Systematic Differences in Bucket Sea Surface Temperature Measurements among Nations Identified Using a Linear-Mixed-Effect Method

DUO CHAN AND PETER HUYBERS

Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts

(Manuscript received 29 August 2018, in final form 28 January 2019)

ABSTRACT

The International Comprehensive Ocean–Atmosphere Dataset (ICOADS) is a cornerstone for estimating changes in sea surface temperatures (SST) over the instrumental era. Interest in determining SST changes to within 0.1°C makes detecting systematic offsets within ICOADS important. Previous studies have corrected for offsets among engine room intake, buoy, and wooden and canvas bucket measurements, as well as noted discrepancies among various other groupings of data. In this study, a systematic examination of differences in collocated bucket SST measurements from ICOADS3.0 is undertaken using a linear-mixed-effect model according to nations and more-resolved groupings. Six nations and a grouping for which nation metadata are missing, referred to as ''deck 156,'' together contribute 91% of all bucket measurements and have systematic offsets among one another of as much as 0.22°C. Measurements from the Netherlands and deck 156 are colder than the global average by −0.10° and −0.13°C, respectively, both at $p < 0.01$, whereas Russian measurements are offset warm by 0.10°C at $p < 0.1$. Furthermore, of the 31 nations whose measurements are present in more than one grouping of data (i.e., deck), 14 contain decks that show significant offsets at $p < 0.1$, including all major collecting nations. Results are found to be robust to assumptions regarding the independence and distribution of errors as well as to influences from the diurnal cycle and spatially heterogeneous noise variance. Correction for systematic offsets among these groupings should improve the accuracy of estimated SSTs and their trends.

1. Introduction

Currently available gridded sea surface temperature (SST) products are based on measurements collected together under the auspices of the International Comprehensive Ocean–Atmosphere Dataset (ICOADS; Freeman et al. 2017). Measurements are of water contained within the upper several meters of the ocean that was variously collected using buckets or engine room intakes, or measured in situ using moored or drifting buoys. Bucket measurements are the dominant source of instrumental SSTs in ICOADS prior to 1941. Measurements from engine-room intakes appear in the 1930s and become common starting in 1941, and buoy and drifter measurements start in the 1980s. Absolute numbers of bucket measurements decline in the 2000s (Kennedy et al. 2011b). We focus on bucket measurements because they are the dominate source of SST measurements prior to 1941 and on systematic biases because of their potential consequences for global trend estimates (Kennedy 2014). Bucket measurements have biases that are estimated to range from −1° to +0.1°C depending on various evaporative, sensible, and solar heat fluxes (Folland and Parker 1995). Given interest in reconstructing regional and global SST changes to the order of 0.1°C, accurate bias correction for bucket measurements is necessary.

An approach widely applied to estimate bucket measurement bias is to thermodynamically model the heating and cooling influences to which water in a bucket is exposed (Folland and Parker 1995). Running such a model with representative meteorological fields yields spatially and seasonally varying correction patterns. The parameters in such a model are, however, generally underconstrained because detailed information regarding parameters such as bucket type, size, and on-deck time are not generally available (Ashford 1948). A common simplification is to assume only two bucket types

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-18-0562.s1.

Corresponding author: Duo Chan, duochan@g.harvard.edu

DOI: 10.1175/JCLI-D-18-0562.1

involving wooden and canvas varieties whose relative proportions vary linearly with time (Folland and Parker 1995; Rayner et al. 2006; Kennedy et al. 2011b). Another approach is to compare SST observations against nighttime marine air temperatures, but where a constant spatial pattern of correction whose amplitude varies slowly in time is typically also used (Smith and Reynolds 2002; Huang et al. 2015). Although commensurate with the limited metadata available for determining observational characteristics, these simplifying assumptions regarding the space–time structure of bucket biases lead to incomplete bias correction (Kennedy 2014).

The presence of biases in specific measurement methods coupled with systematic changes in where and when those methods are applied can impart artificial jumps or trends in SST time series. For example, a sudden drop in global temperature by 0.3°C in 1945 was identified to arise from offsets between engine-room intake and bucket temperature estimates (Thompson et al. 2008; Kennedy et al. 2011b). More recently, difficulty in simulating a slowdown in global warming (Fyfe et al. 2013) was partly reconciled (Medhaug et al. 2017) by adjusting SST bias corrections that led to 0.064°C decade^{-1} more warming between 2000 and 2014 (Karl et al. 2015). Given the significant implications of these past adjustments, identification of any further systematic offsets in SST records could also have important implications for reducing bias and uncertainty in global temperature trends (Jones and Wigley 2010; Kennedy 2014).

There are several indications that certain groups of bucket measurements contain unique bias structures within the ICOADS dataset. Various nations and fleets are documented to have used buckets with distinct rates of sensible, evaporative, and solar heat fluxes, as well as thermal equilibration times (Folland and Parker 1995), and those differences may not be fully captured by dichotomizing the bucket dataset into wooden and canvas types (Folland and Parker 1995; Kennedy et al. 2011b). In the early 1910s, for example, the German navy used a kind of bucket that cooled relatively quickly due to its small size (Ashford 1948). Beyond issues of differing bucket characteristics, groups of ships may follow different measurement protocols. Anecdotal evidence exists, for example, that some Japanese bucket measurements are biased cold (Uwai and Komura 1992), possibly because the Kobe Imperial Marine Observatory prescribed waiting until thermometer readings were stable—thus exposing water samples to increased evaporative cooling—as opposed to stipulating a shorter on-deck measurement time (Folland and Parker 1995). A systematic examination of whether statistically significant offsets exist across groups of bucket SST measurements thus appears warranted.

2. Data and methods

SST measurements used in this study come from ICOADS3.0 (Freeman et al. 2017). Because important details regarding how measurements in ICOADS3.0 were collected and archived are not always available, the specific methodology for identifying bucket SST measurements and screening for outliers is an important feature of this study, which we outline in this section. All bucket data available in ICOADS3.0 between 1850 and 2014 are analyzed. To facilitate intercomparison between bucket measurements, we also describe techniques to account for spatial, seasonal, and diurnal offsets between measurements. As the final part of this section, motivated by the structure of ICOADS3.0, we present a linear-mixed-effect model to evaluate whether systematic differences exist among more-resolved groups of SST data.

a. SST data

1) Identification and quality control

The source of an SST observation within ICOADS3.0 is not always explicitly provided. Following the same procedure used for HadSST3 (Kennedy et al. 2011b), we identify bucket measurements using World Meteorological Organization Report 47 (WMO47; Kent et al. 2007) and ICOADS metadata. Prior to 1941, all SST measurements are assumed to be from buckets unless explicitly recorded otherwise. Analysis of the amplitude of the diurnal cycle in SST before 1941 supports unidentified records as being overwhelmingly from buckets (Carella et al. 2018). From 1941 onward, if the method of measurement is missing in both WMO47 and ICOADS metadata, SST measurements are assumed to come from buckets if the associated nations are reported to have at least 95% of their ships making bucket measurements in WMO47 (Table 1).

Quality control of raw bucket SST data follows three steps. First, 10% of data identified as coming from buckets are omitted because quality-control flags indicate erroneous reports or outlier behavior. Specifically, the SST trimming flag (SF) or the National Climate Data Center (now known as the National Centers for Environmental Information) quality-control flag (SNC) is greater than 5. Second, 0.04% of the remaining data have temperatures warmer than 37°C or colder than −5°C and are removed for being unphysical. These thresholds are based on a physically plausible range from −1.8° to 34°C (Kleypas et al. 2008) along with a

TABLE 1. Nation and deck designations. Decks for which a nation's presence is inferred from deck descriptions, as opposed to being explicitly given, are shown in boldface. The fixed effects of groups that significantly depart from zero in the nation-level analysis are indicated using one asterisk (*) for $p < 0.1$ and two asterisks (**) for $p < 0.01$. Decks that significantly differ from other decks are likewise marked with asterisks.

Abbreviation	Full name	Fixed effect (°C)	ICOADS 3.0 deck
AR	Argentina	0.35**	732, 780, 927
AU	Australia	−0.02	246, **750***, 780, **900**, 926, 927*
BE	Belgium	−0.11	926, 927, 928
BR	Brazil	−0.34**	780, 926, 927
CA	Canada	0.19**	780*, 926, 927
CN	China	0.03	**781**
DE	Germany	0.01	**151**, **192****, **215***, **720**, **721**, 732**, 780, 926, 927
DK	Denmark	−0.06	926, 927
EG	Egypt	0.10	927
ES	Spain	0.31**	926*, 927*
FR	France	0.06	732, 780, 926, 927
GB	Great Britain	0.02	**152**, **184**, 201, 202, 203, **204**, **205***, 206, **216**, 221, 230, **245**, 246, 247, **249***, 732, 780, 926, 927, 928
HK	Hong Kong	0.09	926**, 927**, 928**
HR	Croatia	0.34**	926
IE	Ireland	−0.31**	926, 927
IL	Israel	0.42**	926*, 927*
IN	India	0.38**	780, 926, 927
IS	Iceland	−0.07	926, 927
JP	Japan	−0.06	**118****, **187**, 780, **898**, 926*, 927
KE	Kenya	0.36*	926, 927
MY	Malaysia	0.40**	926, 927
NL	Netherland	−0.10**	**150**, **193**, 732**, 780, 926, 927
NO	Norway	0.17**	**188**, **702**, 780, 926, 927
NZ	New Zealand	0.13*	780, 926, 927
PH	Pakistan	0.16	926, 927
PK	Philippines	−0.07	927
PL	Poland	0.38**	926, 927
PT	Portugal	0.13*	780, 926, 927
RU	Russia	0.10*	**185**, **731**, **732**, **735***, 926, 927
SE	Sweden	0.16*	926*, 927*
SG	Singapore	0.21**	926, 927
TH	Thailand	0.15	926, 927
TZ	Tanzania	−0.37*	927
UG	Uganda	−0.18	927
US	United States	0.03	**116***, **218**, **281**, **701**, **703**, **704**, **705****, **706**, **707**, **710****, 732, 780*, 874*, **888**, 889, 892, 926*, 927
YU	Uruguay	0.41**	926**, 927**
ZA	South Africa	0.07	**899**, 926*, 927
Deck 155	—	−0.13*	—
Deck 156	—	−0.13**	—
Deck 197	—	0.03	—
Deck 201	—	0.02	—
Deck 209	—	0.09	—
Deck 210	—	0.12	—
Deck 255	—	0.20*	—
Deck 700	—	−0.01	—
Deck 740	—	0.08	—
Deck 749	—	0.14	—
Deck 792	—	0.05	—
Deck 849	—	0.16	—
Deck 874	—	0.61	—
Deck 889	—	0.13	—
Deck 896	—	−0.32	—
Deck 901	—	−0.15	—
Deck 926	—	0.30*	—
Deck 927	—	−0.02	—
Deck 992	—	0.10	—

three-standard-deviation observational error (Kent and Challenor 2006). Finally, another 6% of measurements are excluded on the basis of the results obtained by implementing the buddy-check methodology of Rayner et al. (2006). In total 63 million bucket SST measurements pass quality control.

2) PAIRING SST MEASUREMENTS

Similar to a previous study regarding observational errors between 1970 and 1997 (Kent and Challenor 2006), bucket measurements are grouped according to collecting countries under the assumption that countries share more consistent methods. Country information is identified using the ICOADS country code. When the country code is missing, we infer countries from deck information (Kennedy et al. 2011b; see updates in Table 1). Decks initially referred to punch cards used for storing the data and are now the basic unit according to which ICOADS data are organized (Freeman et al. 2017). Such a grouping yields a total of 37 distinct countries. In cases where country information is neither indicated nor inferable (Carella et al. 2018), measurements are grouped and analyzed only according to decks, which accounts for another 19 groups, giving a total of 56 groups at the "national" level.

To assess differences among national-level groups of data, nearby bucket measurements from different groups are paired. Measurements are paired only if they come from distinct nations and are within 300 km and 2 days of one another. Each measurement is used only once in order to prevent the introduction of error covariance between pairs. The algorithm we use to identify pairs prioritizes those measurements that are closest in space. Specifically, all potential pairs within a given month are rank ordered according to distance, and the closest pair is selected. After discarding all pairs involving previously selected measurements, the next closest pair of data is selected, and so on. This search yields 16.2 million pairs of data, involving about half of the 63 million bucket SST measurements that pass quality control. Results are not qualitatively sensitive to using a more stringent spatial threshold of 100 or 200 km for purposes of pairing, or if pairs are also prioritized according to groups having fewer data. Sensitivity to methodological choices is more thoroughly described in section 5.

There are on average 170 000 pairs of data in each year between 1900 and 2000, but with relatively few observations before 1900, during the two world wars, and after 2000 (Fig. 1). Spatially, observations are abundant in the North Atlantic, shipping corridors in the North Pacific, and trade routes to South America (Figs. 1 and 2).

3) CLIMATOLOGICAL OFFSETS

More accurate intercomparison of proximate SST observations is possible through removing systematic offsets associated with spatial and temporal separation. For example, SSTs measured toward higher latitudes or temporally nearer minimum winter temperature may be expected to be colder. To correct for these climatological offsets in pairwise comparisons of SST measurements, we use optimally interpolated SSTs (OI-SST; Reynolds et al. 2007), which provide estimates of daily SST at 0.25° resolution. Under the assumption that oceanic climatological gradients have changed little, we compute the OI-SST average for a given location and day across 1982 –2014 and remove it from each bucket SST measurement.

SSTs obviously also vary diurnally (Kennedy et al. 2007; Morak-Bozzo et al. 2016; Carella et al. 2018). A correction for climatological diurnal variability is made using hourly observations from buoy and drifter data between 1990 and 2014 in ICOADS3.0. The shape of the climatological diurnal cycle can vary substantially, with first-order differences having to do with seasons and latitudinally dependent differences in diurnal variation in incoming solar radiation. We, therefore, follow Morak-Bozzo et al. (2016) and first empirically estimate the shape of the diurnal cycle at 5° latitude bands for each month. The amplitude of the predetermined diurnal shape is then fit for each 5° × 5° grid using least squares. The estimated diurnal anomaly in SST is removed from each bucket SST measurement according to the hour in which it was collected.

Because our analysis is exclusively of differences in paired SST measurements, this removal of the climatological average is equivalent to removing the climatological difference between pairs of measurements according to difference in location and time.

b. *Estimating offsets using a linear-mixed-effect model*

A linear-mixed-effect (LME) model is used to quantify offsets among nations, as well as the significance of these offsets. The model is written

$$\delta\mathbf{T} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}_y\boldsymbol{\beta}_y + \mathbf{Z}_r\boldsymbol{\beta}_r + \boldsymbol{\beta}_\sigma, \tag{1}$$

where $\delta\mathbf{T}$ is a vector of SST differences between paired bucket measurements. As described in previous sections, climatological contributions to differences associated with spatial, seasonal, and diurnal effects are removed from $\delta\mathbf{T}$. The vector $\boldsymbol{\alpha}$ is the "fixed effect" term and represents the SST offset for each nation relative to all other nations. Such systematic offsets may arise for various reasons, including systematic differences in bucket characteristics or measurement protocols (Ashford 1948). We specify national-average offsets in temperature as fixed effects because our interest focuses

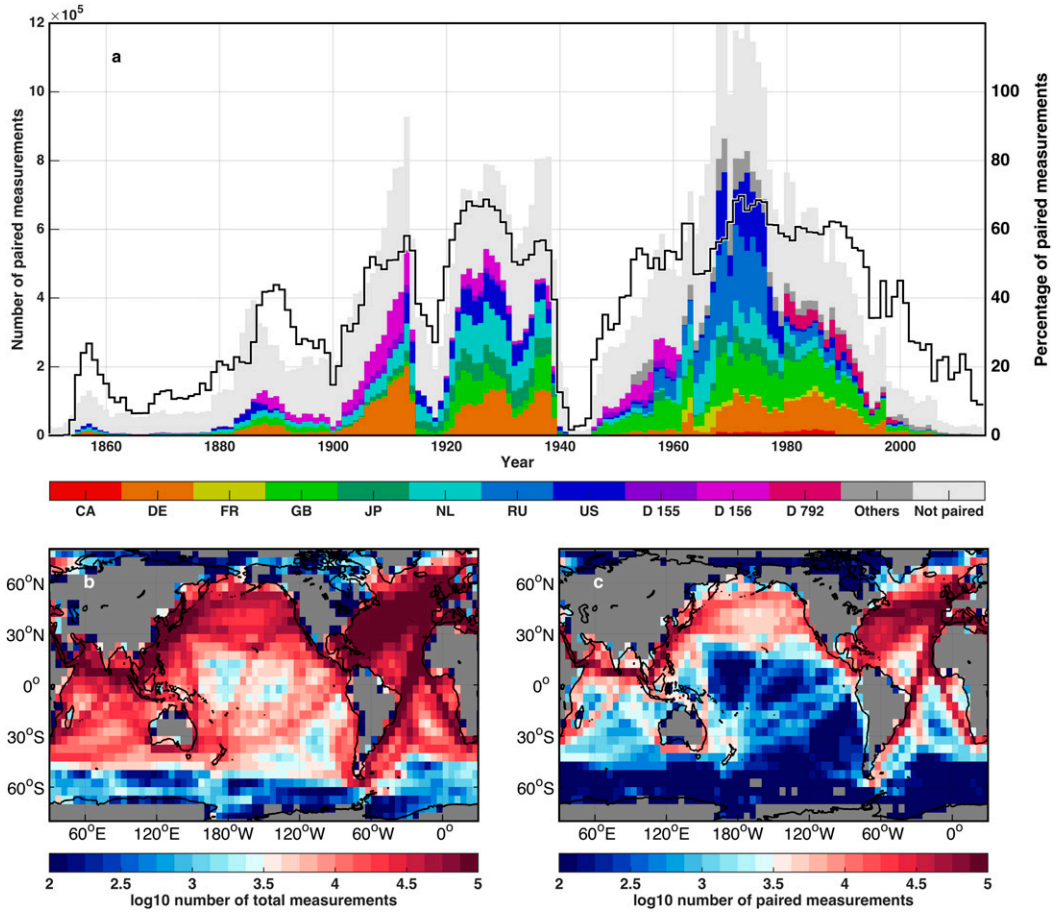

FIG. 1. Statistics of bucket SST measurements. (a) The number of paired measurements per year broken down according to collecting group. Table 1 lists the abbreviations, and "D" denotes deck. Countries that have less than 200 000 paired measurements total between 1850 and 2014 are listed as "Others" and plotted collectively (dark gray). The number of unpaired bucket measurements is plotted as additional data to that of the paired measurements for purposes of completeness (light gray). Also shown is the percentage of SSTs paired in each year (right y axis; black line). (b) A map of the number of bucket SST measurements between 1850 and 2014. (c) A map of the number of paired bucket SST measurements between 1850 and 2014.

on specific nations present in the ICOADS dataset, as opposed to a hypothetical population from which these nations may have been selected (Searle et al. 2009).

Other discrepancies between SST measurements are represented as random effects organized according to years $\boldsymbol{\beta}_y$ and regions $\boldsymbol{\beta}_r$. Regional effects may arise, for example, because bucket biases depend on local weather conditions, such as greater rates of latent cooling in windy regions (Folland and Parker 1995), or from differences in surface and subsurface ocean temperatures and the relative mixture in a bucket sample (Stevenson 1964). Yearly effects may arise from changes in buckets and measurement protocols (Folland and Parker 1995), ship size (Kent et al. 2013) and speed (Carella et al. 2017), or systematic changes in temperature structure (Cowtan et al. 2015) or winds (Vautard et al. 2010). The terms $\boldsymbol{\beta}_y$ and $\boldsymbol{\beta}_r$ are specified as normal with zero mean, an assumption intrinsic to the LME model. The terms $\mathbf{X}$, $\mathbf{Z}_y$, and $\mathbf{Z}_r$ are selection matrices that specify individual measurements that, respectively, belong to common groups, 5-yr blocks, and regions.

The remaining term in Eq. (1), $\boldsymbol{\beta}_\sigma$, represents contributions from SST variance and observational uncertainty. We initially focus on a simple case, in which differences between pairs of SST measurements are assumed independent and identically normally distributed. The variance of these differences σ_p^2 is the sum of observational variance σ_o^2 and the variance associated with displacement of measurements in space and time σ_c^2 after correcting for climatological offsets. Maximum likelihood estimates of the fixed and random effects, along with their variance and covariance, are obtained following Harville (1977). See the appendix for details.

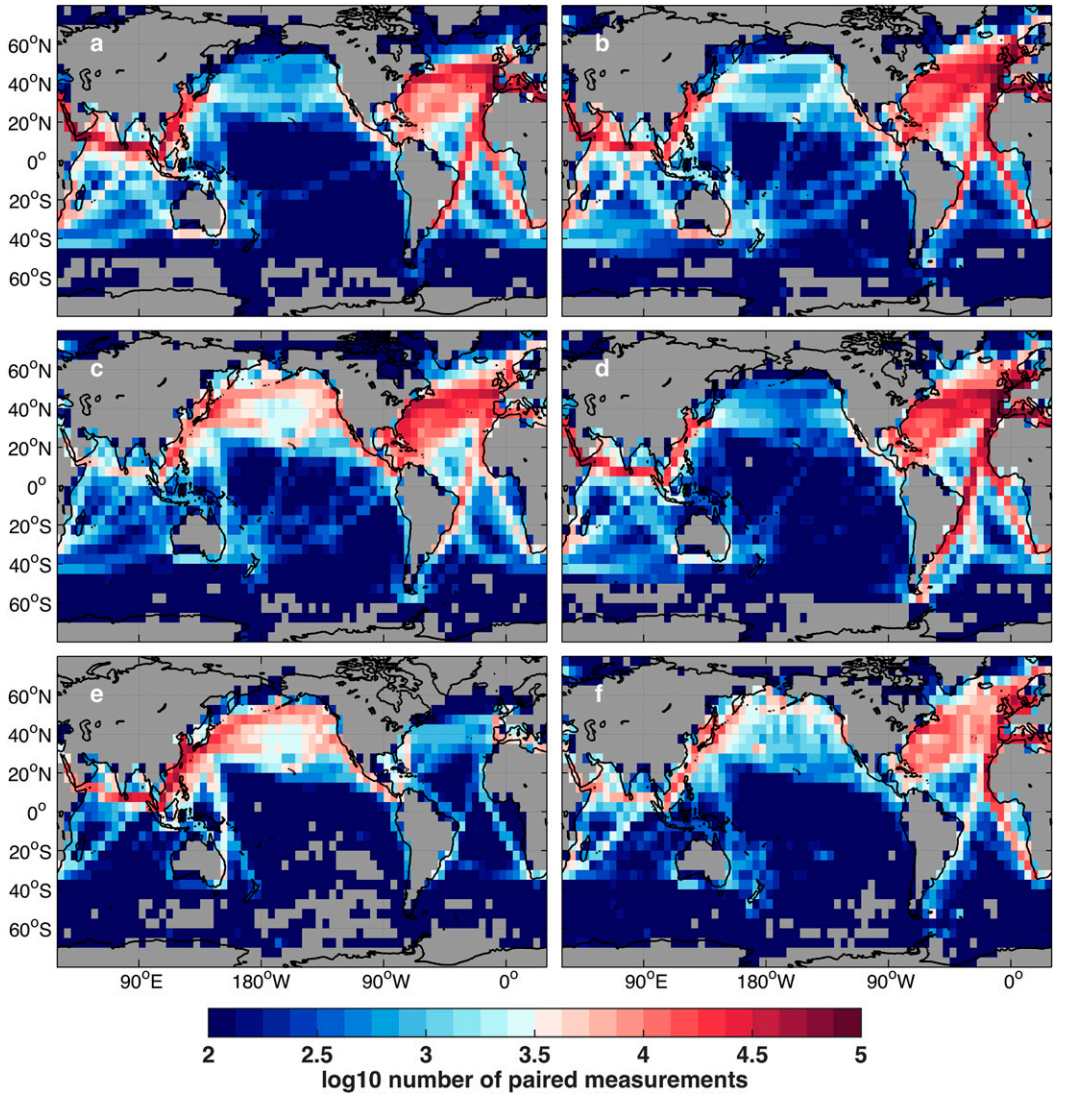

FIG. 2. Paired bucket SST measurements according to nation. Individual panels are for measurements from (a) the Netherlands, (b) the United Kingdom, (c) the United States, (d) Germany, (e) Japan, and (f) Russia. Of the 16.2 million pairs, 86% are in the Northern Hemisphere, and 1% are from regions poleward of 40°S. In addition, 58% of the paired observations are from the Atlantic, 17% from the Pacific, 15% from the Indian Ocean, 7% from the Mediterranean Sea, and 2% from the Arctic. Gray shading indicates regions having no paired measurements.

For purposes of computational efficiency, we average all SST differences associated with a given combination of nations that reside within a given 5-yr increment and region (Fig. 3). For example, all SST differences coming from the United States and United Kingdom in the subpolar North Atlantic between 1950 and 1954 are averaged. Averaging reduces the original 16.2 million SST pairs to 20 775 combinations. To account for this averaging, the diagonal elements in the variance matrix of $\boldsymbol{\beta}_\sigma$ are replaced by $\overline{\sigma}_p^2 = \sigma_p^2/n$, where n is the number of SST differences averaged together for a given combination. As discussed in section 5, our results are not overly sensitive to the degree of averaging or to more comprehensive representations of the error structure.

Our LME approach is similar to an analysis of variance (ANOVA; e.g., chapter 8 in Anderson 1962). Indeed, if $\boldsymbol{\beta}_y$ and $\boldsymbol{\beta}_r$ in Eq. (1) are treated as fixed effects, the LME and ANOVA methods are equivalent, and the problem can be solved using multiple linear regression techniques. LME is expected to outperform ANOVA, however, when random effects are present. Specifically, the LME framework better suits exploration of mean offsets in group-level SST estimates given the presence of yearly and regional variations in differences between bucket SST measurements.

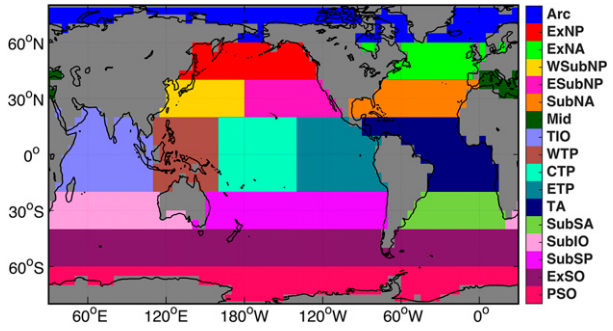

FIG. 3. Regional divisions. Regional offsets of individual groups are estimated at each of 17 subbasins: the Arctic (Arc), subpolar North Pacific (ExNP), subpolar North Atlantic (ExNA), western subtropical North Pacific (WSubNP), eastern subtropical North Pacific (ESubNP), subtropical North Atlantic (SubNA), Mediterranean Sea (Mid), tropical Indian Ocean (TIO), western tropical Pacific (WTP), central tropical Pacific (CTP), eastern tropical Pacific (ETP), tropical Atlantic (TA), subtropical South Atlantic (SubSA), subtropical Indian Ocean (SubIO), subtropical South Pacific (SubSP), subpolar Southern Ocean (SubSO), and polar Southern Ocean (PSO).

3. Nation-level results

The LME methodology is first used to examine whether there exist significant offsets in SST reports among nations (Fig. 4). Offsets are given by the fixed effects and range from −0.37° (Tanzania) to 0.61°C (deck 874, treated as a nation because additional information is missing). The weighted sum of fixed offsets are constrained to equal zero, which is equivalent to computing offsets relative to the average of all paired measurements (see the appendix for details). Out of 56 national groups, 24 groups have fixed effects that significantly ($p < 0.1$) differ from zero, and 15 groups have highly significant ($p < 0.01$) fixed effects. Bucket measurements are expected to be biased cold (e.g., Folland and Parker 1995; Rayner et al. 2006; Kennedy et al. 2011b), but these results indicate that the extent of this bias varies according to nation.

Between 1850 and 2014, 91% of all bucket SST measurements come from six nations (Germany, the United Kingdom, the United States, Japan, Russia, and the Netherlands) as well as deck 156, for which country-specific information is lacking. The spatial distribution of these major reporting groups is shown in Fig. 2. German, U.K., U.S., and Japanese measurements are, on average, consistent with an offset of zero, each having insignificant fixed effects within ±0.06°C. Measurements from the Netherlands, Russia, and deck 156, however, are associated with significant offsets. The Netherlands and deck 156 measurements are highly significantly offset cold by −0.10° and −0.13°C, respectively, and the Russian measurements are significantly offset warm by 0.10°C.

The spread of offsets across major collecting groups, from −0.13° to 0.10°C, can be interpreted in the context of a thermodynamic bucket model (Folland and Parker 1995). We consider a simple case involving only a canvas bucket model. Relative to a standard canvas bucket having an on-deck exposure time of 5 min, the observed range of systematic offsets could be obtained by average on-deck exposure times ranging from 4 to 6.5 min (Fig. 5), or if the bucket size ranges from 64% to 148% of its standard volume. Such ranges in exposure time and volume appear plausible given historical differences in collection methods (Ashford 1948). For example, documents indicate that U.S. (Wyman 1877) and Russian (Hydrometeoizdat Moscow 1941) observers were instructed to have on-deck exposure times of less than 1 min, whereas Japanese (Kobe Imperial Marine Observatory 1925) and Netherlands (KNMI 1937) observers were instructed to wait for an equilibrated reading. More comprehensive interpretation of offsets among these major collecting groups could be achieved by taking into account temporal changes in offsets as well as variations in the ratio of wooden versus canvas buckets, but these are beyond the scope of this current analysis.

Many nations that have large offsets are only present in ICOADS3.0 after 1960 and make small contributions to the total number of bucket measurements. For example, Spain contributes 0.02% and has a highly significant offset of 0.31°C, Malaysia contributes 0.02% and has a highly significant offset of 0.40°C, and Tanzania contributes 0.001% and has a significant offset of −0.37°C. There are, however, potentially important regional contributions. Canada contributes 0.8% of all bucket measurements and has a highly significant fixed effect offset of 0.19°C. This offset could have important implications for regional SSTs because Canadian measurements comprise 80% of all bucket measurements near its shores in the Arctic and North Atlantic since 1960, although the magnitude of the effect will also depend on how non-bucket SST measurements are incorporated in any given analysis. Another notable offset involves Brazil, whose measurements are highly significantly offset by −0.34°C and contribute 40% of bucket measurements near the east coast of South America.

It is also possible to perform pairwise comparisons between nation-level groups. In testing for significant differences between countries, accounting for both fixed effects and random yearly effects becomes important because the intervals covered by various nations typically only partially overlap and yearly random effects can rival the magnitude of the fixed effects. Performing significance tests for offsets between combinations of nations (Fig. 6), we find that among major collecting nations that the Netherlands and deck 156 are

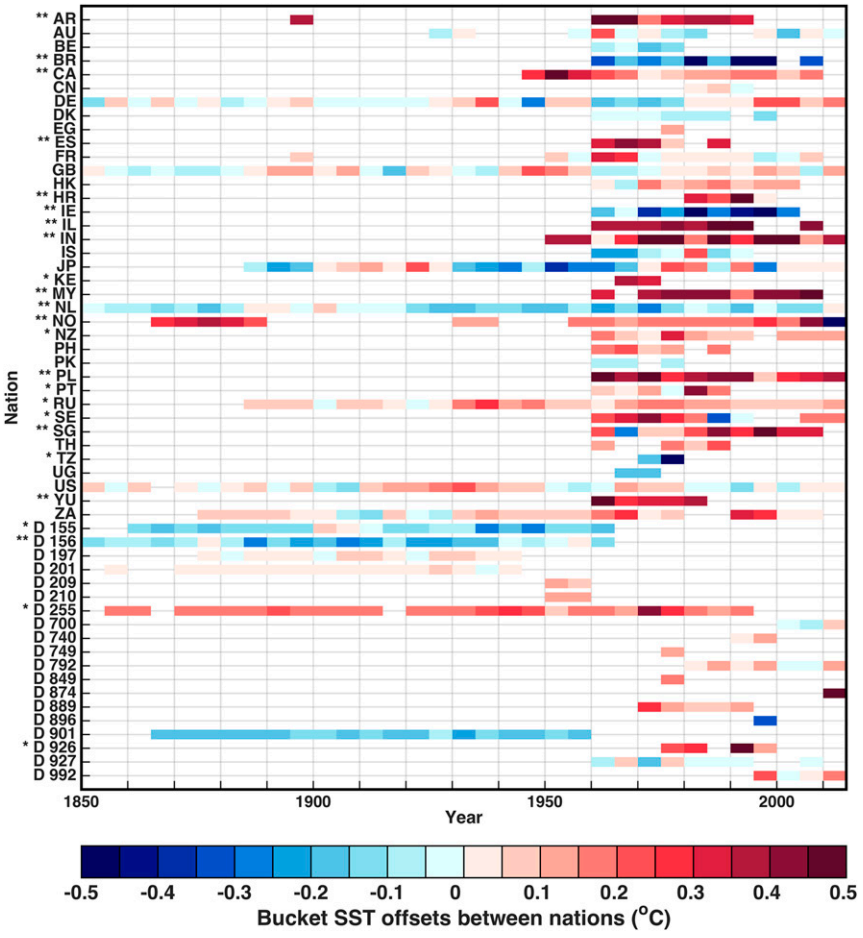

FIG. 4. Fixed plus yearly offsets of bucket SST measurements. Offsets are relative to the average across all paired measurements from all 56 groups. Groups are designated according to nation unless nation information is missing, whereupon groupings are by decks. Using an ordinary t test, 24 groups have fixed offsets that significantly differ from zero at $p < 0.1$ [indicated by one asterisk (*); see Table 1], and 15 groups have significant fixed offsets at $p < 0.01$ [indicated by two asterisks (**)].

significantly colder than German, U.K., and U.S. measurements, whereas Russian SST reports are, on average, significantly warmer than Japanese, Netherlands, and deck 156 measurements. Among the full 1180 combinations of nations, 463 combinations show significant differences at $p < 0.1$ (Fig. 6). Given that we are in a multiple-hypothesis testing regime, we also consider a Bonferroni correction (Bonferroni 1936) to the pairwise test. Pairs are considered significant if their p value is smaller than $a = 0.1/1180$, and 78 of the 1180 combinations remain significant under this stringent criteria (Fig. 6).

It is worth considering whether offsets in relative temperature could indicate that measurements are biased warm in an absolute sense. The absolute temperature correction indicated by the Folland and Parker (1995) bucket model, when driven by the ICOADS climatology and averaged across the globe, is −0.04°C for wooden buckets and −0.4°C for canvas buckets. The greatest cooling occurs in the tropics and western boundary currents, with cooling biases reaching as much as −0.9°C. Offsets among major collecting countries are within ±0.2°C, indicating that measurements from all major groups are, on average, still biased cold in an absolute sense for canvas buckets but may be biased warm for wooden buckets. Furthermore, positive offsets greater than 0.4°C occur for Israel between 1960 and 1995, Malaysia between 1960 and 2010, and Uruguay between 1960 and 1985 (see Table 1 and Fig. 4). A positive absolute bias could be a consequence of buckets that are well insulated but still subject to insolation (Hirahara et al. 2014). Another possibility is that some

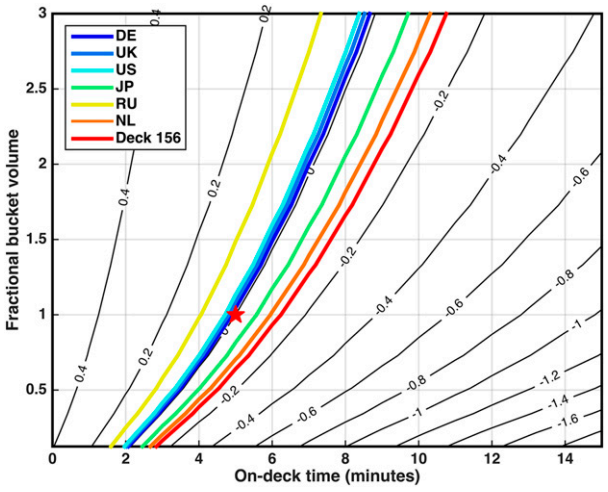

FIG. 5. Simulated SST offsets for a canvas bucket. Offsets are contoured as a function of on-deck time (x axis) and bucket volume (y axis). These offsets are relative to a reference bucket (red star) with a diameter of 16.3 cm and depth of 14 cm that is exposed on deck for 5 min. Other model parameters are listed in Table S1. Offsets estimated for major collecting nations are indicated by colored lines. The bucket model is run at individual 5° × 5° oceanic grids for each hour and month, and indicated offsets are the annual average over 60°S–60°N. OI-SST climatology between 1982 and 2014 is used to initialize the model, and the model is driven by the hourly resolved 2-m air temperature, 2-m dewpoint temperature, 10-m wind speed, and surface insolation climatology from ERA-Interim (Dee et al. 2011) between 1985 and 2014.

measurements are not actually from buckets, but instead from engine room intake, which is generally biased warm (Kennedy et al. 2011b). Further efforts to distinguish between these scenarios, possibly using the amplitude of diurnal variations (e.g., Carella et al. 2018), are warranted but beyond the scope of the present study.

4. Deck-level results

Having identified offsets among nations, it is also relevant to test whether significant offsets exist among decks from the same country. As the basic unit of ICOADS data management (Freeman et al. 2017), some divisions of decks appear purely for storage reasons, with individual ship tracks split between decks (Carella et al. 2017). Some other decks have essentially no metadata from which to infer whether decks are divided into physically meaningful groups. For example, deck 192 was captured by Allied troops during World War II in the form of punch cards and subsequently translated with assistance from the German Meteorological Service, but the original records that might have afforded more metadata were destroyed during the war (Air Weather Service and Weather Bureau 1955).

Metadata associated with some decks clearly indicate distinct sources. For example, deck 118 is the Japanese Kobe collection, whereas deck 187 comprises data from the Japanese Whaling Fleet. Conversely, some distinct decks have similar descriptions. For example, both deck 205 and 211 come from Scottish Fishery Cruisers. Those decks having similar descriptions are combined and treated as a single deck (Table 2). The refined grouping yields 139 decks distributed across 37 nations for a total of 158 distinct groups (Table 1). Measurements without country information are, again, arranged only by decks. Distinguishing groups by decks and nations increases the number of measurements that are paired from 16.2 million in the nation-level analysis to 17.6 million. Averaging according to the same domains as for the nation-only analysis gives 43 813 averaged pairs.

Extending from the nation-only analysis [Eq. (1)], we now assign a hierarchical structure to nations and decks:

$$\begin{aligned} \delta\mathbf{T} &= \delta\mathbf{T}^N + \delta\mathbf{T}^D + \boldsymbol{\beta}_\sigma, \\ \delta\mathbf{T}^N &= \mathbf{X}^N\boldsymbol{\alpha}^N + \mathbf{Z}_y^N\boldsymbol{\beta}_y^N + \mathbf{Z}_r^N\boldsymbol{\beta}_r^N, \\ \delta\mathbf{T}^D &= \mathbf{X}^D\boldsymbol{\alpha}^D + \mathbf{Z}_y^D\boldsymbol{\beta}_y^D + \mathbf{Z}_r^D\boldsymbol{\beta}_r^D, \end{aligned} \tag{2}$$

where superscripts N and D denote nation and deck. Equation (2) is equivalent to Eq. (1) at the national level, but at the deck level additional offsets are allowed. Deck-level fixed effects are constrained to sum to zero for each nation, where the sum is weighted by the number of paired measurements associated with individual decks. No deck-level variability is permitted for nations associated with only one deck. Tests for differences between decks make use of only the deck-level results from the LME and are thus not influenced by nation-level offsets. Significant ($p < 0.1$) deck-level offsets are found for every major collecting country (Fig. 7).

The Netherlands has among the most consistent interdeck SST measurement behavior (Figs. 7a and 8a), despite being offset cold relative to other nations from 1850 to 2014. This consistency suggests that Netherlands measurements could be treated as a single supergroup. The one exception is deck 732, which contains Netherlands measurements that are colder than other Netherlands decks by more than 0.5°C (Fig. 8a). Deck 732 is a Russian Marine Meteorological Dataset for which problematic position reports were identified elsewhere (Kennedy et al. 2011b), leading to exclusion of some of its contents in ICOADS3.0 (Freeman et al. 2017). There are only 499 paired Netherlands measurements in deck 732 in the 1960s (Fig. 8a), and their outlying behavior suggests that position errors or some other data artifact are present. Deck 732 is also an outlier for 1201 German measurements in the 1950s that are

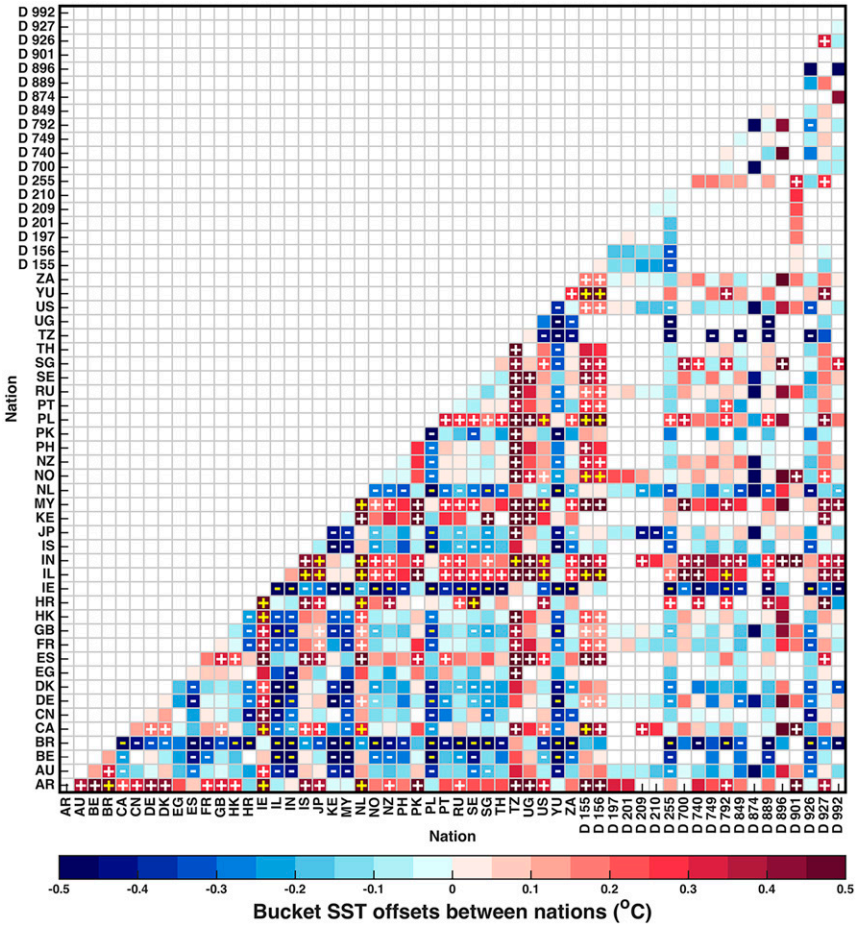

FIG. 6. Offsets between pairs of nation-level groups. Differences between countries on the y axis and countries on the x axis are shown in color. Significance as per Eq. (A5) at $p < 0.1$ is indicated by a plus sign (+) for positive offsets and by a minus sign (−) for negative offsets. Markers in yellow indicate significant offsets between nations after applying the Bonferroni correction for multiple hypothesis testing. White boxes indicate that two groups have no overlapping periods.

more than 0.5°C warmer than all other German decks (Fig. 7d). Measurements from the U.K. are also generally consistent across its 20 decks, except for data from Scottish Fishery Cruisers (deck 205) in the 1940s and 1950s and U.K. Royal Navy Ships during the First World War (deck 249), which are significantly colder than other U.K. decks (Figs. 7b and 8b).

U.S. decks are found to have considerable offsets (Figs. 7c and 8c). Before the 1940s, measurements from U.S. Arctic logbooks (deck 710) are significantly colder than all other U.S. decks. Starting in the 1950s, measurements from the International Maritime Meteorological Data (deck 926) and World Ocean Database (deck 780) are significantly colder than those from International Marine archives (deck 927). Furthermore, from the 1920s to the 1940s, U.S. Naval measurements (deck 281) are significantly warmer than those from Merchant Marine ships (decks 705, 706, and 707). It may be the case that U.S. Naval measurements do not actually come from buckets, as is assumed to be the case before 1941 and is indicated in ICOADS or WMO47 metadata afterward, but instead come from engine room intake measurements, as suggested previously (Kennedy et al. 2011b; Carella et al. 2018). There are also differences between U.S. Merchant Marine decks, with deck 705 being significantly colder than deck 706 and 707.

Rayner et al. (2006) found that U.S. Merchant Marine deck 705 was generally unbiased, whereas Merchant Marine deck 706 was offset cold in the Pacific, and deck 707 was offset cold in both the Pacific and Atlantic relative to Met Office Historical Sea Surface Temperatures (MOHSST). The discrepancy between our results and those of Rayner et al. (2006) may arise because of

TABLE 2. Decks that are combined in the deck-level analysis. Decks in each row are combined because they have similar descriptions and are assumed to have similar bias structures. Boldface numbers indicate the name used when referring to the combined decks elsewhere.

Description	Decks
British Navy (HM) Ships	**204**, 229, 239
Deutsche Seewarte Marine	**192**, 196
Great Britain Marine	**184**, 194, 902
Japanese Kobe Collection	**118**, 119, 762
Japanese Whaling Ship	**187**, 762
Netherlands Marines	189, **193**
Scottish Fishery Cruiser, MARIDS	**205**, 211
U.K. Met Office Selected Ships	**203**, 207, 209, 213, 223, 227, 233
U.S. Navy	**281**, 195, 555, 709
U.S. NCEP: ship data	792, **892**
International Marine (U.S.- or foreign-keyed ship data)	128, 254, **927**

updates to the data contained within ICOADS relative to MOHSST. A comprehensive set of codes for processing ICOADS data and testing for offsets among nations and decks are made available with this publication and can be downloaded from https://github.com/duochanatharvard/SST-LME-compact. This code should facilitate checking how subsequent changes to the ICOADS dataset influence relative offsets.

5. Sensitivity of results to changes in model formulation

Several simplifying assumptions were made in implementing Eq. (1) involving whether errors are homogeneous, independent, and normally distributed. The sensitivity of our results to these assumptions is explored in this section first through introducing a more comprehensive error model and then examining the sensitivity of results to various changes in model structure and data screening procedures.

a. A more complete error model

The variance of averages of SST pairs is expressed as the sum of three contributions:

$$\overline{\sigma}_p^2 = \frac{2\sigma_o^2}{n} + \frac{\sum \sigma_c^2(i)}{n^2} + \frac{\sigma_s^2}{s_1} + \frac{\sigma_s^2}{s_2}. \tag{3}$$

The first term on the right-hand side represents uncorrelated observational error σ_o^2 and is unchanged from our original error model. The second term represents the variance of differences in physical SSTs σ_c^2, where a technique to account for spatial and temporal differences in SST variance and covariance between measurements is described in section 5a(1). The final contribution comes from ship-level biases σ_s^2, and a technique for estimating the number of ships contained within each group s_1 and s_2 is detailed in section 5a(2).

1) VARIANCE OF DIFFERENCES IN PHYSICAL SSTS

The variance of the difference between two SST measurements is

$$\sigma_c^2(i) = \sigma_T^2(p) + \sigma_T^2(q) - 2C_T(p,q), \tag{4}$$

where σ_T^2 is the SST variance associated either with measurement p or q. Also, $C_T(p, q)$ is the covariance between the two measurements, which we assume to decay exponentially:

$$C_T(p,q) = \sigma_T(p)\sigma_T(q)\exp[-\phi'(p,q)\gamma_\phi - \theta'(p,q)\gamma_\theta - t'(p,q)\gamma_t]. \tag{5}$$

Displacement in longitude $\phi'(p, q)$, latitude $\theta'(p, q)$, and time $t'(p, q)$ is associated with exponential decay of covariance at a rate governed by the corresponding γ terms. Values for the σ_T and γ terms are estimated as a function of location and calendar month using OI-SST data between 1982 and 2014 (Reynolds et al. 2007).

SST variance is computed at 0.25° resolution on each calendar day using year-to-year values, with results then averaged to monthly resolution and the square root taken to provide estimates of σ_T. As anticipated, σ_T is larger near western boundary currents, across the Antarctic Circumpolar Current, and over the eastern equatorial Pacific (Fig. 9a). Seasonally, the variance over the eastern Pacific is higher during boreal winter, which is partly related to stronger variations associated with El Niño–Southern Oscillation. Interestingly, midlatitude SST variance is lower in winter than summer for each hemisphere (Fig. 9b), possibly because the climatological meridional SST gradient is lower in winter over extratropical open oceans (Schneider et al. 2015). Examination of data from 17 nearly continuously monitored buoys (Hervey 2014) confirms that daily average midlatitude variance is generally lower during winter.

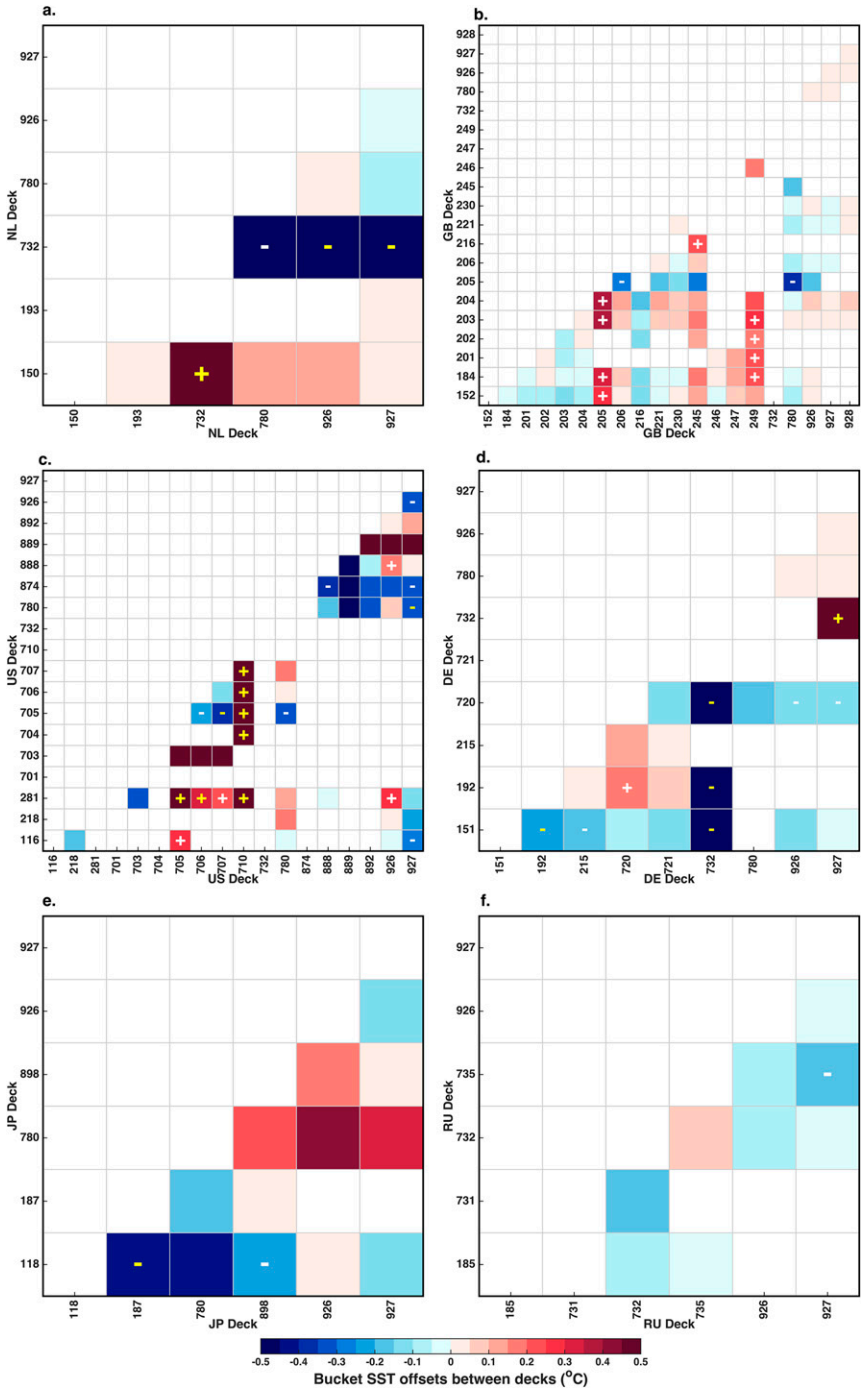

FIG. 7. Offsets between pairs of decks within the same nation. Individual panels are similar to Fig. 6, but are for (a) the Netherlands, (b) the United Kingdom, (c) the United States, (d) Germany, (e) Japan, and (f) Russia. We find significant offsets at $p < 0.1$ in all six major collecting countries.

Values for the γ terms in Eq. (5) are estimated using sample correlations between each SST estimate in OI-SST and its neighbors within a vicinity of 5° in latitude and longitude and 5 days in time across years from 1982 to 2014. Each γ value is estimated at daily resolution and then averaged to monthly values. Daily SST covariance generally decays at similar rates zonally (Fig. 9c) and meridionally (Fig. 9d), with more rapid declines near

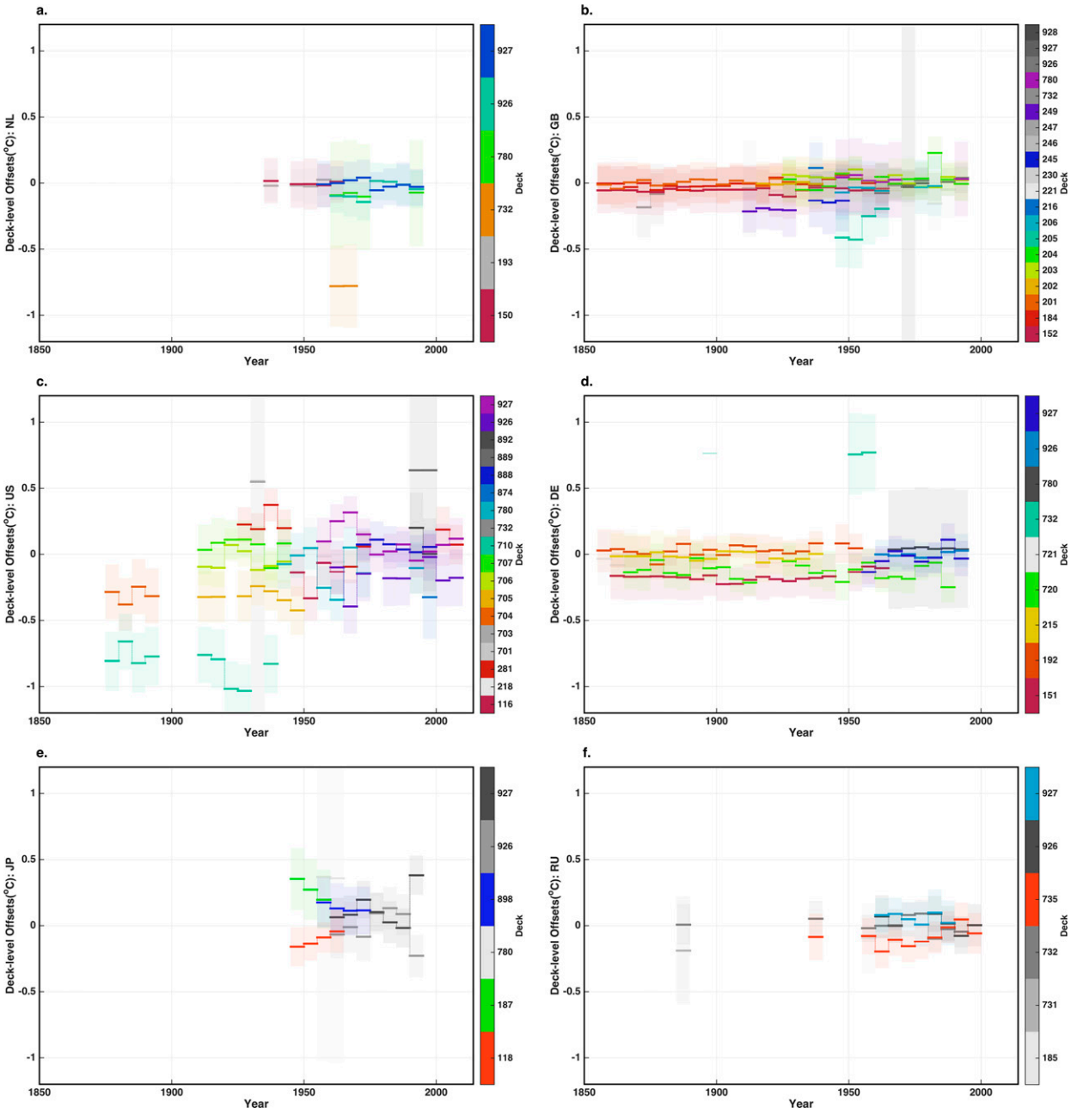

FIG. 8. Fixed and yearly offsets of decks for (a) the Netherlands, (b) the United Kingdom, (c) the United States, (d) Germany, (e) Japan, and (f) Russia. The 90% confidence interval of each offset is shown in shading. Decks that significantly differ from another deck from the same country at $p < 0.1$ are colored, whereas others are gray. Offsets of individual decks are relative to the overall offset for each nation.

western boundary currents and across the Antarctic Circumpolar Current. To account for the fact that decorrelation rates are estimated using OI-SST daily averages (Fig. 9e), whereas buckets obtain seawater samples within minutes, we examine the decorrelation of hourly and daily data at the 17 buoys noted above. Hourly SST decorrelation rates average 55% larger than the corresponding daily SSTs (see Table S2 in the online supplemental material), and we increase values of γ_t by this percentage.

2) VARIANCE OF SHIP-LEVEL BIASES

A second potentially important error structure involves the correlation associated with measurements taken from the same ship. Ship-level biases were previously estimated by comparing bucket and engine room

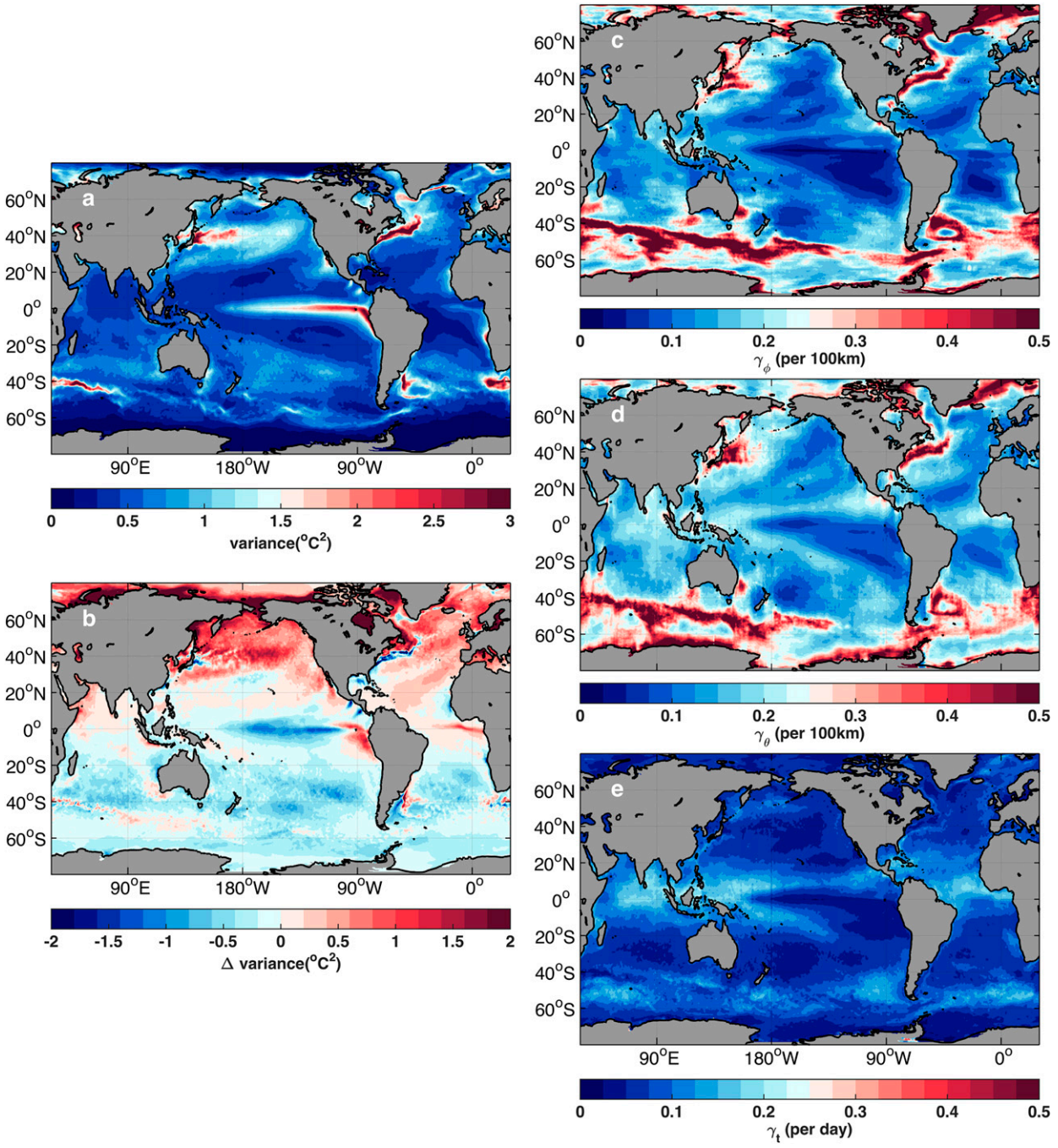

FIG. 9. Spatial distribution of SST variance and decorrelation rates. (a) The variance of daily average SST σ_T^2, averaged over the entire year, is largest along western boundary currents, the eastern equatorial Pacific, and parts of the Southern Ocean. (b) The seasonal difference in SST variance (JJA − DJF) shows that the summer hemisphere tends to have greater daily average SST variance. Also shown are decorrelation rates in (c) longitude γ_ϕ, (d) latitude γ_θ, and (e) time γ_t. Statistics are based on daily OI-SST between 1982 and 2014.

intake SST measurements to numerical weather prediction results (Kent and Berry 2008) and satellite observations (Kennedy et al. 2012; Xu and Ignatov 2010) using data taken since the 1980s. These studies indicate roughly equal contributions of variance from independent measurement error and ship-level biases [see Table 4 in Kennedy (2014)]. Ship-level biases were estimated to contribute more than half of the uncertainty of monthly estimates of global-mean SST (Kennedy et al. 2011a). If the ships associated with individual measurements were known, Eq. (3) could be directly

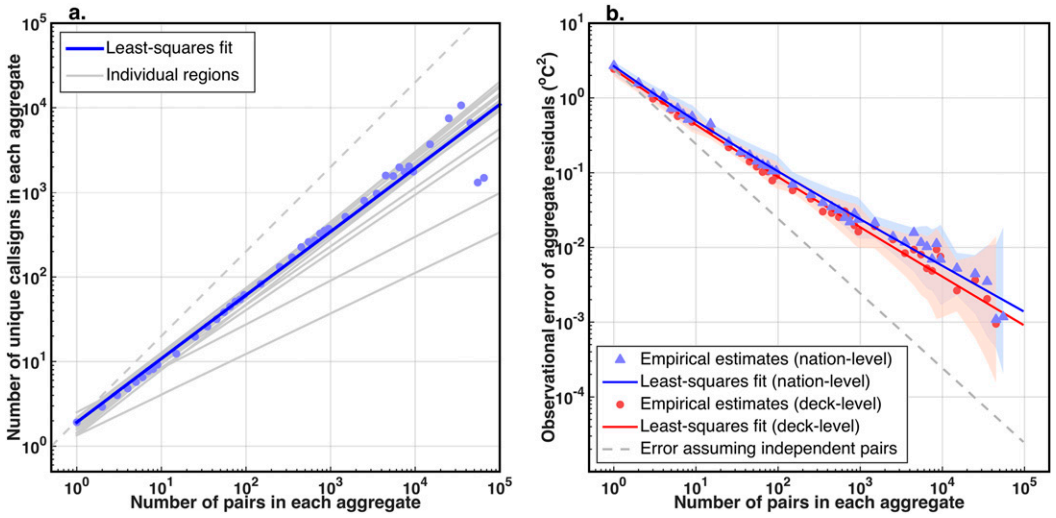

FIG. 10. Ship-level statistics. (a) The number of unique ships scales with the number of pairs of data to a power of ~0.75 (blue dots and blue line). Regional variations in this power-law scaling range from 0.48 to 0.83 (solid gray lines), where pairs are averaged within the 17 regions shown in Fig. 3. All scaling relationships indicate that the number of ships increase at a rate that is smaller than proportional to the number of observations (dashed gray line). (b) Variance scales with the number of pairs of data to a power of ~−0.61 for nation-level analysis (blue dots and blue line) and −0.65 for deck-level analysis (red dots and red line), thus decreasing more slowly than expected if measurements are independent (dashed gray line). Power laws are estimated using a least squares fit in log space, and 95% confidence intervals for each bin (shading) are estimated from the residual of the fit and numbers of groups in each bin.

estimated, but among the 17.6 million bucket SST pairs used in this study, only 6.0 million pairs are associated with unique call signs for both measurements.

Absent an absolute count of ships, we instead use those 6.0 million SST pairs for which ship information is present to estimate a scaling relationship. SST pairs are first binned by nation, region, and 5-yr increment. The number of call signs in each bin is counted and then averaged over bins with similar sample sizes. A least squares fit in log space yields that the number of ships s scales approximately exponentially with the number of measurements n^x, where x is estimated to be 0.75 ± 0.02 (Fig. 10a). Scaling is generally consistent when independently estimated across regions (Fig. 3). Note that this estimation approach leads to equivalence between s_1 and s_2 in Eq. (3).

We also estimate how measurement errors associated with SST differences scale with numbers of pairs. The nation-level and deck-level offsets estimated in sections 3 and 4 are first removed, as are the climatological SST variance estimates σ_c^2. Residual SST differences that are averaged over national, yearly, and regional groups are then used to compute variance according to number of pairs in a bin. For small sample sizes, the measurement variance of averaged pairs are close to those estimated assuming independence, but measurement variance is 30 times larger for sample sizes of 20 000 (Fig. 10b). These results indicate that ship-level biases dominate the errors associated with average offsets in major collecting groups.

Equation (3) can be rewritten using the approximation that number of ships scales with the number of observations raised to an exponent,

$$\log\left[\overline{\sigma}_p^2 - \frac{2\sigma_o^2}{n} - \frac{\sum\sigma_c^2(i)}{n^2}\right] = \log(2\sigma_s^2) - x\log(n), \quad (6)$$

where the natural log is taken and terms rearranged to illustrate that estimates can be obtained using a linear least squares approach. Fits are weighted according to the number of pairs averaged together for each sample variance.

For the nation-level analysis the variance of uncorrelated independent measurement errors σ_o^2 is estimated to equal 0.55° ± 0.15°C^2 (one standard deviation); ship-level biases σ_s^2 to equal 0.78° ± 0.12°C^2; and the exponent x to equal 0.61 ± 0.02. For the deck-level analysis σ_o^2 is 0.42° ± 0.15°C^2, σ_s^2 is 0.80° ± 0.12°C^2, and x is 0.65 ± 0.02. A higher exponent is estimated directly from ship numbers than from the scaling of paired SST variance, possibly because of dependencies across ships. Another potential contribution is from incorrectly identifying the number of ships present, which are assigned based on multiple different sources in ICOADS metadata (Carella et al. 2017).

TABLE 3. Estimates of measurement error of ship-based SST for studies in which measurement error is decomposed into an independent measurement error σ_o^2 and a ship-level systematic error σ_s^2.

Study	σ_o^2 (°C^2)	σ_s^2 (°C^2)
Nation level (this study)	0.55 ± 0.15	0.78 ± 0.12
Deck level (this study)	0.42 ± 0.15	0.80 ± 0.12
Kent and Berry (2008)	0.49	0.64
Kennedy et al. (2012)	0.57	0.50
Brasnett (2008)	1.35	0.48
Xu and Ignatov (2010)	0.65	0.28

3) PARSING SST MEASUREMENT VARIANCE

Our observational error estimates can be compared with estimates from previous studies (as summarized in Table 3). Using a variogram technique, Kent and Challenor (2006) estimated measurement error of 1.21°C^2 (68% confidence interval: 0.64°–1.96°C^2), which is essentially equal to the combined measurement error of 1.21° ± 0.04°C^2 that we obtain for the deck-level analysis. In a later analysis, Kent and Berry (2008) compared ICOADS measurements with numerical weather prediction results to estimate 0.49°C^2 for σ_o^2 and 0.64°C^2 for σ_s^2. Kennedy et al. (2012) compared ICOADS data with along-track scanning radiometer SST retrievals and obtained 0.57°C^2 for σ_o^2 and 0.50°C^2 for σ_s^2. Brasnett (2008) compared ship measurements with an SST analysis that incorporated both in situ and satellite retrievals and estimated 1.35°C^2 for σ_o^2 and 0.48°C^2 for σ_s^2. Xu and Ignatov (2010) compared ship data with multisensor satellite SST fields and got 0.65°C^2 for σ_o^2 and 0.28°C^2 for σ_s^2. Our ship-level bias estimates are generally larger than previous estimates, possibly because our estimates implicitly account for covariance across ships as opposed to previous studies that assumed biases to be independent across ships.

For purposes of better identifying where opportunities exist to improve estimation of SSTs, it is also useful to consider the various contributions to SST uncertainty in somewhat greater detail. The original SST pairs have an average variance of 3.15°C^2, where 23% of the variance is related to climatological differences associated primarily with spatial offsets and, to a lesser extent, seasonal offsets. For the remaining 2.45°C^2 of variance, 98% is attributable to independent or ship-level measurement errors and physical SST variability. Though systematic offsets account for only the other 2% of the variance, this percentage increases with averaging. Offsets have an expected variance that exceeds those from the combination of independent measurement error and SST variability once 250 paired differences are averaged. The number of pairs typically average according to location, epoch, and group is 820 for the nation-level analysis and 390 for the deck-level analysis. The variance contributed by intergroup offsets can be further decomposed. For the deck-level analysis, which we consider to be the more complete analysis, 40% of the variance in offsets comes from fixed offsets, 40% from yearly effects, and 20% from regional effects.

4) SENSITIVITY TO UPDATED ERROR MODEL

We refit the LME model taking into account the heterogeneous structure of SST variance and covariance between pairs due to ship-level biases. Specifically, the averages of SST differences are weighted by the inverse of the variance obtained from Eq. (3). The updated error structure does not influence the central estimate of the averaged SST difference but leads to greater variance, especially for bins containing large numbers of pairs. Greater variance is also assigned to bins in regions and from seasons that are intrinsically more variable or have SST pairs that are more separated. On account of our error model being empirically based and itself subject to uncertainty, we also introduce a scaling factor multiplying the total error $k\overline{\sigma}_p^2$. The maximum-likelihood solution obtained from the LME indicates that k equals 0.98 for the nation-level analysis and 0.97 for the deck-level analysis, suggesting consistency between the LME solution and Eq. (3).

The fixed and random effects estimated using our updated error model are generally consistent with results obtained assuming that measurement errors are independent and identically distributed. The median absolute change in fixed and yearly offsets is 0.03°C for nations in 5-yr increments with less than 100 measurements and is only 0.01°C for those having more than 10 000 measurements. Similarly, the median absolute change in the uncertainty of offsets is 0.007°C for nations in 5-yr increments with less than 100 measurements and is only 0.003°C for those having more than 10 000 measurements. Accordingly, the median absolute change in p values of fixed effects is small at only 0.027 for the nation-level analysis and 0.033 for the deck-level analysis, although in some cases marginal p values are flipped between significant or insignificant values (Table 4).

b. *Sensitivity to other assumptions*

A number of other methodological choices and simplifying assumptions have been made whose implications are not necessarily obvious, and these are addressed through a series of sensitivity tests. Results are found to be robust with respect to degree of averaging, assumptions regarding normality and homogeneity of errors, and the influence of diurnal variability. Details of changes in results to each of these factors are summarized in Table 4. All comparisons are relative to the model formulation presented in section 5a.

TABLE 4. Sensitivity to alternate LME configurations. The columns show the seven different configurations that are considered: (column 1) assuming independent and identically normally distributed (i.i.d.) pairs, (column 2) accounting for spatially heterogeneous SST variance and correlation across pairs, (column 3) outlier pairs are trimmed at three sigmas for each 5° × 5° box, (column 4) using spatially heterogeneous measurement errors from Kent and Challenor (2006), (column 5) pairs are split into two equal parts and averaged separately, (column 6) pairs are averaged globally, and (column 7) pairs are binned by 10-yr increments. Changes in significance occur because some p values are marginal at the 90% or 99% level. All changes are relative to the baseline analysis given in section 5a.

	i.i.d. pairs	Updated errors	Trim	KC06	Less average	No regional effects	Decadal average
Significant fixed effects (90% level, out of 56)	24	24	25	25	23	25	21
Significant fixed effects (99% level, out of 56)	15	16	17	16	14	15	13
Fixed offsets between RU and deck 156 (°C)	0.22	0.20	0.20	0.21	0.19	0.25	0.22
Min fixed effect among nations (°C; nation level)	−0.37	−0.43	−0.39	−0.43	−0.45	−0.45	−0.41
Max fixed effect among nations (°C; nation level)	0.61	0.61	0.62	0.61	0.66	0.63	0.61
Median change of p values (nation level)	0.027	—	0.018	0.006	0.016	0.038	0.028
Median change of p values (deck level)	0.033	—	0.022	0.012	0.012	0.046	0.039
Nations with outlier decks (90% level; out of 31)	14	15	14	15	15	12	12

1) AVERAGING

To examine the implication of averaging pairs of observations, we rerun the analysis using different numbers of averages. Results are not qualitatively sensitive to doubling the number of averages by randomly separating bins into two groups (Table 4). Increasing the degree of averaging through removing regional effects and conducting the analysis at a global level also leads to similarly small sensitivities, as does increasing the temporal binning from 5 to 10 years.

Results are presumably robust to the degree of aggregation of the dataset because averages are over random errors that are independent of national- or deck-level offsets (Breckling et al. 1994; King et al. 2004). Our analysis does not account for covariance across averages as would arise from a ship traversing across regions, but given the small changes found when accounting for covariance (section 5a), we do not expect this source of covariance to be consequential. A more comprehensive analysis using estimates of ship tracks (Carella et al. 2017) may nonetheless be useful.

2) NONNORMALITY AND INHOMOGENEITY

As is standard for a LME model, the random effect and error terms in Eq. (4) are assumed to follow normal distributions with zero mean. ICOADS data have excess kurtosis (Kennedy et al. 2012), however, with the paired SST differences that we consider having a sample kurtosis of 7.5, as compared with a value of 3 for a normal distribution. The fact that the variance of paired SST differences is spatially heterogeneous (Kent and Challenor 2006) explains part of the excess kurtosis. If kurtosis is instead estimated at the level of a 5° × 5° grids, kurtosis averaged across the grid drops to 4.6. Some contributions to kurtosis were thus already addressed when considering inhomogeneity of SST variance in the more complete error model presented in section 5a. Segmenting SST distributions according to seasons would only marginally further decreases kurtosis to 4.5, and segmenting according to nations would have negligible effect.

Some of the excess kurtosis in SST observations may be intrinsic to the measurement as a result of bucket cooling depending on environmental parameters multiplied by measurement time. If cooling rates and measurement time are independently normally distributed, such multiplicative uncertainty tends to generate excess kurtosis (Oliveira et al. 2016). To check the effects of nonnormality, we rerun the analysis after trimming pairs of SST differences that are more than three sample standard deviations away from the mean within their respective 5° × 5° boxes. Although still not normal, the trimmed dataset has a kurtosis of 3.5 and permits for assessment of whether outliers influence the results. Our results are robust to the deviations from normality found in the paired SSTs (Table 4).

Although we account for regional differences in SST variance and different numbers of observations [Eqs. (4)

and (5)], measurement errors σ_o^2 and σ_s^2 are assumed homogeneous. Kent and Challenor (2006) estimated a spatial pattern of the sample variance of SST measurements. To examine the implications of regional variability in measurement variance, as opposed to only SST variance, we prescribe the error variance estimated by Kent and Challenor (2006) at 30° × 30° resolution. When estimating the uncertainty of averaged pairs [Eq. (6)], measurement variance is partitioned into contributions from independent measurement errors σ_o^2 and correlated ship-level biases σ_s^2 using the ratios given in Table 3 for national- and deck-level analyses. Results are again insensitive to the inclusion of heterogeneous measurement errors (Table 4).

3) DIURNAL VARIABILITY

Finally, there are concerns that the diurnal cycle of SST could compromise intercomparisons (Kent et al. 2010; Kennedy 2014). By way of a cautionary analogy, Kennedy et al. (2007) concluded that 13% of the observed trend in tropical tropospheric temperatures could be attributed to the slow drift of satellite orbits that altered the sampling of the diurnal cycle. We have factored out a climatological diurnal estimate based on buoy and drifter data, as described in section 2a. Offset estimates are not qualitatively sensitive to whether or not climatological diurnal variability is removed. This result can be understood in that removing the climatology of diurnal offsets reduces the variance of raw SST differences by less than 0.5%. Removing spatial and seasonal dependence, in contrast, decreases the variance of SST differences by 23%. Furthermore, the diurnal cycle is not expected to introduce offsets because the hour offset between paired SSTs is not systematic.

As one further test, we note that estimating the diurnal cycle directly from bucket measurements leads to systematically larger diurnal amplitude estimates (Carella et al. 2018). Factoring out these larger estimates of diurnal variability results in similarly small changes to intergroup offset estimates. Thus, we conclude that offset estimates are not qualitatively sensitive to whether or not climatological diurnal variability is removed.

6. Further discussion and conclusions

It is useful to explicitly consider whether analyses of SST offsets ought to make distinctions among decks of data in addition to national-level offsets. There exists a trade-off between possibly making unwarranted divisions among data versus failing to distinguish among groups having real offsets. Increasing the number of divisions of the data according to both nations and decks allows the LME model to explain 18% more variance of the paired SST differences. In addition, the interquartile range of fitted offsets for individual pairs widens from a range of −0.20° to 0.07°C to one of −0.22° to 0.09°C. These increases are expected as a result of more degrees of freedom afforded to the model fit. Nevertheless, 14 out of 31 nations show decks with significant departures at $p < 0.1$. Furthermore, Germany deck 732, U.S. deck 705, and U.S. deck 710 have highly significant deck-level offsets at the $p < 0.01$ level. This frequency of significant outliers exceeds the expected false-positive rate.

If decks that have actual offsets are instead combined into one group, the unresolved deck-level offsets are treated as independent measurement error, leading to increased estimates of uncertainty and smaller amplitude offsets. For example, if we combine Japanese Kobe Collection deck 118 and Japanese Whaling Ships deck 187, despite the latter appearing significantly warmer, the overall estimate of the cooling of Japanese temperature offsets during 1920–50 is around 20% weaker. Conversely, if homogeneous measurements are arbitrarily divided into distinct decks, their offset estimates remain consistent with one another. For example, if Japan decks 118, 119, and 762, which are all described as the Kobe Collection, are separated into distinct decks and the analysis rerun, the three decks show a similar pattern to the case in which they are merged. Similar results hold for grouping and ungrouping other decks, including U.S. measurements from International Marine decks 128, 254, and 927.

Failure to identify systematic offsets among decks thus appears the greater liability. For this reason, we consider our deck-level analysis to be both a more complete and likely a more accurate estimate of offsets associated with groups of ICOADS3.0 bucket data. Furthermore, we recommend that analyses accounting for systematic offsets should typically admit for nation- and deck-level variations. These results have implications for estimates of SST trends that will be taken up elsewhere in further detail.

A limitation of our analysis is that offset estimates are relative to an unknown mean bucket bias. The mean bias of bucket temperatures has been estimated using thermodynamic models of canvas and wooden buckets and the relative fraction of these two types of buckets through time (Folland and Parker 1995). Global-average bucket biases are estimated to range from −0.1° to −0.4°C between 1850 and 1940 (Kennedy et al. 2011b). Parametric uncertainty associated with the fraction of canvas versus wooden buckets range from as much as 0.2°C warmer if all measurements are from wooden buckets to −0.3°C cooler if only canvas buckets are used (Kennedy et al. 2011b). Accounting for other uncertainties—such as those associated with the cooling rate of different models

of wooden and canvas buckets and on-deck time during which a measurement is taken—would lead to a broader range of uncertainties. The uncertainty in mean bucket corrections is, therefore, comparable to the range of offsets across major collecting nations.

An alternative to solving for global bucket bias would be to select a single nation as a baseline against which all other countries are compared. In this framework, absolute SST would only need to be estimated for a single nation. One obvious candidate is the United Kingdom, which contributes 17% of all bucket observations in ICOADS3.0. U.K. observations are generally consistent across decks except for those from Scottish Fisheries and the Royal Navy. Another candidate is the Netherlands, whose observations are the most consistent across decks of any nation, although the national average is colder than the ICOADS3.0 average. Of course, multiple reference nations could be used as a means of checking the accuracy of any individual estimate.

A complimentary strategy to decreasing offsets among nations would be to better constrain variations in bucket parameters and measurement practices in order to better determine absolute temperatures. Although it is unclear whether such a level of detail could be obtained, measurement models might be developed according to nation and deck, if not also for individual ships and years. If offsets occur because of different bucket attributes or measurement techniques, they may be variously accentuated in different environments, such as where and when winds are stronger, solar insolation greater, subsurface temperature gradients larger, or contrasts between SSTs and air temperature larger (Folland and Parker 1995). Examination of offsets in bucket SSTs conditional on various environmental factors may help in identifying bucket attributes or measurement practices. Intercalibration of other measurements—such as temperature measurements coming from engine room intakes, radiosondes, satellites, or near-surface atmospheric thermometers (e.g., Thorne et al. 2005)—may also benefit from an LME approach.

Acknowledgments. We thank ICOADS for making data publicly available, C. Wunsch and P. Chan for helpful discussion of the results, and four anonymous reviewers for their comments and suggestions. Support was provided by the Harvard Global Institute.

APPENDIX

The linear-mixed-effect (LME) model [Eq. (1)] presented in the main text, $\delta\mathbf{T} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}_y\boldsymbol{\beta}_y + \mathbf{Z}_r\boldsymbol{\beta}_r + \boldsymbol{\beta}_\sigma$, can be equivalently written as

$$\begin{bmatrix} \delta\mathbf{T} \\ \boldsymbol{\beta}_y \\ \boldsymbol{\beta}_r \\ \boldsymbol{\beta}_\sigma \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{X}\boldsymbol{\alpha} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{Z}_y\mathbf{Z}_y^{\mathrm{T}}\sigma_y^2 + \mathbf{Z}_r\mathbf{Z}_r^{\mathrm{T}}\sigma_r^2 + \mathbf{U} & \mathbf{Z}_y\sigma_y^2 & \mathbf{Z}_r\sigma_r^2 & \mathbf{U} \\ \sigma_y^2\mathbf{Z}_y^{\mathrm{T}} & \mathbf{I}_y\sigma_y^2 & \mathbf{0} & \mathbf{0} \\ \sigma_r^2\mathbf{Z}_r^{\mathrm{T}} & \mathbf{0} & \mathbf{I}_r\sigma_r^2 & \mathbf{0} \\ \mathbf{U} & \mathbf{0} & \mathbf{0} & \mathbf{U} \end{bmatrix} \right), \tag{A1}$$

where $\delta\mathbf{T}$ is the difference between two SSTs in each pair but with climatological SST differences removed; $\boldsymbol{\alpha}$ is the fixed effect, denoting mean offsets of individual nations; $\boldsymbol{\beta}$ terms are random effects, representing yearly, $\boldsymbol{\beta}_y$, and regional, $\boldsymbol{\beta}_r$, variations in each nation; and $\mathbf{X}$ and $\mathbf{Z}$ are selection matrices. Also, $\mathbf{U}$ is the error matrix, which takes different forms based on various assumptions (see sections 3 and 5). Note that groups of random effects are assumed to follow distinct normal distributions. With this assumption, the LME algorithm accounts for shared information across regions and years, and the algorithm conservatively estimates random effects by shrinking them toward zero based on the relative contributions of random effects and random errors (Gelman and Hill 2006). In addition, reflecting the fact that there is no absolute calibration, we add the constraint that the weighted sum of all fixed effects must be zero, namely $\sum w_g\alpha_g = 0$, where w_g is the number of unique measurements in group g.

The maximum likelihood estimate of this model is obtained through maximizing a log-likelihood function (Harville 1977):

$$\mathbf{L}(\sigma_y^2, \sigma_r^2, \mathbf{U}|\delta\mathbf{T}) = -\frac{1}{2}\log(|\mathbf{V}|) - \frac{1}{2}(\delta\mathbf{T} - \mathbf{X}\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{V}^{-1}(\delta\mathbf{T} - \mathbf{X}\boldsymbol{\alpha}), \tag{A2}$$

where $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathrm{T}} + \mathbf{U}$, $\alpha = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\delta\mathbf{T}$, $\mathbf{Z} = [\mathbf{Z}_y \quad \mathbf{Z}_r]$, and,

$$\mathbf{G} = \begin{bmatrix} \mathbf{I}_y\sigma_y^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_r\sigma_r^2 \end{bmatrix}.$$

Equation (A2) has three parameters, σ_r^2, σ_y^2, and $\mathbf{U}$, that are iteratively estimated using a quasi-Newton method until a local maximum in likelihood is obtained.

The maximum likelihood estimate of $\boldsymbol{\beta}$ conditional on the fixed effects $\boldsymbol{\alpha}$ is obtained by

$$\boldsymbol{\beta} = \mathbf{G}\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{-1}(\delta\mathbf{T} - \mathbf{X}\boldsymbol{\alpha}). \tag{A3}$$

A standard t test is used to test whether the fixed effects associated with individual groups significantly differ from zero. Beyond testing whether the fixed effects of individual groups significantly differ from zero, we also examine whether two groups, i and j, of data have fixed effects that significantly differ. We use a two-sided z test, instead of a two-group t test, because the large

number of degrees of freedom in our analysis leads to the relevant distributions converging toward a standard normal. The z coefficient is

$$z = \frac{\alpha_i - \alpha_j}{(C^{\alpha}_{ii} + C^{\alpha}_{jj} - 2C^{\alpha}_{ij})^{1/2}}, \tag{A4}$$

where the covariance between all fixed effects is estimated by $\mathbf{C}^{\alpha} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$.

When yearly effects are included to account for the comparison being conducted over several multiyear bins in which both groups have yearly effect estimates, we equally weigh individual bins and extend Eq. (A4) to the following:

$$z = \frac{\left|\alpha_i - \alpha_j + \frac{1}{N}\left(\sum_{n=1}^{N}\beta_{in} - \beta_{jn}\right)\right|}{\left[C^{\alpha}_{ii} + C^{\alpha}_{jj} - 2C^{\alpha}_{ij} + \frac{1}{N^2}\left(\sum_{n=1}^{N}\sum_{m=1}^{N} C^{\beta}_{iinm} + C^{\beta}_{jjnm} - 2C^{\beta}_{ijnm}\right)\right]^{1/2}}. \tag{A5}$$

Here $\mathbf{C}^{\beta}$ is the covariance matrix of all random effects conditional on fixed effects, estimated by $\mathbf{G} - \mathbf{G}\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}$. Indices i and j indicate groups, and n and m indicate 5-yr bins.

In addition to the above pairwise significance tests, we also test whether a deck significantly differs from other decks within a nation. The test is performed iteratively, where one deck [i in Eq. (A5)] is compared against all others [grouped into j in Eq. (A5)]. In the first iteration, each deck is compared against all other decks, after which the deck having the smallest p value is removed as an outlier if $p < 0.1$. The test is then repeated until either all remaining decks insignificantly differ from one another or only one deck remains.

REFERENCES

Air Weather Service and Weather Bureau, 1955: *Reference Manual: 192 Deutsche Seewarte Marine 1859.* HQ, Air Weather Service, Data Control Division, and Weather Bureau, Climatological Services Division, 14 pp.

Anderson, T. W., 1962: *An Introduction to Multivariate Statistical Analysis*. 2nd ed. John Wiley and Sons, 721 pp.

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

Bonferroni, C. E., 1936: *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze, Vol. 8, 3–62.

Brasnett, B., 2008: The impact of satellite retrievals in a global sea-surface-temperature analysis. *Quart. J. Roy. Meteor. Soc.*, **134**, 1745–1760, https://doi.org/10.1002/qj.319.

Breckling, J., R. Chambers, A. Dorfman, S. Tam, and A. Welsh, 1994: Maximum likelihood inference from sample survey data. *Int. Stat. Rev.*, **62**, 349–363.

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, https://doi.org/10.1002/joc.4492.

——, J. Kennedy, D. Berry, S. Hirahara, C. J. Merchant, S. Morak-Bozzo, and E. Kent, 2018: Estimating sea surface temperature measurement methods using characteristic differences in the diurnal cycle. *Geophys. Res. Lett.*, **45**, 363–371, https://doi.org/10.1002/2017GL076475.

Cowtan, K., and Coauthors, 2015: Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophys. Res. Lett.*, **42**, 6526–6534, https://doi.org/10.1002/2015GL064888.

Dee, D., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

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

Freeman, E., and Coauthors, 2017: ICOADS release 3.0: A major update to the historical marine climate record. *Int. J. Climatol.*, **37**, 2211–2232, https://doi.org/10.1002/joc.4775.

Fyfe, J. C., N. P. Gillett, and F. W. Zwiers, 2013: Overestimated global warming over the past 20 years. *Nat. Climate Change*, **3**, 767–769, https://doi.org/10.1038/nclimate1972.

Gelman, A., and J. Hill, 2006: *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, 625 pp.

Harville, D. A., 1977: Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Stat. Assoc.*, **72**, 320–338, https://doi.org/10.1080/01621459.1977.10480998.

Hervey, R. V., 2014: Meteorological and oceanographic data collected from the National Data Buoy Center Coastal-Marine Automated Network (C-MAN) and moored (weather) buoys during 2014-03 (NODC Accession 0117682), version 1.1. National Oceanographic Data Center, accessed 1 May 2016, https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.nodc:0117682.

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

Huang, B., and Coauthors, 2015: Extended reconstructed sea surface temperature version 4 (ERSST. v4). Part I: Upgrades and intercomparisons. *J. Climate*, **28**, 911–930, https://doi.org/10.1175/JCLI-D-14-00006.1.

Hydrometeoizdat Moscow, 1941: *Instructions for Taking Hydrometeorological Observations at Sea by the Ship's Crew*. Hydrometeorological Service of the USSR.

Jones, P. D., and T. Wigley, 2010: Estimation of global temperature trends: What's important and what isn't. *Climatic Change*, **100**, 59–69, https://doi.org/10.1007/s10584-010-9836-3.

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

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

——, P. Brohan, and S. Tett, 2007: A global climatology of the diurnal variations in sea-surface temperature and implications for MSU temperature trends. *Geophys. Res. Lett.*, **34**, L05712, https://doi.org/10.1029/2006GL028920.

——, N. Rayner, R. Smith, D. 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, https://doi.org/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, https://doi.org/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, https://doi.org/10.1016/j.rse.2010.11.021.

Kent, E. C., and P. G. Challenor, 2006: Toward estimating climatic trends in SST. Part II: Random errors. *J. Atmos. Oceanic Technol.*, **23**, 476–486, https://doi.org/10.1175/JTECH1844.1.

——, and D. Berry, 2008: Assessment of the marine observing system (ASMOS): Final report. 55 pp., http://nora.nerc.ac.uk/id/eprint/150260/.

——, 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, https://doi.org/10.1175/JTECH1949.1.

——, J. J. Kennedy, D. I. Berry, and R. O. Smith, 2010: Effects of instrumentation changes on sea surface temperature measured in situ. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 718–728, https://doi.org/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.*, **118**, 1281–1298, https://doi.org/10.1002/jgrd.50152.

King, G., M. A. Tanner, and O. Rosen, Eds., 2004: *Ecological Inference: New Methodological Strategies.* Cambridge University Press, 421 pp.

Kleypas, J. A., G. Danabasoglu, and J. M. Lough, 2008: Potential role of the ocean thermostat in determining regional differences in coral reef bleaching events. *Geophys. Res. Lett.*, **35**, L03613, https://doi.org/10.1029/2007GL032257.

KNMI, 1937: Handleiding voor het verrichten van meteorologische waarnemingen op zee. KNMI, 98 pp.

Kobe Imperial Marine Observatory, 1925: The mean atmospheric pressure, cloudiness and sea surface temperature of the North Pacific Ocean and the neighbouring seas for the lustrum 1916 to 1920. Kobe Imperial Marine Observatory.

Medhaug, I., M. B. Stolpe, E. M. Fischer, and R. Knutti, 2017: Reconciling controversies about the global warming hiatus. *Nature*, **545**, 41–47, https://doi.org/10.1038/nature22315.

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

Oliveira, A., T. Oliveira, and A. Seijas-Macías, 2016: Evaluation of kurtosis into the product of two normally distributed variables. *AIP Conf. Proc.*, **1738**, 470002, https://doi.org/10.1063/1.4952232.

Rayner, N., P. Brohan, D. Parker, C. Folland, J. Kennedy, M. Vanicek, T. Ansell, and S. 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, https://doi.org/10.1175/JCLI3637.1.

Reynolds, R. W., T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, 2007: Daily high-resolution-blended analyses for sea surface temperature. *J. Climate*, **20**, 5473–5496, https://doi.org/10.1175/2007JCLI1824.1.

Schneider, T., T. Bischoff, and H. Płotka, 2015: Physics of changes in synoptic midlatitude temperature variability. *J. Climate*, **28**, 2312–2331, https://doi.org/10.1175/JCLI-D-14-00632.1.

Searle, S. R., G. Casella, and C. E. McCulloch, 2009: *Variance Components*. Vol. 391, Wiley Series in Probability and Statistics, John Wiley and Sons, 501 pp.

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

Stevenson, R. E., 1964: The influence of a ship on the surrounding air and water temperatures. *J. Appl. Meteor.*, **3**, 115–118, https://doi.org/10.1175/1520-0450(1964)003<0115:TIOASO>2.0.CO;2.

Thompson, D. W., 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, https://doi.org/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, https://doi.org/10.1175/BAMS-86-10-1437.

Uwai, T., and K. Komura, 1992: The collection of historical ships' data in Kobe marine observatory. *Proc. Int. COADS Workshop*, Boulder, CO, NOAA, 13–15.

Vautard, R., J. Cattiaux, P. Yiou, J.-N. Thépaut, and P. Ciais, 2010: Northern Hemisphere atmospheric stilling partly attributed to an increase in surface roughness. *Nat. Geosci.*, **3**, 756–761, https://doi.org/10.1038/ngeo979.

Wyman, R. H., 1877: Revised instructions for keeping the ship's logbook and for compiling the new meteorological returns. U. S. Navy Hydrographic Office, 28 pp.

Xu, F., and A. Ignatov, 2010: Evaluation of in situ sea surface temperatures for use in the calibration and validation of satellite retrievals. *J. Geophys. Res.*, **115**, C09022, https://doi.org/10.1029/2010JC006129.

Yang, X., 2003: Manual on sediment management and measurement. World Meteorological Organization Operational Hydrology Rep. 47, WMO-No. 948, 158 pp., https://library.wmo.int/doc_num.php?explnum_id=1709.