AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER
10.1002/2017GL076475

Key Points:
- SST measurement method is identified from characteristic differences in diurnal variations under similar wind and solar radiation conditions
- Mean SST anomaly differences between the different measurement methods vary on scales from global to regional and seasonal to decadal
- Method-dependent bias adjustments used in global SST gridded analyses do not fully capture the observed differences between the methods

Supporting Information:
- Supporting Information S1

Correspondence to:
E. C. Kent,
eck@noc.ac.uk

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

Received 6 NOV 2017
Accepted 15 DEC 2017
Accepted article online 26 DEC 2017

Estimating Sea Surface Temperature Measurement Methods Using Characteristic Differences in the Diurnal Cycle

G. Carella[1,2], J. J. Kennedy[3], D. I. Berry[1], S. Hirahara[4], C. J. Merchant[5,6], S. Morak-Bozzo[5], and E. C. Kent[1]

[1]National Oceanography Centre, Southampton, UK, [2]Laboratoire des Sciences du Climat et de l'Environnement (LSCE-IPSL, CNRS/CEA/UVSQ), Centre d'Etudes de Saclay, Orme des Merisiers, Gif-sur-Yvette, France, [3]Met Office Hadley Centre, Exeter, UK, [4]Japan Meteorological Agency, Tokyo, Japan, [5]Department of Meteorology, University of Reading, Reading, UK, [6]National Centre for Earth Observation, University of Reading, Reading, UK

Abstract Lack of reliable observational metadata represents a key barrier to understanding sea surface temperature (SST) measurement biases, a large contributor to uncertainty in the global surface record. We present a method to identify SST measurement practice by comparing the observed SST diurnal cycle from individual ships with a reference from drifting buoys under similar conditions of wind and solar radiation. Compared to existing estimates, we found a larger number of engine room-intake (ERI) reports post–World War II and in the period 1960–1980. Differences in the inferred mixture of observations lead to a systematic warmer shift of the bias adjusted SST anomalies from 1980 compared to previous estimates, while reducing the ensemble spread. Changes in mean field differences between bucket and ERI SST anomalies in the Northern Hemisphere over the period 1955–1995 could be as large as 0.5°C and are not well reproduced by current bias adjustment models.

Plain Language Summary The sea surface temperature (SST) is an important indicator of climate change but its uncertainty affects our confidence in estimates of global surface temperature change. A main systematic component of SST uncertainty (or bias) is caused by changes in SST observational practice onboard ships. Historically, SST measurements have been made either using buckets to collect water samples or recording the temperature of pumped seawater used to cool the ship engines (engine room-intake (ERI)). Because SST observational biases vary by measurement method, empirical models to quantify SST biases must be applied separately to bucket and ERI SSTs. This approach is hampered by the lack of reliable information on the adopted measurement method. We present an independent assessment of SST measurement practice derived comparing diurnal SST variations from individual ships with those computed from drifting buoys under similar conditions. The newly inferred mixture of observations leads to a systematic warmer shift of the bias-adjusted SST anomalies from 1980 compared to previous estimates, while reducing the uncertainty in these estimates. Changes in bucket-ERI mean SST anomaly differences are not fully reproduced by current bias adjustment models. These results have important implications for estimates of uncertainty in SST trends and variability changes.

1. Introduction

Sea surface temperature (SST) is a crucial parameter for climate change assessments (Hartmann et al., 2013) but its uncertainty affects our confidence in estimates of surface-temperature change. Biases in SST data are the largest contributor to uncertainty in large-scale global surface temperature time series (Jones, 2016). The systematic component of SST uncertainty comes from changes in the SST observational practice and affects the assessment of long-term trends and variability. SST observational biases for observations made by ships vary by measurement method (Kent et al., 2017). Before World War II (WWII) most SST measurements were made using buckets to collect water samples: during collection and hauling the temperature of the water in the bucket is affected by heat exchange with the atmosphere and by any direct solar heating. Typical atmospheric conditions over the ocean combine to produce a global mean cool bias relative to the true SST, ranging from about −0.1°C in 1850s to −0.4°C in 1940s (Folland & Parker, 1995; Smith & Reynolds, 2002). This increase in bias is due to changing bucket types, from wooden to canvas, and increasing ship speed (e.g., Kennedy et al., 2011b, hereafter K11b) and hence increasing airflow past the buckets. The

practice of recording the SST as the temperature of pumped seawater used to cool the engines (engine room-intakes or ERIs) became common after ~1930. While ERI measurements sample water at greater depth than buckets (Kent, Woodruff, & Berry, 2007), and therefore are expected to be on average colder than surface SSTs, they are generally biased warm (Kent & Kaplan, 2006; K11b), likely due to heating of water in the intake pipes. Existing studies (summarized in K11b) estimate typical mean ERI biases between +0.1°C and +0.3°C, although there is some evidence that ERI biases have reduced over time (Kent & Kaplan, 2006). From around 1970 ships increasingly use dedicated hull-mounted sensors that are probably of higher quality than ERIs (Kent et al., 1993).

Commonly used SST bias adjustments are based either on empirical models of observational practice (Hadley Centre SST data set (HadSST3) (Folland & Parker, 1995; Kennedy et al., 2011a; K11b) and Centennial Observation-Based Estimates of SST version 2 (COBE-SST2) (Hirahara et al., 2014)), or on the assumption of an invariant relationship between SST and night marine air temperature (Extended Reconstructed Sea Surface Temperature, Version 5) (Huang et al., 2017; Smith & Reynolds, 2002). Lack of reliable observational metadata represents a key barrier to understanding SST biases and their uncertainties (Kennedy, 2014; Kent et al., 2017). For example, Thompson et al. (2008) reported a sudden drop in the residual global-average SST in late 1945 after known climate signals were removed, which coincided with a change in data source. The drop was hypothesized to arise from a rapid change from ERI measurements to uninsulated bucket measurements at the end of WWII. The details of this transition are uncertain, as metadata in this period are sparse.

Two flags indicating SST measurement method are available in the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) (Freeman et al., 2017; Woodruff et al., 2011). The first, the SST method indicator (SI), contains information recorded in the ships' logbooks or contained in the weather report. The second, the SST measurement method (SIM) is derived from WMO Publication 47—hereafter Pub47 (e.g., Kent et al., 2007)—and is available from ~1960 onward.

Unfortunately, the available metadata is not always correct; both SI and SIM can lead to misidentified reports (Kent & Taylor, 2006; K11b). For discussion on the reliability of the metadata see section S1 of the supporting information.

Both HadSST3 and COBE-SST2 bias adjustments require estimation of the fraction of observations associated with each measurement method, inferred using either SI or SIM. In K11b, additional information was derived from the literature and from the typical method adopted by the country that recruited the ship. By contrast, in Hirahara et al. (2014)—hereafter H14—time-varying ratios of bucket to ERI observations were calculated such that global mean SST anomalies, following method-dependent bias adjustment, agreed for the data sets with extant and missing metadata.

In this study, we present a new assessment of SST measurement methods from 1855 to 2010. The method (section 2) relies on characteristic differences in the observed SST diurnal cycle between measurement methods. Results are presented in section 3 and conclusions in section 4.

2. Method

Diurnal variability is one of the dominant variations in SST (Clayson & Weitlich, 2005; Kennedy, Brohan, & Tett, 2007; Stuart-Menteth et al., 2003). Variations in solar heating and wind mixing lead to large differences in the magnitude of the SST diurnal cycle (Morak-Bozzo et al., 2016). Solar radiation and wind speed represent the most important factors affecting the diurnal warming in SST (Price et al., 1986). The size of the diurnal cycle decreases with depth (Kawai & Wada, 2007). Buckets, which sample at shallow depths, are therefore expected to show stronger diurnal variability than ERIs. Bucket observations may also show a diurnal signal from direct solar heating of the water sample in the bucket, in addition to real variations in SST. We therefore estimated the measurement method based on the characteristics of the diurnal cycle shown by subsets of observations compared to the expected diurnal cycle seen under similar solar radiation and wind speed conditions. This has been estimated following Morak-Bozzo et al. (2016)—hereafter MB16—modeling the SST diurnal anomaly, relative to the daily mean SST, as a function of the time of day, 10° latitude band, season, wind speed, and cloud cover based on drifting buoy observations and reanalysis model output.

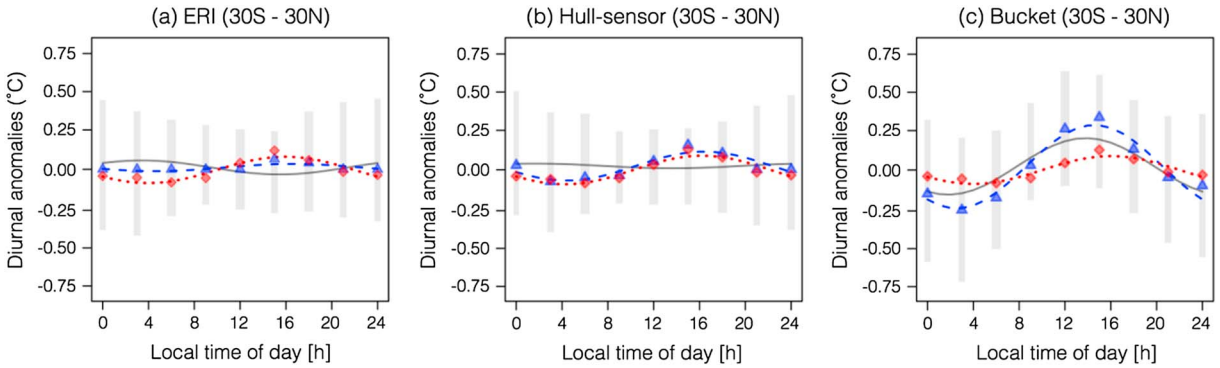

Figure 1. Diurnal anomalies (30°S–30°N) by measurement method, 1990–2006. (a) Diurnal cycle anomalies for ERI observations classified from SI(M) with SI = SIM, nonmissing (*blue triangles*) with fit overlaid (*dashed blue line*). Co-located MB16 diurnal anomalies (*red diamonds*) with fit overlaid (*dotted red line*). Residual fitted diurnal anomalies (*solid grey line*) and interquartile range (*grey bars*). (b) As Figure 1a but for hull-sensor observations. (c) As Figure 1a but for bucket observations. The widths of the grey bars indicate the relative size of each bin.

For individually identifiable ships, based on the ship identifier (ID) derived following Carella, Kent, and Berry (2017), local time diurnal SST anomalies relative to the local time daily mean SST were calculated. Only IDs with at least one observation in each quarter-day (local time 00 to 06 h, 06 to 12 h, etc.) were retained. For each observation, the expected diurnal SST anomaly was calculated following MB16 to accounting for local conditions at the time of observation. The differences between the observed and expected diurnal anomalies, hereafter residual anomalies, were then calculated.

Figure 1 shows the median observed and expected diurnal anomalies for ERI (Figure 1a), hull-sensors (Figure 1b), and buckets (Figure 1c) for three-hourly local time bins for 30S–30°N and 1990–2006. The median is calculated using all observations across all IDs, and the measurement method is assigned using the available metadata. The SI flag is used in preference to SIM, and SIM only used where SI is either missing or ambiguous—hereafter designated SI(M). Observations without a valid SI(M) have been discarded. Also shown are the parameterization of the diurnal anomalies and residual, calculated using a polynomial function of local time of day

$$\mathrm{SST}_{\mathrm{anomaly}}(t) = \alpha_0 + \alpha_1 \cdot \sin(\omega t) + \alpha_2 \cdot \cos(\omega t) \quad (1)$$

where $\alpha_0 \ldots \alpha_2$ are the fitting coefficients, t is the local time in hours, and $\omega = 2\pi/24\ \mathrm{h}^{-1}$. The fit is computed by ordinary least squares minimization, weighted by the number of observations in each three-hourly bin. As expected, ERI measurements and hull-sensors are characterized by "reversed" or near-zero residual anomalies. Bucket observations show an enhanced diurnal cycle, approximately twice the magnitude of the expected diurnal range from MB16 and also reported by previous authors (Clayson & Weitlich, 2007, report a median diurnal range of ~0.3°C in the tropics). Moreover, the maximum excursion occurs closer to local noon, suggesting that this excess diurnal cycle in bucket observations might be partly attributed to direct solar heating during measurement or to a residual signal from the bucket's on-deck temperature due to inadequate time spent in the water.

The analysis relies therefore on identifying diurnal variations that are either significantly larger, or smaller, than those seen on average by drifting buoys under similar conditions of wind speed and solar radiation. The approach does not require an estimate of the diurnal excursion on any individual day, which would require quantification of many additional terms that are typically unknown for these observations (Bernie et al., 2005).

The measurement method for well-sampled subsets of observations (defined as more than 10^4 observations and more than 10^3 observations in each quarter-day) was estimated from the characteristic shape of the residual diurnal variability fitted using equation (1). When the fit was significant at the 5% level (p-value < 0.05) and characterized by a local maximum (first derivative zero and second derivative positive) and no local minima during daylight hours (9–18 h local time), the subset was classified as buckets, otherwise (no local maxima but with a local minimum, or, no local maxima or minima) as ERI. When the fit was not significant at the 5% level, the subset was counted as ERI only when the range of the residual diurnal anomalies was

close to zero (<0.05°C); otherwise, the measurement method for that subset was recorded as unknown. This last criterion accounts for ERI/hull-sensor observations showing diurnal anomalies very similar to MB16, that otherwise would remain unassigned.

Measurements from research vessels (see Table S2.1 in the supporting information) and hull-sensors show a diurnal cycle similar to MB16 regardless of their SI(M) flag so we retained their original classification. Our method cannot distinguish between ERI and hull-sensor observations so our ERI classification contains both methods.

This fitting method is used to assign measurement methods to SST observations from ICOADS Release 2.5 (Woodruff et al., 2011). Subsets of data expected to use the same measurement method were identified and analyzed. Typically, each ship-operating country has a preference for a particular measurement method (K11b), and some ICOADS data sources (known as decks) derive from archives of particular countries. Country is expected to be more consistently associated with method than deck, but is not available for every observation, so subsets are also analyzed according to deck. The approach is described in more detail in section S2 in the supporting information. To account for variations over time and regionally, the analysis was performed for 10 year intervals and by 60° latitude band. Five different realizations were performed changing the start date by 2 years (1850, 1852, 1854, 1856, and 1858).

3. Results and Discussion

Typically for more than 75% of the observations, we could diagnose the measurement method from the diurnal anomalies without relying on SI(M), although higher percentages of unclassified observations are present at the beginning of the record and during WWII (see section S2 in the supporting information). These periods are characterized by poor sampling of the groups used to partition the data and/or by observations that are too noisy to derive a clean signal. The method is robust to the choice of 10 year time windows, with the results obtained for different start dates consistent for most of the record (Figure 2a). However, during WWII differences among ensemble members are larger, more than 30% of the number of reports, with also a larger number of remaining "unknown" measurements, suggesting that in this period rapid changes in the observation practice are occurring. Moreover, apart from few cases discussed later, the method typically identifies the same measurement method as that reported by SI(M), giving us confidence that our classification is correct (see also Table S2.2 in the supporting information). The remaining unknown measurements were then randomly reassigned to ERIs and buckets.

Figure 2b compares the percentage of reports here identified as buckets and in K11b and H14 from 1930 onward. Before 1930, the presence of ERI metadata is negligible, and before about 1940 both K11b and H14 assume all unknown observations to be buckets (also verified by our method, see section S2 in the supporting information). Recall that K11b assumed that the mean proportion (bucket:ERI) of observations with missing SI(M) was the same as that for observations with known method from SI(M) supplemented with method preferences by country or deck (section 1) and the assumption that ERI would be preferred on large or fast ships; similar to the approach taken here, the remaining unclassified observations were randomly reallocated to buckets and ERIs, with the uncertainty in this assumption explored within an ensemble.

Differences between the classification derived here and in K11b are largest during and just after WWII and in the 1960s and 1970s. UK observations during and immediately after WWII classified by SI(M) as buckets were identified by our method as ERI measurements, compatible with the idea of a switchover from buckets to ERI during this period. K11b assumed that these observations were from buckets giving a rapid increase in the bucket proportion immediately post-WWII.

During WWII, our method identifies a larger percentage of buckets than both K11b and H14, although in this period more than half the observations remain unclassified (see Figure S2.2 in the supporting information).

During the 1960s and 1970s, we found that the majority of U.S. observations—including those classified as buckets by SI(M)—were classified as ERI. In contrast, our results confirmed the presence of a small percentage (less than 4%) of U.S. bucket observations in the 1980s and 1990s, according to SI(M). K11b assumed that all U.S. observation post-WWII were made by ERI. In the period 1960–1975 observations from Dutch and Russian ships have sparse method information, which indicates bucket observations. However, we identified the much larger proportion with unknown method from these countries as ERI, explaining much of the

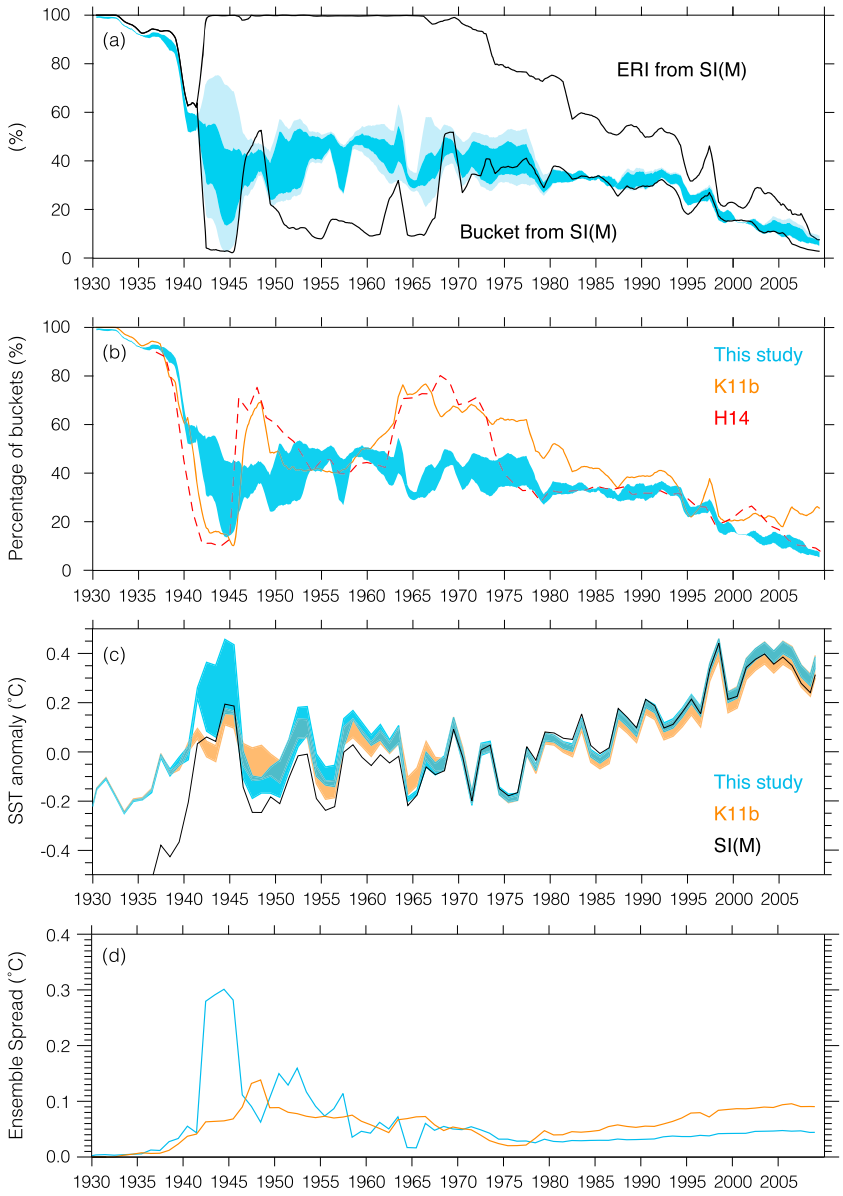

Figure 2. (a) Percentage of observations identified as ERIs and buckets from SI(M) (*black lines*) and in this study (*dark blue shading*: ensemble range (buckets: within/below; ERIs: within/above); *light blue shading*: ensemble mean percentage of the unknown measurements, randomly reassigned to ERIs and buckets). (b) Percentage of buckets identified in this study (*dark blue shaded area*, as Figure 2a), in K11b (*orange solid line*, median of the ensemble) and in H14 (*red dashed line*). (c) SST anomaly (°C) for bias adjusted observations classified according to this study (*dark blue shaded area*, uncertainty given at the 95% confidence level), K11b (*orange shaded area*, uncertainty given at the 95% confidence level) and from SI(M). (*d*): As Figure 2c but for the ensemble spread in this study and in K11b. All lines represent 12 month running means.

reduction in bucket measurements in this period. Data from Russian deck 732 were analyzed separately (see section S3 in the supporting information) and were also classified as ERI.

Different classifications will impact on long-term SST trends because different observational biases characterize each SST method (Kent et al., 2017). To test the sensitivity of the SST field on the inferred mix of methods, we compared the effect of our classifications and K11b on global average SST anomalies computed combining SST observations from ships and, for the most recent period, buoy-derived SSTs (Figure 2c). Because we are interested only in those components of the bias adjustment ensemble that are associated with metadata uncertainty, for each classification, the SST record was bias adjusted following a simplified version of the bias

adjustment method applied in K11b. We used area-weighted global averages with a fixed bias of +0.2°C for ERI observations and the ensemble median of the bias field realizations for bucket observations; the uncertainty in the timing of the transition from canvas to rubber buckets is explored with an ensemble following K11b. After 1980 there is an offset between the two estimates. During the climatological period (1961–1990) K11b classification is characterized by more buckets and fewer ERIs (Figure 2b). Our different mix of observations leads to a colder climatology, which means the modern period anomalies are warmer. Earlier on, the two estimates show also some differences but the uncertainty in our assignments is much larger (Figure 2d), especially during WWII. In contrast, from 1980 onward, the ensemble spread of the SST anomalies for our classification is narrower, as a result of fewer unknown measurements after 1980, possibly fewer misassignments and fewer buckets in the climatology period, which implies a diminished impact of the uncertainty in the transition dates from canvas to rubber buckets (Kent et al., 2017). Moreover, compared to K11b, our method reduces the uncertainty-on-the-uncertainty, testing some of K11b assumptions and existing metadata, and with uncertainty estimates that are traceable to objective criteria.

In H14 the fraction of reports for observations with unknown measurement method is derived requiring the bias-adjusted global mean of SST anomalies to be equivalent to that for the known types. However, because different recruiting countries prefer different measurement types and because of the difference in national shipping routes, there are regional variations in the bucket:ERI ratio (see Kent & Taylor, 2006, for regional maps). This inhomogeneous distribution of SST measurement methods is likely to affect both global and hemispheric means and represents a significant limitation to the method of H14 (and indeed any method), which is based on the consistency of the global means only.

Improved metadata will also help to refine assessments of bucket and ERI biases by ensuring a cleaner separation of the two groups. A mean negative bias is expected for bucket measurements (Carella et al., 2017; Folland & Parker, 1995), and ERI observations are more likely to be characterized by a mean warm bias (K11b; Kent et al., 2017). Figure 3a shows for the period 1955–1995 the global mean difference between bucket and ERI SST monthly anomalies, which are computed relative to a climatology derived from Merchant et al. (2014). The anomalies were computed according to the classification of the observations derived both from the results of this analysis and from SI(M). The results obtained agree well, both in changes over time and in seasonal variability, despite the increased number of observations included adopting the classification derived in this study (Figure 2a). Moreover, compared to the results derived from K11b, before mid-1960s and after about 1970, our classification shows a better defined annual cycle and a clearer offset between the methods. From mid-1970s, when the comparison is possible, our method shows better agreement with the classification from SI(M). On the other hand, between 1964 and 1968, the large drop in bucket observations relative to both K11b and H14 (Figure 2a) in our classification does not result in an improved separation between the methods, as differences become smaller. This suggests that in this period the SST data might be of poorer quality compromising some of the assignments.

Despite these differences, for both analyses, the differences between bucket and ERI SST anomalies show variability on seasonal, interannual, and longer time scales, which is not well reproduced by the bias model applied in K11b and other existing SST analyses (e.g., H14), as Figure 3b shows. The most obvious signal in the difference time series is the reduction in magnitude of the difference before 1965 from about −0.5°C to close to −0.25°C around mid-1970s, followed by a sudden drop between 1980 and 1983. After that, the difference reduces gradually to around zero, or slightly positive by 1995. Lack of bucket measurements makes later comparisons impossible. A reduction in difference between the methods could be attributed to a decrease in magnitude of either the warm bias typically seen in ERI measurements, or the cold bias typically seen in bucket measurements. Seasonal differences indicate, as expected, that bucket measurements are relatively cooler in winter than in summer, because of larger heat loss. Before 1960, the size of the seasonal cycle of the differences is increased, which might suggest the use of uninsulated canvas buckets. However, the time series of ERI anomalies (not shown) indicates that in this period ERI anomalies are both noisier and warmer than in later periods, which could lead to a poorly defined seasonal cycle in the ERI measurements, therefore explaining some of the excess seasonal cycle in the difference. This result also suggests that ERI observations were of poorer quality before 1960 and that the typically warm biases characterizing ERI observations may reduce over time (Kent & Kaplan, 2006). After 1970, the size of the annual cycle in the bucket-ERI differences does not increase over time, suggesting that, although of different designs (Kent et al., 2017), the buckets adopted in this period were mostly insulated, with uninsulated canvas buckets largely out of use by the

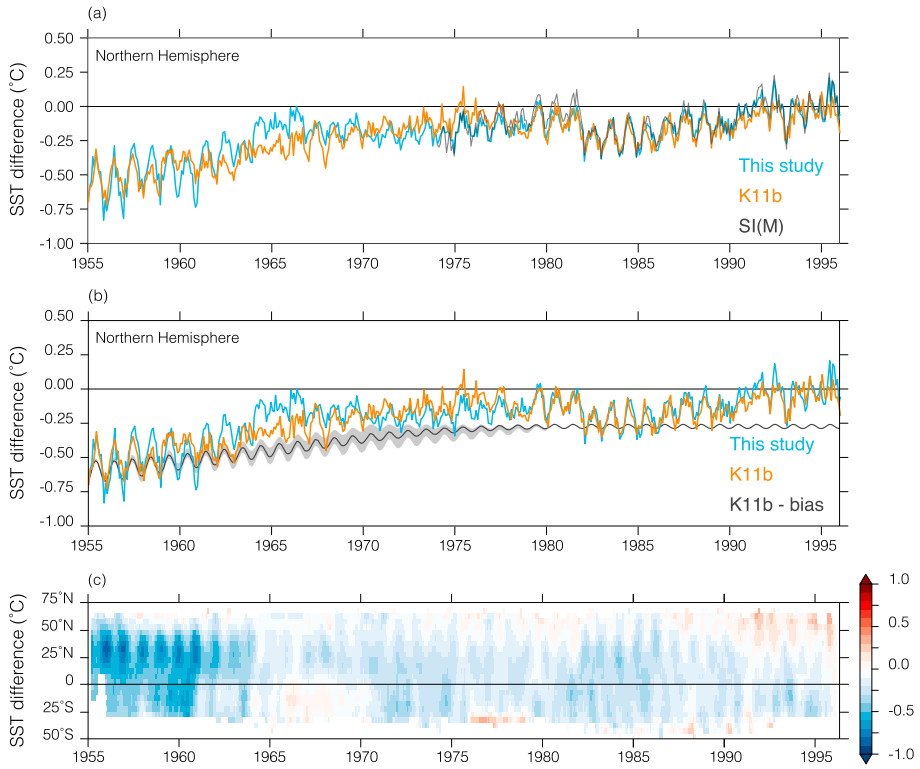

Figure 3. (a) Northern Hemisphere mean difference between bucket and ERI SST anomalies (°C) computed from the classification derived in this study (*blue line*, representing the mean of the ensemble), from SI(M) (*black line*) and from K11b (*orange line*, representing the mean of the ensemble). (b) As Figure 3a and for the bias ensemble (spread and median) computed assuming the classification derived in this study and with the method described in the text (*grey shaded area and grey line*). (c) Hovmöller diagram of the difference between bucket and ERI SST anomalies (°C) computed from the classification derived in this study, averaged on a 5° latitude by 5° longitude grid and smoothed with a 3 × 3 moving window. The anomalies were computed on a 5° grid using values only for grid cells and months where there were at least 10 observations from each method and using a monthly climatology derived from the gap-filled, daily blend ESA SST CCI analysis (Merchant et al., 2014), available for the period 1991–2010 on a 0.05° grid. In K11b the anomalies were averaged on a 5° grid with any number of observation in a grid box and relative to HadSST2 climatology for 1961 to 1990 (Rayner et al., 2006).

1970s, consistent with James and Fox (1972). This result is consistent with the findings in H14, where the phasing out of uninsulated canvas buckets is estimated to be nearly complete around 1962 (K11b; Kent et al., 2017). In K11b the latest switchover dates to insulated (rubber) buckets allowed in the HadSST3 ensemble is around 1980 and the phasing out of canvas bucket happens more slowly.

These decadal variations can be observed at all latitudes (Figure 3c) and could be linked to changes in the ships' measurement systems. After 1970, it is likely that all the buckets are insulated so the most likely cause of changes is in the ERI measurements (e.g., caused by changes in the mixture of individual vessels and fleets). However, in addition to any changing biases due to changing measurement systems, these simple differences will show some variability due to either large-scale changes in atmospheric conditions affecting the heat exchange experienced by the bucket samples, or any real changes between temperatures at the typical depths sampled by the different methods.

4. Conclusions

Changing observational practice and measurement methods are responsible for pervasive systematic biases in the SST historical record similar in magnitude to the climatic signal (Jones, 2016; Kent et al., 2017). Empirically based bias adjustment models require classification of measurement methods; however, this information is often missing and sometimes unreliable (Kennedy, 2014; Kent et al., 2010). Current methods to estimate the number of bucket and ERI reports (K11b and H14) rely heavily on known metadata and on the characteristics of reports for which the metadata are available. Here we develop a method to diagnose

the measurement practice by comparing observed SST diurnal anomalies from ships with a reference derived from drifting buoys (MB16) under similar conditions of wind speed and solar radiation. Bucket measurements are characterized by a larger diurnal cycle (compared to MB16), and the maximum excursion between the observed and estimated diurnal cycle occurs close to local noon (see Figure 2c). In contrast, the diurnal cycle observed by ERIs is reduced because typically ERI sample seawater at a greater depth.

Compared to existing estimates, the method suggests a larger proportion of bucket reports during WWII (although relatively few observations are identified in this period) and a larger number of ERI reports post-WWII and in the period 1960–1980. Confidence in the classification derived here comes from demonstrating that our method typically infers the same measurement type as that derived from SI(M). Additionally, from mid-1970s, when the comparison is possible, differences between bucket and ERI SST anomalies derived from our classification and from SI(M) agree better in changes over time and in seasonal variability than previous results (K11b). Anomaly differences in K11b tend also to underestimate seasonal variations and have a smaller offset during the 1970s, indicating a poorer separation of the measurement methods. Based on this diagnostic, our method performs well before mid 1960s, and after 1970, but gives poor results between 1964 and 1968. This may indicate data quality problems in this period affecting the representation of diurnal variations.

These findings have important implications for the estimation of biases in the SST record, and reconciliation of differences between measurements will improve estimates of SST trends and variability. Differences in the inferred mixture of observations lead to differences in the bias adjusted SST record, with our findings indicating a systematic warmer shift of the SST anomalies from 1980 onward compared to the results derived applying the same bias adjustment method to the record classified according to K11b. From 1980 onward, on the very carefully defined subset of uncertainties explored in this study, our method also reduces the ensemble spread of the bias adjusted SST anomalies.

Our results highlight the shortcomings of current bias adjustment models. By assuming a fixed bias for ERI observations these models are not able to reproduce the changes in mean field differences between bucket and ERI SST anomalies, which most likely reflect changes in biases in ERI measurements that occurred in the period 1955–1995. Our results also suggest that the phasing out of uninsulated (canvas) buckets was probably completed before the 1970s, and therefore earlier than currently assumed in the K11b bias adjustment ensemble.

The classification of SST observational method presented in this study shows the need for, and enables, further work to understand and reconcile the differences in SST from different methods and thereby reduce the uncertainty in the SST record and hence in estimates of global surface temperature change.

Acknowledgments

The authors wish to acknowledge use of the R software for analysis and graphics in this paper (R Development Core Team, 2016). G.C., E.K., and D.B. were funded by the Natural Environment Research Council (NERC) through grant NE/J020788/1. S.M.-B. and C.J.M. were funded by NERC through grant NE/J02306X/1. J.K. was supported by the Joint UK BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101). S.H. was supported by the Environment Research and Technology Development Fund of the Ministry of the Environment, Japan (2-1506). The data used in this study can be obtained from https://rda.ucar.edu/datasets/ds540.0/.

References

Bernie, D. J., Woolnough, S. J., Slingo, J. M., & Guilyardi, E. (2005). Modeling diurnal and Intraseasonal variability of the ocean mixed layer. *Journal of Climate*, *18*(8), 1190–1202. https://doi.org/10.1175/JCLI3319.1

Carella, G., Kent, E. C., & Berry, D. I. (2017). A probabilistic approach to ship voyage reconstruction in ICOADS. *International Journal of Climatology*, *37*(5), 2233–2247. https://doi.org/10.1002/joc.4492

Carella, G., Kent, E. C., Berry, D. I., Morak-Bozzo, S., & Merchant, C. J. (2017). Measurements and models of the temperature change of water samples in sea surface temperature buckets. *Quarterly Journal of the Royal Meteorological Society*, *143*(706), 2198–2209. https://doi.org/10.1002/qj.3078

Clayson, C. A., & Weitlich, D. (2005). Interannual variability of tropical Pacific diurnal sea surface temperature warming and nighttime cooling. *Geophysical Research Letters*, *32*, L21604. https://doi.org/10.1029/2005GL023786

Clayson, C. A., & Weitlich, D. (2007). Variability of tropical Diurnal Sea surface temperature. *Journal of Climate*, *20*(2), 334–352. https://doi.org/10.1175/JCLI3999.1

Folland, C. K., & Parker, D. E. (1995). Correction of instrumental biases in historical sea surface temperature data. *Quarterly Journal of the Royal Meteorological Society*, *121*(522), 319–367. https://doi.org/10.1002/qj.49712152206

Freeman, E., Woodruff, S. D., Worley, S. J., Lubker, S. J., Kent, E. C., Angel, W. E., … Smith, S. R. (2017). ICOADS release 3.0: A major update to the historical marine climate record. *International Journal of Climatology*, *37*(5), 2211–2232. https://doi.org/10.1002/joc.4775

Hartmann, D. L., Klein Tank, A. M. G., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y., … Zhai, P. M. (2013). Observations: Atmosphere and surface. In T. F. Stocker, et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change* (Chap. 2, pp. 159–254). Cambridge, United Kingdom and New York: Cambridge University Press.

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

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., … Zhang, H.-M. (2017). Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, validations, and Intercomparisons. *Journal of Climate*, *30*(20), 8179–8205. https://doi.org/10.1175/JCLI-D-16-0836.1

James, R., & Fox, P. (1972). Comparative sea surface temperature measurements in WMO reports on marine science affairs. Tech. Rep. 336, 5
Jones, P. D. (2016). The reliability of global and hemispheric surface temperature records. *Advances in Atmospheric Sciences*, *33*(3), 269–282. https://doi.org/10.1007/s00376-015-5194-4
Kawai, Y., & Wada, A. (2007). Diurnal sea surface temperature variation and its impact on the atmosphere and ocean: A review. *Journal of Oceanography*, *63*(5), 721–744. https://doi.org/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. *Reviews of Geophysics*, *52*, 1–32. https://doi.org/10.1002/2013RG000434
Kennedy, J. J., Brohan, P., & Tett, S. F. B. (2007). A global climatology of the diurnal variations in sea-surface temperature and implications for MSU temperature trends. *Geophysical Research Letters*, *34*, L05712. https://doi.org/10.1029/2006GL028920
Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E., & Saunby, M. (2011a). Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *Journal of Geophysical Research*, *116*, D14103. https://doi.org/10.1029/2010JD015218
Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E., & Saunby, M. (2011b). Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *Journal of Geophysical Research*, *116*, D14104. https://doi.org/10.1029/2010JD015220
Kent, E. C., & Kaplan, A. (2006). Toward estimating climatic trends in SST. Part 3: Systematic biases. *Journal of Atmospheric and Oceanic Technology*, *23*(3), 487–500. https://doi.org/10.1175/JTECH1845.1
Kent, E. C., Kennedy, J. J., Berry, D. I., & Smith, R. O. (2010). Effects of instrumentation changes on sea surface temperature measured in situ. *WIREs Climate Change*, *1*(5), 718–728. https://doi.org/10.1002/wcc.55
Kent, E. C., Kennedy, J. J., Smith, T. M., Hirahara, S., Huang, B., Kaplan, A., … Zhang, H.-M. (2017). A call for new approaches to quantifying biases in observations of sea-surface temperature. *Bulletin of the American Meteorological Society*, *98*(8), 1601–1616. https://doi.org/10.1175/BAMS-D-15-00251.1
Kent, E. C., & Taylor, P. (2006). Toward estimating climatic trends in SST. Part I: Methods of measurement. *Journal of Atmospheric and Oceanic Technology*, *23*(3), 464–475. https://doi.org/10.1175/JTECH1843.1
Kent, E. C., Taylor, P. K., Truscott, B. S., & Hopkins, J. S. (1993). The accuracy of voluntary observing ship's meteorological observations—Results of the VSOP-NA. *Journal of Atmospheric and Oceanic Technology*, *10*(4), 591–608. https://doi.org/10.1175/1520-0426(1993)010%3C0591:TAOVOS%3E2.0.CO;2
Kent, E. C., Woodruff, S. D., & Berry, D. I. (2007). Metadata from WMO publication no. 47 and an assessment of voluntary observing ships observation heights in ICOADS. *Journal of Atmospheric and Oceanic Technology*, *24*(2), 214–234. https://doi.org/10.1175/JTECH1949.1
Merchant, C. J., Embury, O., Roberts-Jones, J., Fiedler, E., Bulgin, C. E., Corlett, G. K., … Donlon, C. (2014). Sea surface temperature datasets for climate applications from phase 1 of the European Space Agency Climate Change Initiative (SST CCI). *Geoscience Data Journal*, *1*(2), 179–191. https://doi.org/10.1002/gdj3.20
Morak-Bozzo, S., Merchant, C. J., Kent, E. C., Berry, D. I., & Carella, G. (2016). Climatological diurnal variability in sea surface temperature characterized from drifting buoy data. *Geoscience Data Journal*, *3*(1), 20–28. https://doi.org/10.1002/gdj3.35
Price, J. F., Weller, R. A., & Pinkel, R. (1986). Diurnal cycling: Observations and models of the upper ocean response to diurnal heating, cooling, and wind mixing. *Journal of Geophysical Research*, *91*(C7), 8411–8427. https://doi.org/10.1029/JC091iC07p08411
R Development Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rayner, N. A., Brohan, P., Parker, D. E., Folland, C. K., Kennedy, J. J., Vanicek, M., … Tett, S. F. (2006). Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset. *Journal of Climate*, *19*(3), 446–469. https://doi.org/10.1175/JCLI3637.1
Smith, T. M., & Reynolds, R. W. (2002). Bias corrections for historic sea surface temperatures based on marine air temperatures. *Journal of Climate*, *15*(1), 73–87. https://doi.org/10.1175/1520-0442(2002)015%3C0073:BCFHSS%3E2.0.CO;2
Stuart-Menteth, A. C., Robinson, I. S., & Challenor, P. G. (2003). A global study of diurnal warming using satellite-derived sea surface temperature. *Journal of Geophysical Research*, *108*(C5), 3155. https://doi.org/10.1029/2002JC001534
Thompson, D. W. J., Kennedy, J. J., Wallace, J. M., & Jones, P. D. (2008). A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature*, *453*(7195), 646–649. https://doi.org/10.1038/nature06982
Woodruff, S. D., Worley, S. J., Lubker, S. J., Ji, Z., Freeman, E. J., Berry, D. I., … Wilkinson, C. (2011). ICOADS release 2.5: Extensions and enhancements to the surface marine meteorological archive. *International Journal of Climatology*, *31*(7), 951–967. https://doi.org/10.1002/joc.2103