ESSAY

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

Bias estimation for sea surface temperature is discussed and recommendations for improving data, observational metadata, and uncertainty modeling are given.

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 supplement , which is available online at http://dx.doi.org/10.1175/BAMS-D-15-00251.2). Both SST and land air temperature require adjustments to account for changes in, for example, 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).

AFFILIATIONS: Kent, Berry, and Carella—National Oceanography Centre, Southampton, United Kingdom; Kennedy, Parker, Atkinson, and Rayner—Met Office Hadley Centre, Exeter, United Kingdom; Smith—NOAA/NESDIS/STAR, College Park, Maryland; Hirahara—Global Environment and Marine Department, Japan Meteorological Agency, Tokyo, Japan, and ECMWF, Reading, United Kingdom; Huang and Zhang—NOAA/National Centers for Environmental Information, Asheville, North Carolina; Kaplan—Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York; Fukuda—Japan Meteorological Agency, Tokyo, Japan; Ishii—Climate Research Division, Meteorological Research Institute, Tsukuba, Ibaraki, Japan; Jones—Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, United Kingdom, and Department of Meteorology, Center of Excellence for Climate Change Research, King Abdulaziz University, Jeddah, Saudi Arabia; Lindgren—University of Edinburgh, Edinburgh, United Kingdom; Merchant and Morak-Bozzo—University of Reading, Reading, United Kingdom; Venema—University of Bonn, Bonn, Germany; Yasui—Global Environment and Marine Department, Japan Meteorological Agency, Tokyo, Japan
CORRESPONDING AUTHOR: Elizabeth C. Kent, eck@noc.ac.uk

The abstract for this article can be found in this issue, following the table of contents.

DOI:10.1175/BAMS-D-15-00251.1

A supplement to this article is available online (10.1175/BAMS-D-15-00251.2)

In final form 16 December 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: Hadley Centre SST dataset, version 3 (HadSST3; Kennedy et al. 2011a,b); Extended Reconstructed SST, version 4 (ERSSTv4; Huang et al. 2015); and Centennial Observation-Based Estimates of SST, version 2 (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

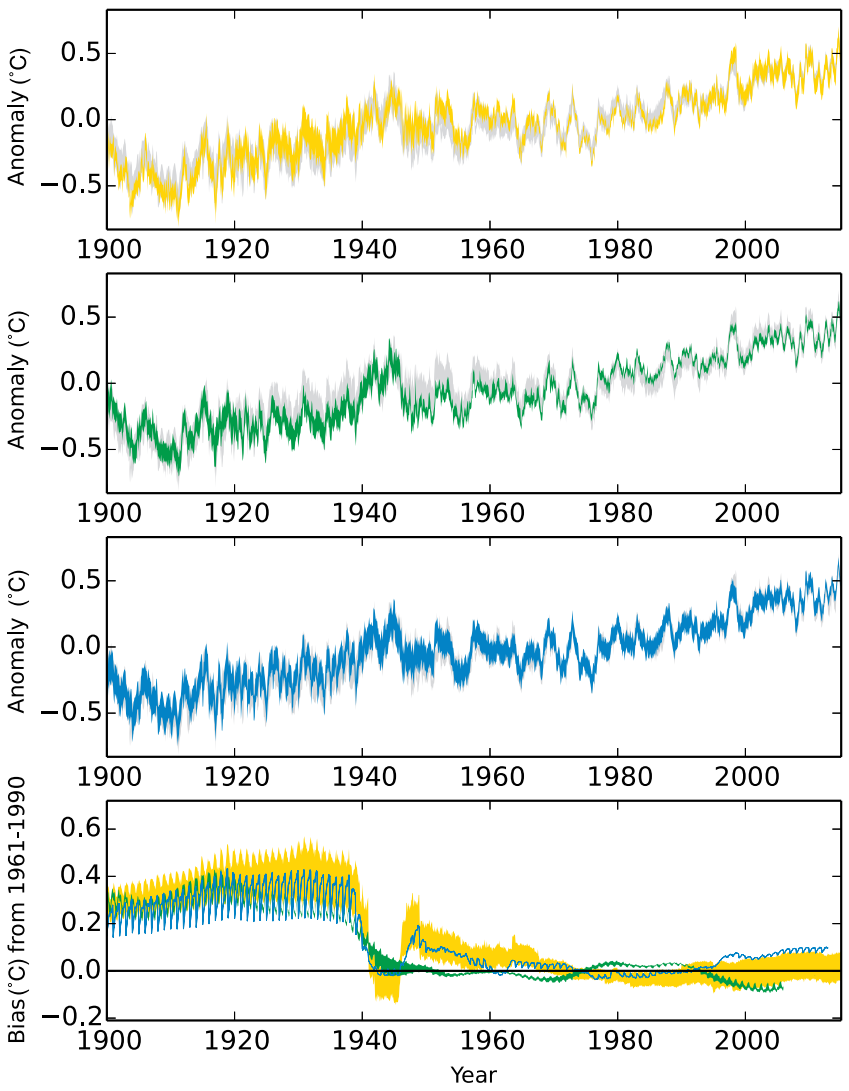


Fig. 1. Global-average SST anomaly from HadSST3, ERSSTv4, and COBE-SST2. In each panel the shaded region is the approximate 95% uncertainty range, and the gray areas are the other two datasets and their uncertainty ranges for comparison. Biases and anomalies have been set to average zero over the period 1961–90. (top) Time series of global-average SST anomalies from HadSST3 (yellow). (top middle) As in (top), but from ERSST v4 (green). (bottom middle) As in (top), but for COBE-SST2 (blue). (bottom) Estimated bias adjustments and their uncertainties from each dataset using the same color scheme.

LOST DATASETS—CAN YOU HELP?

Over the years there have been several studies comparing either SST measurements made by different methods or detailed wind tunnel– and ship-based assessments of temperature change from buckets. We have learned 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 have tried and failed, but we still hope they are out there and that someone knows where they are. And, of course, if you know the whereabouts of any similar measurements, we would be delighted to hear from you.

James and Fox (1972): Approximately 16,000 log entries, each containing at least two measurements of SST and ancillary data, and metadata collected under the auspices of the World Meteorological Organization (WMO) and analyzed at the U.S. Naval Oceanographic Office in Washington, D.C.

Roll (1951a,b): Wind-tunnel measurements of the temperature change of a German SST bucket made at the Meteorological Office for northwestern Germany, Central Office, Hamburg. Also pairs of SST measurements made on the Fishery 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, United Kingdom.

Brooks (1926, 1928): Paired measurements of SST made on the Royal Mail Ship *Empress of Britain* and other ships in the 1920s. Analysis was at Clark University, Worcester, Massachusetts, and at least a subset of the data was filed with the Library of the U.S. Weather Bureau in 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.

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 (see the sidebar for more information about lost datasets). 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 150+ years are known. We know that adjustments are required and the direction and approximate size of the change at very large scales. However, a 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 better understand the biases, to improve adjustment methods, and to refine the uncertainty estimates.

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

WHAT IS SST AND HOW IS IT MEASURED?

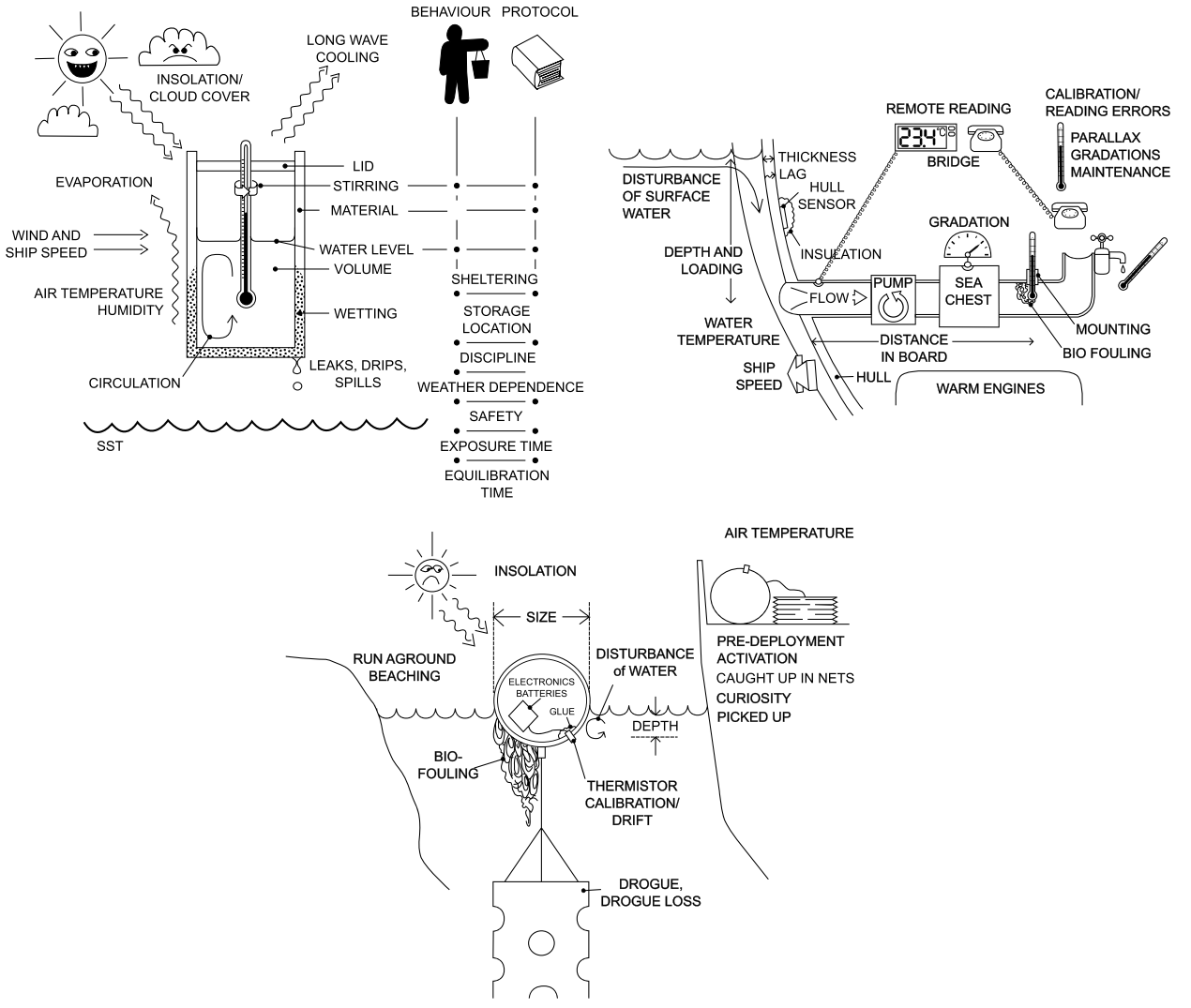
What is SST? The temperature of the water near the sea surface varies on all space and time scales. The term SST has typically been used to describe the mean temperature of the upper few meters of the ocean. Historically measurements taken at depths from the surface and down to about 20 m have all been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

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 2 (left) 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 it 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 the 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 to 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 on board to cool the engines can be used as an estimate of SST (Fig. 2, right). 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 m inboard from the inlet. Technological developments have likely resulted in thermometers placed nearer to the hull (possible with remote-reading automatic sensors) and farther 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 thermometers. 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 Fig. 2, right) 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 (Fig. 2, bottom) measure at shallow depths, nominally 10–20 cm. Biases in drifter measurements might arise due to error in sensor calibration, temperature calibration "drift" while deployed, or biofouling 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. 2012), 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 predeployment 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

◂ Fig. 2. Illustrations of factors affecting SST measurements made using different methods. (top left) 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 of 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. (top right) 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 affect the measurement, and biofouling 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 by the temperature contrast between the water temperature and the internal temperature of the ship. (bottom) 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 the drifter sphere is 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 the motion of the buoy. Biofouling can be significant in some regions and has the potential to affect the temperature measurement. Detailed quality control is required to identify predeployment activation, beaching, and degradation over time, especially at the end of the drifter life.

control flags appended, but checks of the International Comprehensive Ocean–Atmosphere Data Set (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 1 m 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.

Availability of observations and ancillary information. SST observations were first made available in the nineteenth 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 U.S. national collection developed into a publicly available databank (Woodruff et al. 1987) that became ICOADS, currently release 3.0 (Freeman et al. 2017). 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 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, and complete information of methods, instruments, and protocols used, and of the ambient conditions (Fig. 2). ICOADS contains some of the information required (described in section S3 of the supplement), 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).

CURRENT APPROACHES TO SST BIAS ESTIMATION.

Physics-based bias models. The factors affecting bucket SST measurements are well known (Fig. 2, top) 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 supplement. 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 (Fig. 2, right). 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 the daytime 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 subsurface measurements (Embury et al. 2012). Such models could be used to inform comparisons of in situ measurements made at different depths.

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 supplement 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 (Publ. 47); Kent et al. 2007], and other documentary information, to estimate measurement methods and their uncertainties (Fig. 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 an 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 (Fig. 3b). Notable discrepancies include estimates of the rate of transition from uninsulated to insulated buckets (Kennedy 2014).

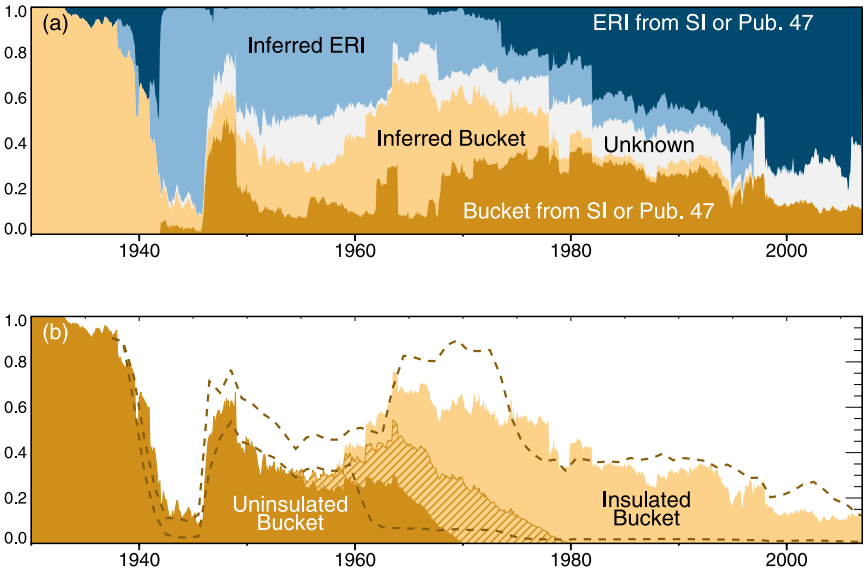


Fig. 3. (a) Estimates of measurement method composition for ship data only from ICOADS release 2.5 for the period Jan 1930–Jan 2007 after Kennedy et al. (2011b). Darker shading represents the measurement method obtained by the SST measurement method indicator in ICOADS (SI) or from a match to an entry via call sign to Publ. 47. Lighter shading represents the measurement method obtained indirectly, either through country preference or inferred bucket for the earliest observations. (b) As in (a), 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 (bottom portion), insulated buckets (middle portion), and ERI (top portion).

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.

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; section S1 of the supplement) 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 section S6 of the supplement.

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 differences 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 II (WWII) period dominated by bucket measurements (Fig. 3). Huang et al. (2015) extended the method throughout the record

and generated an ensemble to explore uncertainty (described in section S6 of the supplement).

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

COMPARISON AND EVALUATION OF ESTIMATES OF SST BIAS. *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 calculation of the ensembles is described in sections S4 and S6 of the supplement). 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 do not yet fully understand the biases and their uncertainties at all times throughout the record. Maps showing the average spatial variation of the biases averaged over 1890–1919 (Figs. 4a,c) show differences that exceed the range of their combined uncertainty ensembles over large regions (Fig. 4e). Even in the more recent period of 1995–2004 (Figs. 4b,d), there are regions where the difference exceeds the ensemble range (Fig. 4f). Zonal-mean (Fig. 4g) and global-average differences (Fig. 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 is when the proportion of SST observations made by ERI is increasing (Fig. 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 WWII 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 WWII 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 WWII period shows rapid variations in the difference between the bias estimates (Figs. 4g,h) but also a large ensemble range, so by this metric these differences are understood, albeit very uncertainly. 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 WWII), 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 (Figs. 4e–g; also Huang et al. 2015; section S7 of the supplement). Uncertainty due to undersampling 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) datasets 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.

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—since the mid-1990s—there are multiple sources of validation data for

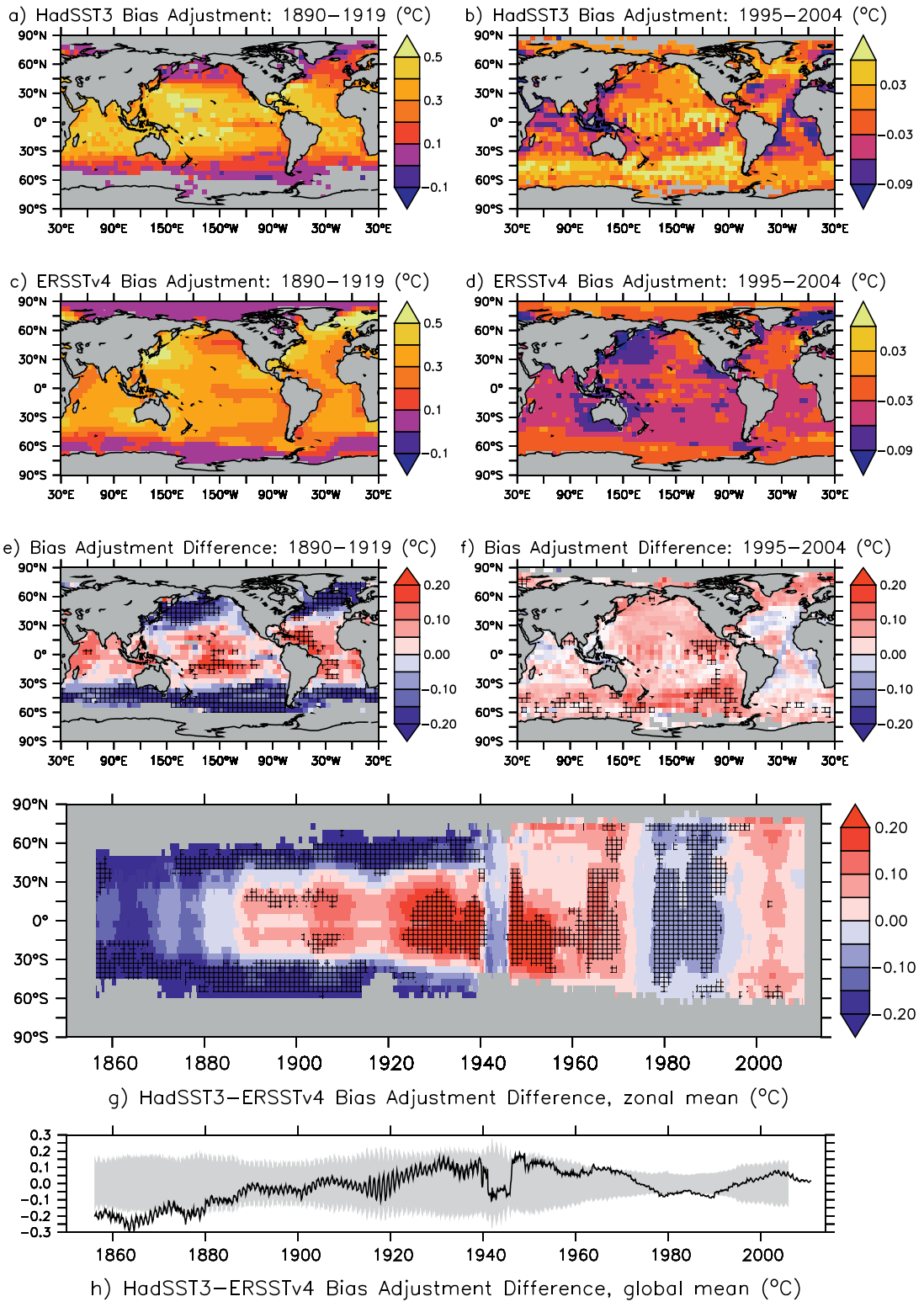


FIG. 4. Comparison of SST bias adjustments used in HadSST3 and ERSSTv4 (°C). (a) Averaged bias adjustment from HadSST3, 1890–1919. (b) Averaged bias adjustment from HadSST3, 1995–2004. (c) As in (a), but for ERSSTv4. (d) As in (b), but for ERSSTv4. (e) Bias adjustment difference (HadSST3 minus ERSSTv4), 1890–1919; hatching indicates 5°C areas where the difference exceeds half the sum of the full range of the ensemble estimates of bias uncertainty. (f) As in (e), but for 1995–2004. (g) As in (e), but the zonal mean is smoothed with a 12-month running-mean filter. Gray shaded areas in (a)–(g) are unsampled. (h) Global-mean bias adjustment difference (black) and full range of ensemble differences (gray).

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 (Argo 2000) provide accurate data, but low sampling rates, and can be used for validation after about 2005. Some satellite datasets covering the 1990s to the present are of the desired accuracy and are largely independent of the in situ record (Merchant et al. 2012, 2014); therefore, they are 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 they 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 datasets would be valuable. Careful analysis is required if comparisons are made with different parameters (such as air temperature), with coastal observations (which 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 SST biases adjusted in different ways (e.g., Folland 2005). An overview of potential validation data is given in section S8 of the supplement. 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.

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, that 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. Kennedy et al. (2011b) showed that adjustments applied to ERI and bucket measurements improved agreement between these two subsets of data from the 1950s onward.

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 also can 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 dataset 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, which are not yet widely used in historical SST datasets, might be used to validate modern periods. Newly digitized data could be used for historical assessments. A degree of independence should also be maintained between the institutions producing bias adjustments and those performing validation. This could be achieved if validation were carried out by an organization independent of the dataset developers, or by using a standard set of widely agreed criteria and comparisons.

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

PRIORITIES FOR THE FUTURE. *Improvements to data and metadata.* Fundamentally, there is scope for improvements to ICOADS. Although ICOADS is often thought of as "raw" data, it is derived from a larger, more heterogeneous underlying databank from diverse sources. Further reprocessing of the databank could help to better resolve duplicate observations, incomplete ship identifiers, scale conversions, missing metadata, and positional errors among other basic problems (recommendation 2). The recent

addition (release 2.5.1 and later) of unique identification (UID) to each report in ICOADS is tremendously helpful. Tying quality control information and metadata studies back to ICOADS via the UID and sharing code and methods will improve traceability, promote collaboration, and help new researchers enter the field (recommendation 9).

Much is to be gained from improvements to metadata (recommendations 1–3). Ship tracking—the association of individual reports into coherent voyages (Carella et al. 2017)—will enable the better characterization of ship-by-ship biases and other errors. Bringing together known sources of metadata into a single repository would be a step toward a more holistic synthesis. A start has been made on inferring absent metadata (Kent et al. 2007, 2010; Kennedy et al. 2011b; Hirahara et al. 2014; Carella et al. 2017) 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 Publ. 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, for example, 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 multivariate records are often fragmented (Thorne et al. 2017). 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).

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 and their uncertainties, and allow for 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 spatiotemporal 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 our 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 (Fig. 2, right), so these ship-specific biases may need to be characterized statistically.

Improved statistical approaches. SST biases are statistically and computationally challenging. There are several hundred million in situ observations in ICOADS. This amount of data 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).

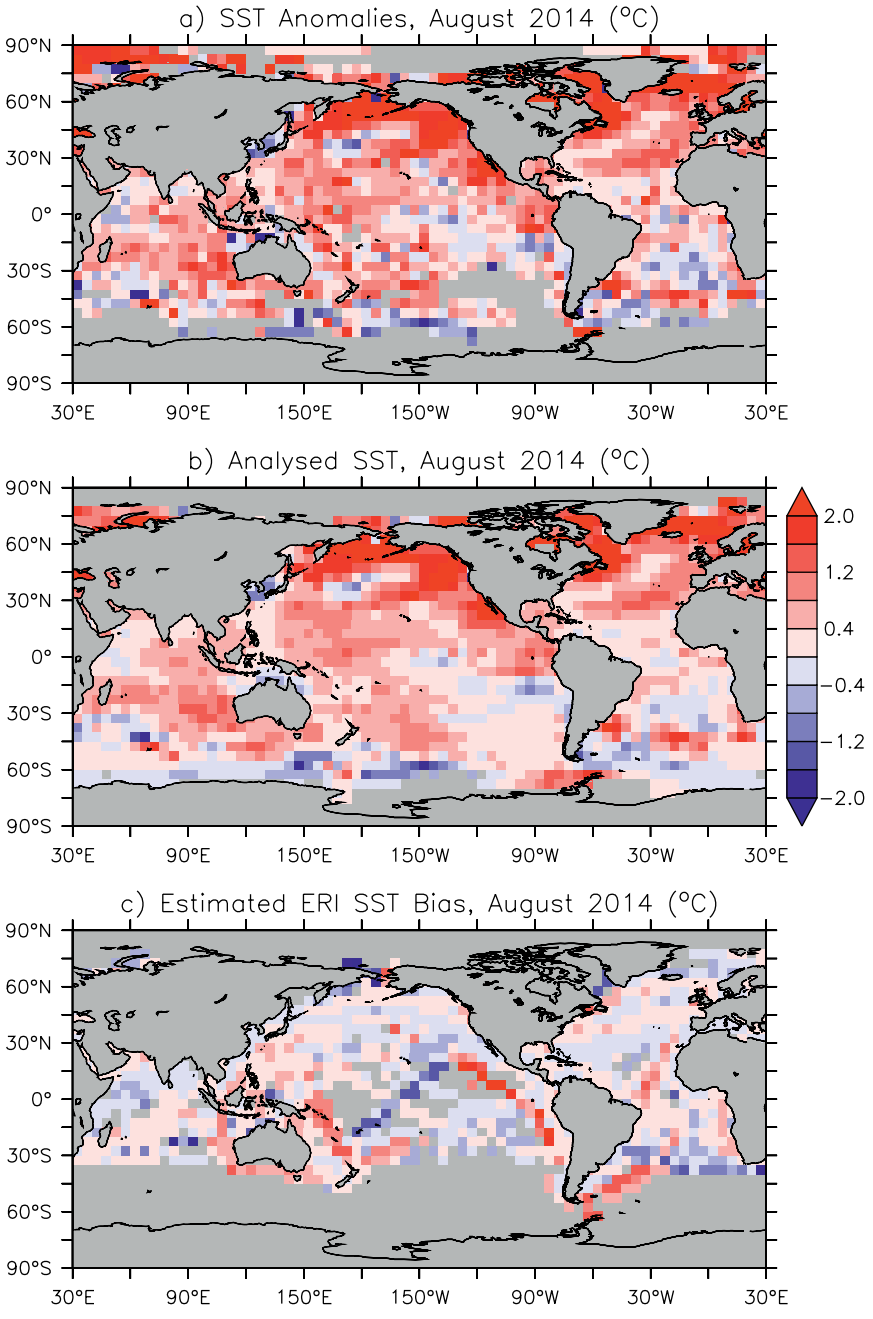


Fig. 5. (a) SST anomalies (°C) relative to 1961–90 for Aug 2014 based on an ICOADS real-time extension based on data for ships, drifting and moored buoys, quality controlled and gridded according to Rayner et al. (2006). Gray areas indicate regions with no observations. (b) SST anomalies for Aug 2014 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 (c). Details on the method used can be found in the supplement.

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 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 (Fig. 5). This relatively simple implementation, described in more detail in section S9 of the supplement, 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 learned 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 also could 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.

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/Hadley Centre Sea Ice and Sea Surface Temperature dataset (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 are 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 multivariate observations together.

Recommendation 2: Reprocess existing ICOADS records. Older ICOADS acquisitions are often lacking metadata and are 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 call-sign recovery and reprocessing of WMO Publ. 47 metadata will help link observations to metadata from 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 they 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 our 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 the 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 been tackled by only 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, among others. Providing modular software tools and 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.

ACKNOWLEDGMENTS. We thank the three reviewers for their help in 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

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.

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. 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. Oper. Oceanogr.*, **5**, 59–73, 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**, 987–1001, doi:10.1002/joc.2059.

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

Bitterman, D. S., and D. V. Hansen, 1993: Evaluation of sea surface temperature measurements from drifting buoys. *J. Atmos. Oceanic Technol.*, **10**, 88–96, doi:10.1175/1520-0426(1993)010<0088:EOSSTM>2.0.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.

Brooks, C. F., 1926: Observing water-surface temperatures at sea. *Mon. Wea. Rev.*, **54**, 241–253, doi:10.1175/1520-0493(1926)54<241:OWTAS>2.0.CO;2.

——, 1928: Reliability of different methods of taking sea surface temperature measurements. *J. Wash. Acad. Sci.*, **18**, 525–545.

Bunker, A. F., 1976: Computations of surface energy flux and annual air–sea interaction cycles of the North Atlantic Ocean. *Mon. Wea. Rev.*, **104**, 1122–1140, doi:10.1175/1520-0493(1976)104<1122:COSEFA>2.0.CO;2.

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.

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**, 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 Sens. Environ.*, **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.

——, 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.

——, ——, and F. E. Kates, 1984: Worldwide marine temperature fluctuations 1856–1981. *Nature*, **310**, 670–673, doi:10.1038/310670a0.

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.]

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.

Hartmann, D. L., and Coauthors, 2013: Observations: Atmosphere and surface. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 159–254.

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.

Houghton, J. T., G. J. Jenkins, and J. J. Ephraums, Eds., 1990: *Climate Change: The IPCC Scientific Assessment*. Cambridge University Press, 365 pp.

Huang, B., 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.

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

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

Jones, P. D., 2016: The reliability of global and hemispheric surface temperature records. *Adv. Atmos. Sci.*, **33**, 269–282, doi:10.1007/s00376-015-5194-4.

——, T. M. L. Wigley, and P. B. Wright, 1986: Global temperature variations between 1861 and 1984. *Nature*, **322**, 430–434, doi:10.1038/322430a0.

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 P. K. Taylor, 2006: Toward estimating climatic trends in SST. Part I: Methods of measurement. *J. Atmos. Oceanic Technol.*, **23**, 464–475, doi:10.1175/JTECH1843.1.

——, 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.

——, ——, B. S. Truscott, and J. S. Hopkins, 1993: The accuracy of voluntary observing ship's meteorological observations—Results of the VSOP-NA. *J. Atmos. Oceanic Technol.*, **10**, 591–608, doi:10.1175/1520-0426(1993)010<0591:TAOVOS>2.0.CO;2.

——, S. D. Woodruff, and D. I. Berry, 2007: Metadata from WMO Publication No. 47 and an assessment of Voluntary Observing Ship 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.

Lumpkin, R., N. Maximenko, and M. Pazos, 2012: Evaluating where and why drifters die. *J. Atmos.*

Oceanic Technol., **29**, 300–308, doi:10.1175/JTECH-D-11-00100.1.

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.]

Merchant, C. J., 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.

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. Ansell, and S. B. F. 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.

Rennell, J., 1832: *An Investigation of the Currents of the Atlantic Ocean and of Those which Prevail between the Indian Ocean and the Atlantic*. J. G. & F. Rivington, 375 pp.

Reverdin, G., and Coauthors, 2013: Near–sea surface temperature stratification from SVP drifters. *J. Atmos. Oceanic Technol.*, **30**, 1867–1883, doi:10.1175/JTECH-D-12-00182.1.

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

——, 1951b: The accuracy of measuring water temperature with the water scoop thermometer. *Ann. Meteor.*, **4**, 480–482.

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.

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

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., D. E. Parker, J. R. Christy, and C. A. Mears, 2005: Uncertainties in climate trends: Lessons from upper-air temperature records. *Bull. Amer. Meteor. Soc.*, **86**, 1437–1442, doi:10.1175/BAMS-86-10-1437.

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

Trewin, B., 2010: Exposure, instrumentation, and observing practice effects on land temperature measurements. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 490–506, doi:10.1002/wcc.46.

Venema, V., and Coauthors, 2012: Benchmarking homogenization algorithms for monthly data. *Climate Past*, **8**, 89–115, doi:10.5194/cp-8-89-2012.

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

Woodruff, S. D., R. J. Slutz, R. L. Jenne, and P. M. Steurer, 1987: A Comprehensive Ocean-Atmosphere Data Set. *Bull. Amer. Meteor. Soc.*, **68**, 1239–1250, doi:10.1175/1520-0477(1987)068<1239:ACOADS>2.0.CO;2.

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