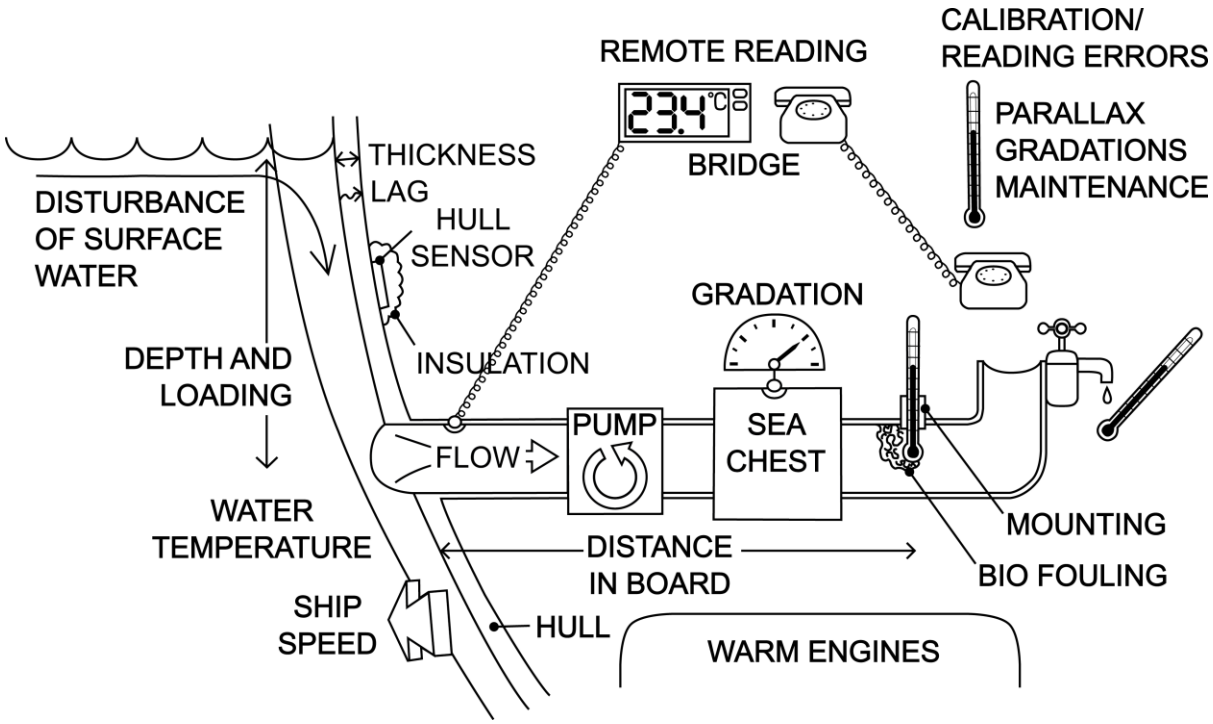
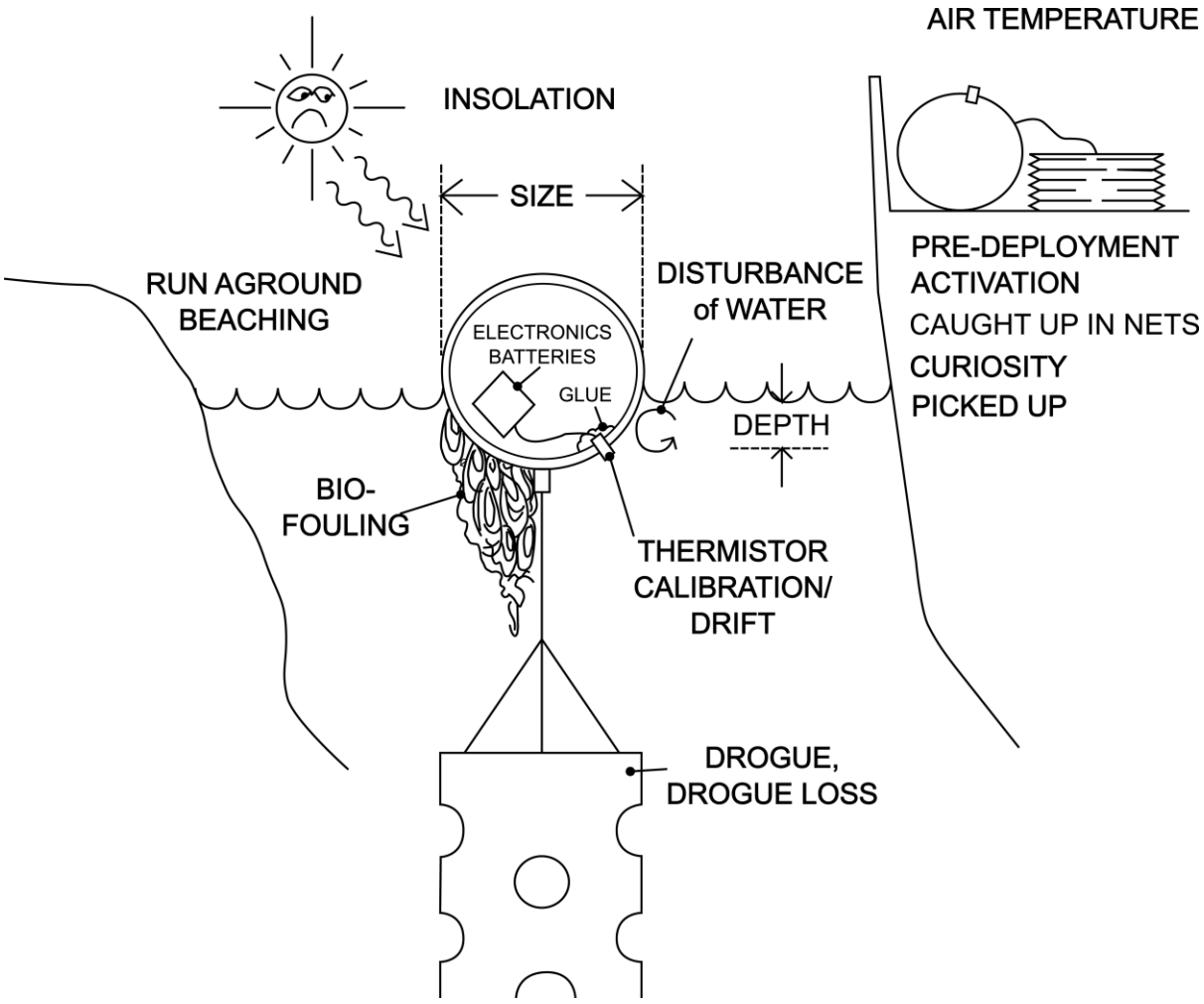


Figure 2: Illustrations of factors affecting SST measurements made using different methods. a) Bucket measurements of SST are affected by ambient conditions (solar radiation, wind speed, temperature, humidity and air-sea temperature difference) that control the thermodynamic forcing. The construction of the bucket is important: different materials will insulate the water sample from the external thermal forcing to varying extents; the volume and water level affect the heat capacity; a lid may reduce heat exchange from the top. Observing protocol may prescribe how long the bucket should remain in the sea, whether the sample is to be stirred, whether the bucket should be shaded from the sun or sheltered from the wind, how it should be stored and how long an exposure time should be allowed for the thermometer to reach equilibrium. And of course important aspects of observing protocol may be either undefined or not followed by an observer;



b) Both engine intake and hull contact sensor measurements of SST are made at depths that may vary with ship loading. The ship may mix the water or draw down surface water and this may vary with ship speed. The temperature of the pumped water at the measurement site will depend on the flow rate and the properties of any sea chest, the distance inboard, the amount of insulation of the pipe and the temperature difference between the water and the ship interior. The type of thermometer and its mounting affects the measurement and bio-fouling may build up with certain types of installation. How the thermometer is read is important. Remote reading permits thermometer installation near the inlet which may not be easily accessible. The thermometers used may have coarse gradations (particularly dial thermometers) and are subject to parallax errors if inconveniently sited. Observations may have been relayed from the engine room to the bridge, possibly incurring delay and communication errors. Hull sensor-derived SST observations may be affected by the thickness and construction of the hull, by the amount of insulation and the temperature contrast between the water temperature and the internal temperature of the ship.;



c) Drifting buoys are expected to give the best quality SST observations overall, but there are still several problems that may be encountered, including drift of the calibration over time. Solar radiation on the drifter body may cause errors, either through direct heating or through temperature effects on the electronics: the size of any effect will vary with buoy design. The depth of measurement may vary: the drogue is designed to keep the drifter sphere largely submerged, if lost the measurement will be closer to the surface (Reverdin *et al.* 2013) and the buoy might not remain correctly oriented. Water may be disturbed by motion of the buoy. Bio-fouling can be significant in some regions and has the potential to affect the temperature measurement. Detailed quality control is required to identify pre-deployment activation, beaching and degradation over time, especially at the end of the drifter life.

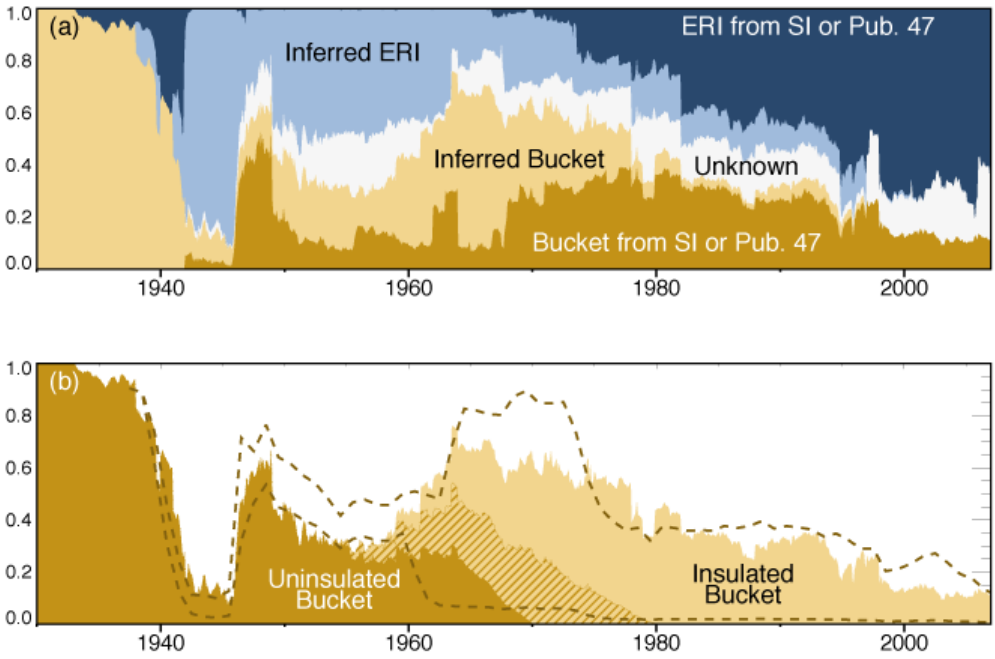
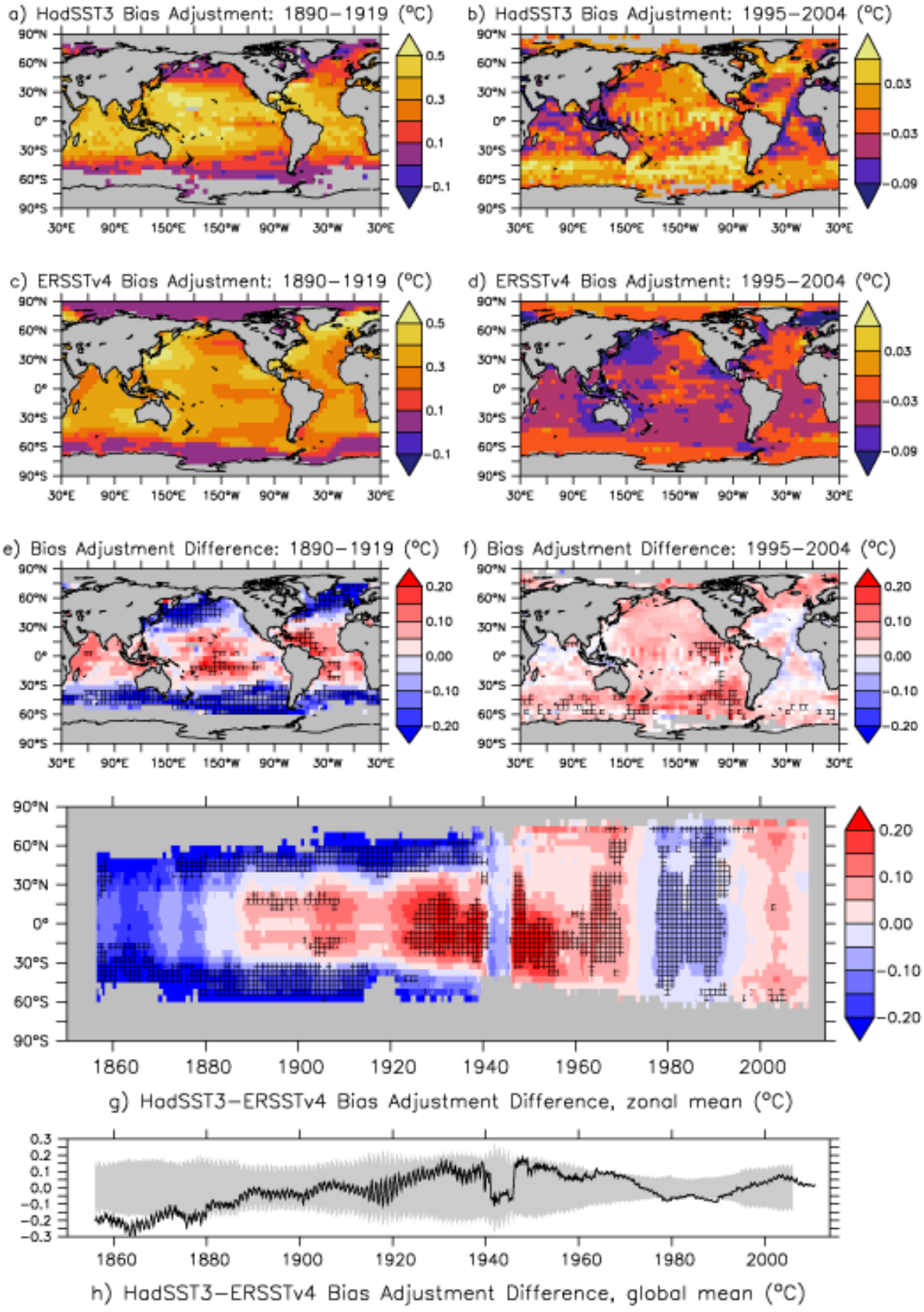


Figure 3: a) Estimates of measurement method composition for ship data only from ICOADS Release 2.5 for the period January 1930 to January 2007 after Kennedy *et al.* (2011b). Darker shading represents measurement method obtained by the SST measurement method indicator in ICOADS (SI) or from a match to an entry via callsign to Pub. 47. Lighter shading represents measurement method obtained indirectly, either through country preference or inferred bucket for the earliest observations.

b) As 3a but also splitting the bucket observations indicating whether the observation was likely to be taken with an uninsulated (canvas) or insulated (rubber or plastic) bucket. The hatched area indicates the estimated uncertainty in that assignment. The white area represents ERI and measurements of unknown source. The dashed lines show the measurement method assignments following (Hirahara *et al.* 2014) partitioning between uninsulated buckets (lower portion), insulated buckets (center portion) and ERI (top portion).



912

913 **Figure 4**

Figure 4: Comparison of SST bias adjustments used in HadSST3 and ERSSTv4 (°C). Grey shaded areas in panels a-g are unsampled.

a) averaged bias adjustment from HadSST3, 1890- 1919;

b) averaged bias adjustment from HadSST3, 1995-2004;

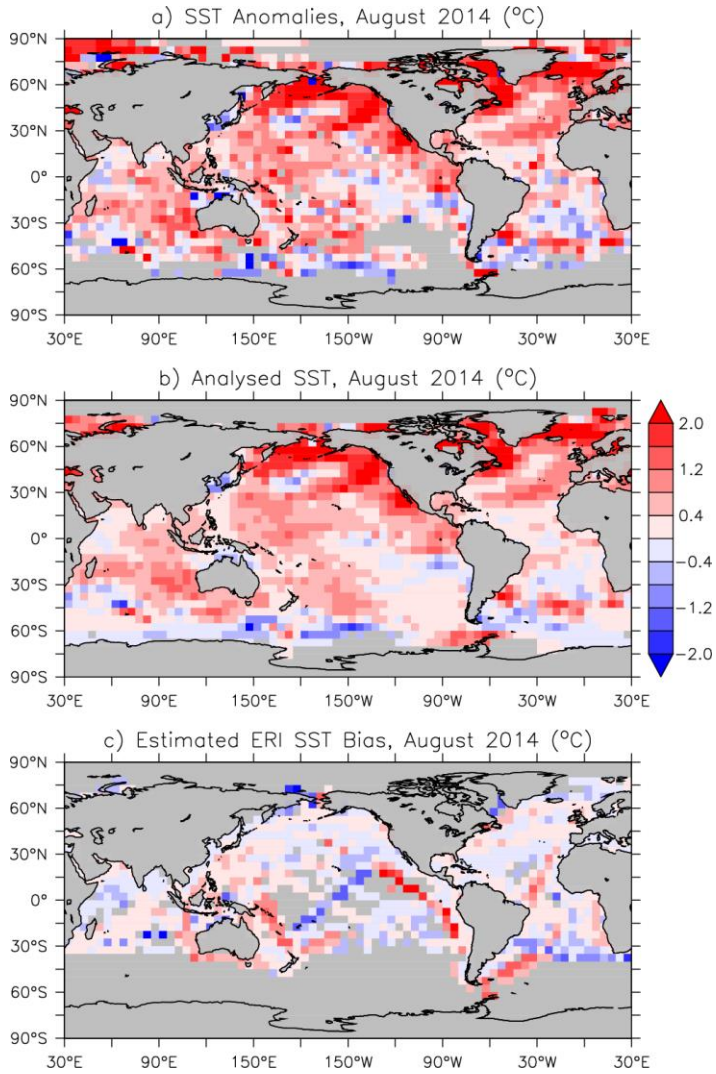
c) as a) but for ERSSTv4; d) as b) but for ERSSTv4;

e) bias adjustment difference (HadSST3 - ERSSTv4), 1890- 1919, hatching indicates 5˚ areas where the difference exceeds half the sum of the full range of the ensemble estimates of bias uncertainty.

f) as e) but for 1995-2004

g) as e) but zonal mean smoothed with a 12-month running mean filter.

h) global mean bias adjustment difference (black) and full range of ensemble differences (grey)



926

927 **Figure 5:**

928 a) SST anomalies (˚C) relative to 1961-1990 for August 2014 based on ICOADS real time
929 extension based on data for ships, drifting and moored buoys, quality controlled and
930 gridded according to Rayner *et al.* (2006). Grey areas indicate regions with no
931 observations.

932 b) SST anomalies for August 2014 after interpolation using a local optimal interpolation with
933 varying length scales and successively assimilating buoy and ship measurements.

934 c) Estimated average biases in gridded engine room measurements assessed using the
935 residual of the interpolation scheme from the previous panel. Details on the method
936 used can be found in the Supplemental Material.