Systematic Differences in Bucket Sea Surface Temperatures Caused by Misclassification of Engine Room Intake Measurements

DUO CHAN AND PETER HUYBERS

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

(Manuscript received 21 December 2019, in final form 12 May 2020)

ABSTRACT

Differences in sea surface temperature (SST) biases among groups of bucket measurements in the International Comprehensive Ocean–Atmosphere Dataset, version 3.0 (ICOADS3.0), were recently identified that introduce offsets of as much as 1°C and have first-order implications for regional temperature trends. In this study, the origin of these groupwise offsets is explored through covariation between offsets and diurnal cycle amplitudes. Examination of an extended bucket model leads to expectations for offsets and amplitudes to covary in either sign, whereas misclassified engine room intake (ERI) temperatures invariably lead to negative covariance on account of ERI measurements being warmer and having a smaller diurnal amplitude. An analysis of ICOADS3.0 SST measurements that are inferred to come from buckets indicates that offsets after the 1930s primarily result from the misclassification of ERI measurements in points of five lines of evidence. 1) Prior to when ERI measurements become available in the 1930s, offset–amplitude covariance is weak and generally positive, whereas covariance is stronger and generally negative subsequently. 2) The introduction of ERI measurements in the 1930s is accompanied by a wider range of offsets and diurnal amplitudes across groups, with 3) approximately 20% of estimated diurnal amplitudes being significantly smaller than buoy and drifter observations. 4) Regressions of offsets versus amplitudes intersect independently determined end-member values of ERI measurements. 5) Offset-amplitude slopes become less negative across all regions and seasons between 1960 and 1980, when ERI temperatures were independently determined to become less warmly biased. These results highlight the importance of accurately determining measurement procedures for bias corrections and reducing uncertainty in historical SST estimates.

1. Introduction

Accurate estimates of historical sea surface temperature (SST) variability are needed for a wide range of climate studies. Applications include assessing the historical relationship between climate variability and tropical cyclones (Vecchi et al. 2011), exploring whether the characteristics of El Niño–Southern Oscillation have changed (Yeh et al. 2009), attributing internal versus externally forced climate variability (Ting et al. 2014), and determining which radiative feedbacks have historically participated in driving climate change (Armour et al. 2013). It is thus of broad relevance that recently identified systematic offsets among groups of bucket SST measurements alter estimates of regional, multidecadal SST variability by as much as 0.5°C and increase the associated uncertainty estimates by an order of magnitude relative to foregoing estimates (Chan et al. 2019; Chan and Huybers 2019).

A wide variety of factors could potentially explain the presence of errors in bucket measurements (Kent et al. 2017); these factors can be divided into physical and nonphysical categories. Physical processes are defined as those causing differences between temperatures of measured water and those at the surface of the ocean and are generally related to solar heating and evaporative and sensible cooling. Relative contributions to heating and cooling of a bucket will depend on bucket characteristics, environmental conditions, and measurement protocols (Ashford 1948; Folland and Parker 1995; Carella et al. 2017b). Nonphysical processes that can influence SST reports include miscalibration or errors in thermometer readings (Kent et al. 2017), misclassification of engine room intake (ERI) measurements as coming from buckets (Carella et al. 2018), or record-keeping errors. As an example of the latter case, SST estimates originally reported to tenths of a degree Celsius in the Japanese Kobe collection were truncated in the process of digitization, causing biases in the northwest Pacific Ocean of 0.45°C (Chan et al. 2019).

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

DOI: 10.1175/JCLI-D-19-0972.1

There are widely used methods to correct for certain systematic biases associated with bucket SST measurements. The fact that more evaporative cooling is expected from canvas than wooden buckets is, for example, accounted for in HadSST estimates using a temporally linearly varying but spatially uniform proportion of canvas to wooden buckets (Folland and Parker 1995; Rayner et al. 2006; Kennedy et al. 2011). The ERSST5 estimate from NOAA instead applies a fixed spatial pattern of corrections derived by comparing SSTs against nighttime marine air temperatures (Huang et al. 2017). A method similar to that of NOAA's was recently proposed where bucket SSTs are instead compared against coastal and island weather station measurements (Cowtan et al. 2018). The HadSST3 bucket corrections at the level of individual grid boxes range from −0.1° to +1°C and for the global average range from −0.05° to 0.45°C [99% uncertainty range from Kennedy et al. (2011)].

Uncertainties in bias corrections are a major contribution to the uncertainty in global warming over the last century (Jones 2016). A major issue with the foregoing methods for correcting bucket temperatures is difficulty in accounting for regional changes in measurement details. For example, from 1900 to 1913, most SST measurements in the South Pacific and the South Atlantic Oceans come from German compilations, averaging 231 000 measurements per year. However, from 1914 to 1920, contributions from German compilations drop to only 38 000 measurements per year, and the United Kingdom becomes the dominant source of SST data in these ocean basins. Both German and U.K. compilations include sources from a variety of nations, but the composition of observations differs between these sources. Changes in the mixture of bucket designs or measurements protocols present in the compilations could, for example, lead to distinct biases and, thus, offsets among data sources.

Chan and Huybers (2019) used a linear-mixed-effect (LME) model to detect offsets among groups of SSTs that are thought to be bucket measurements. Groupwise offsets are relative to the mean of all measurements used in the LME analysis, and the range of associated SST adjustment is from −1.0° to 1.3°C at the level of individual grid boxes (Chan et al. 2019). Because these corrections are systematic across space and time, they can have major implications for regional trends. For example, a trend in North Pacific SST between 1908 and 1941 changes from 0.31° to 0.56°C over the 34 years when applying offset corrections (Chan et al. 2019). We note that the recently published HadSST4 dataset (Kennedy et al. 2019) may also implicitly account for groupwise SST offsets after 1941 by comparing bucket measurements with XBT and CTD measurements at a monthly 5° resolution. Although many offsets are statistically highly significant (Chan and Huybers 2019), the origins of these offsets are generally unknown. Lack of metadata makes using features of the temperature measurements themselves attractive for purposes of further exploring the origins of observed offsets.

One indicator of bucket characteristics comes from the diurnal cycle of SST measurements, where the diurnal cycles of bucket measurements generally have a larger amplitude and are more in phase with diurnal insolation variability than drifter, buoy, and ERI measurements. Carella et al. (2018) used diurnal amplitudes to better distinguish between measurements coming from buckets and ERIs. They inferred nearly 100% accuracy after the 1990s but that approximately 10%–20% of the bucket measurements available from between the 1930s and 1980s are misclassified. Given opposing offsets associated with warm ERI measurements and cool bucket measurements, such misclassification has the potential to cause substantial variation in the mean offsets associated with different groups.

Herein a method to evaluate mean groupwise offsets against the amplitude of diurnal cycles among groups of bucket SSTs is developed. After introducing data and methods, we develop baseline expectations of offset–amplitude relationships by examining the response of a thermodynamic model of a wooden bucket to plausible parameter changes. We note that a wooden bucket model will behave like one for canvas buckets if the wood is prescribed to be 2 mm thick (Folland and Parker 1995), as explored in section 3. We then diagnose offset–amplitude relationships from version 3.0 of the International Comprehensive Ocean–Atmosphere Dataset (ICOADS3.0) and consider physical and nonphysical contributions to bucket SST offsets.

2. Data and methods

A portion of the data processing and methods that we apply build from the approaches used in recent work (Chan and Huybers 2019; Chan et al. 2019). In the following section we highlight improvements to some procedures and adaptations of others to focus on characterizing diurnal SST variability in relation to mean offsets.

a. Identification of bucket measurements

In situ SST observations used in this study are from ICOADS3.0 (Freeman et al. 2017). There exist incomplete and sometimes contradictory indications regarding which of these measurements were obtained by buckets. We generally follow Kennedy et al. (2011) in identifying measurements likely to have come from buckets, and in this respect our procedure exactly corresponds to that of

TABLE 1. List of decks that are combined when grouping SST data. 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**, 761
Netherlands Marines	189, **193**
Scottish Fishery Cruiser; MARIDs	**205**, 211
Met Office selected ships	**203**, 207, 209, 213, 223, 227, 233
U.S. NCEP: ship data	792, **892**
International Maritime Meteorological Data/International Marine	128, 254, 926, **927**

Chan and Huybers (2019) and Chan et al. (2019). First, following Kennedy et al. (2011), before 1941, all SSTs are assumed to be bucket measurements unless explicitly indicated otherwise in ship log books (Freeman et al. 2017), referred to as ICOADS-SI. Second, from 1941 onward, if ICOADS-SI metadata are missing, metadata from World Meteorological Organization Publication 47 (WMO No. 47 hereinafter; Kent et al. 2007) are instead relied upon. In total 22.4% of quality-controlled and ship-based SSTs after 1941 are identified as coming from buckets using ICOADS-SI and another 3.9% using WMO No. 47 metadata.

If metadata are missing from both WMO No. 47 and ICOADS-SI, a measurement is assumed to come from a bucket if more than 95% of ships from the same country for which metadata are available used buckets in the same year (Kennedy et al. 2011), which identifies another 5.2% of all ship-based SSTs to be bucket measurements. If country metadata are missing, country information is inferred from ship call signs (Chan et al. 2019) or, otherwise, from deck number (Kennedy et al. 2011). The ''deck number'' refers to batches of punch cards associated with early digitization of much of the ICOADS data and, although not specifically organized according to physical or procedural methods, temperatures reported across certain decks contain highly significant offsets ($P < 0.1$; Chan and Huybers 2019).

To compare subsets of bucket measurements, we assign groups according to combinations of deck numbers and nations. Measurements not associated with a nation are combined into separate groups according to deck number. Following Chan and Huybers (2019), we combine decks that have the same descriptions in ICOADS, but we now also combine decks 254 and 926 (International Maritime Meteorological Data) with decks 128 and 927 (International Marine—U.S. or foreign-keyed ship data; Table 1). Data from East Germany and West Germany are now treated as separate groups, a distinction omitted by Chan and Huybers (2019) and Chan et al. (2019). Analyses are conducted in 20-yr intervals, and groups contributing less than 6000 pairs in such a period are excluded for purposes of computational efficiency. We describe the general insensitivity of our conclusions to plausible changes in these procedures in appendix A.

b. *Diurnal amplitudes*

Our analysis also makes use of recent estimates of individual ship tracks (Carella et al. 2017a) that are available for 82% of bucket measurements between 1880 and 2009 in ICOADS3.0. The ability to obtain sequences of observations from a single ship permits for estimating the diurnal cycle under conditions where observing characteristics, such as bucket type and on-deck time, are expected to be more homogeneous. We perform analyses for 20-yr periods starting at 1880–99 and move the analysis forward at annual increments until 1990–2009. Ship tracks are not available after 2009 (Carella et al. 2017a), but neither are bucket measurements as common (Kennedy et al. 2011, 2019). For each 20-yr period, similar to Morak-Bozzo et al. (2016) and Carella et al. (2017a), measurements covering a single day from a single ship are used to compute a diurnal anomaly relative to daily-average SST if there are SST observations corresponding to each 6-hourly interval from local midnight. Diurnal SST anomalies are aggregated by local hours for each nation-deck group and are averaged annually for the tropics (20°S–20°N) and seasonally outside the tropics (20°–40° and 40°–60°N). Moreover, only observations that meet these specifications for inclusion in estimating the diurnal cycle are used for purposes of computing intergroup offsets.

Diurnal amplitudes are calculated by fitting the phase and amplitude of a sinusoid having a one-day period, where fitting is weighted by the sample size in each hourly bin. Note that unlike Carella et al. (2018), who examined the amplitude of excess diurnal cycles relative to a climatology obtained from drifting buoys, we apply

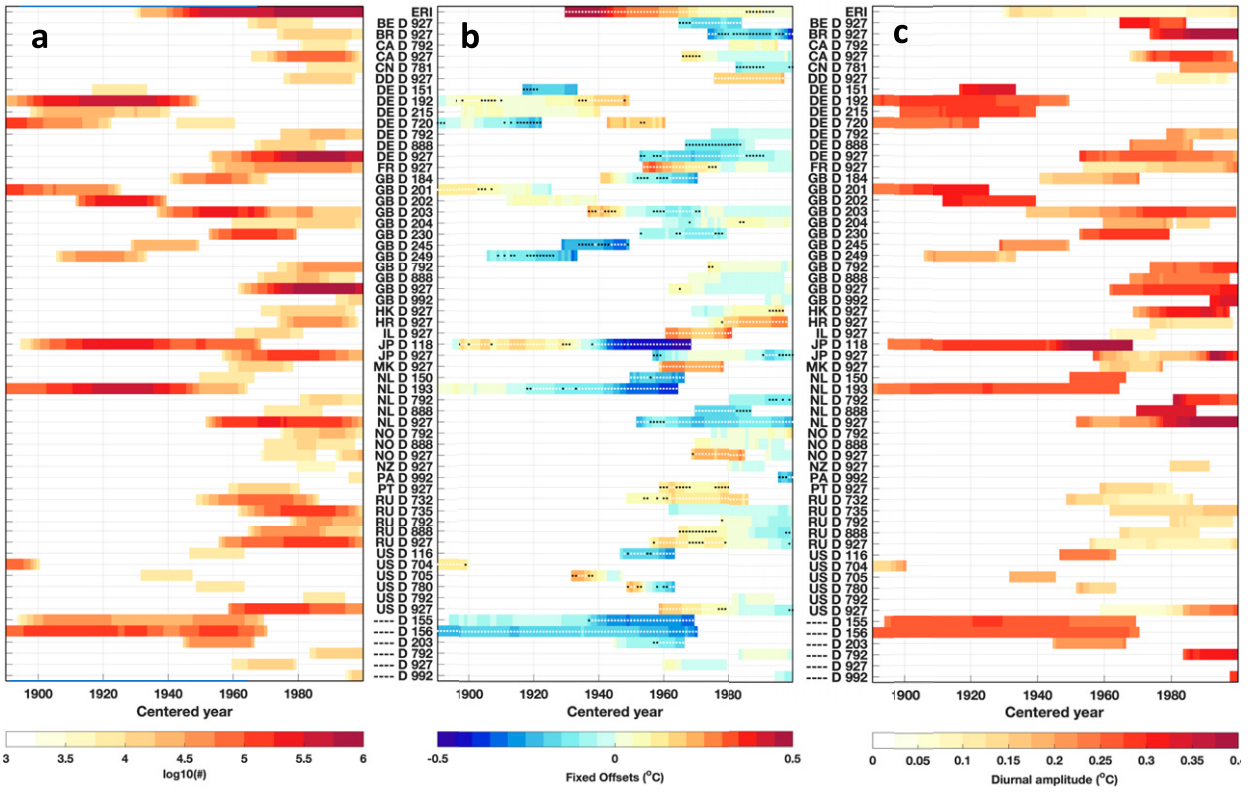

FIG. 1. Number of measurements, offsets, and diurnal amplitudes: (a) Numbers of measurements (shading) from individual groups for each 20-yr LME analysis. (b) As in (a), but for annual-mean fixed offsets. Fixed offsets are relative to an unknown mean bias across all measurements in each 20-yr analysis. Statistically significant offsets are indicated by black dots ($p < 0.05$), and offsets that remain significant after a Bonferroni correction for multiple hypothesis testing are indicated by white dots. (c) As in (b), but for amplitudes of annual-average diurnal cycles in the tropics (20°S–20°N). Groups that do not sample the tropical oceans, such as those associated with Norway, are omitted for the purposes of this figure only.

a harmonic fitting to full diurnal cycles of bucket SSTs. We also explored fitting higher-frequency sinusoids, as implemented elsewhere (Kennedy et al. 2007). The amplitude of a twice-per-day sinusoid was found to be highly correlated to that of the first-order harmonics across bucket groups ($r = 0.85$ in the tropics in 1970–89), however, indicating that the shape and amplitude of diurnal cycles differ in a consistent manner and that the first-order harmonic fitting is sufficient for capturing and summarizing differences in diurnal amplitudes among groups. The same procedure is also used to estimate diurnal amplitudes associated with ERI measurements.

c. *Linear-mixed-effect model*

An LME model is used to identify offsets among groups of bucket measurements relative to the overall average of all paired measurements. The LME model is described in detail by Chan and Huybers (2019) but three changes are made here. First, to explore plausible seasonality in offset-diurnal relationships, we include seasonal effects for December–February (DJF), March–May (MAM), June–August (JJA), and September–November (SON) over the latitude bands 0°–20°, 20°–40°, 40°–60°, and 60°–90°, leading to as many as 16 seasonal parameters for each group. Southern Hemisphere measurements are shifted by one-half year to account for the seasonal asymmetry between hemispheres. Second, unlike our previous analyses in which decadal variations were controlled when estimating variations between 1850 and 2014, the present analysis is performed over a sliding 20-yr interval without additional controls for decadal variations. ERI observations are included as a single group to empirically constrain offsets relative to bucket temperatures. Figure 1 summarizes the number of data from individual groups in each 20-yr analysis, together with estimated offsets and diurnal amplitudes.

Because both groupwise offsets and diurnal amplitudes are uncertain, a York fit is used for purposes of estimating trends in offsets as a function of diurnal amplitude (York et al. 2004). The associated 95% confidence intervals are obtained by a bootstrapping technique that randomly resamples nation-deck groups with replacement and repeats for 10 000 times. Although ERI measurements are incorporated in the analysis, only bucket SST groups are used in York regressions. Under

the assumption of a linear relationship, the intergroup variability explained by diurnal amplitudes is quantified as the square of their Pearson's correlation coefficient (r^2). Confidence intervals are estimated following Lane et al. (2013). Codes for reproducing all results are available online (https://github.com/duochanatharvard/LME-Offsets-vs-Diurnal-Amplitudes).

3. Bucket simulations

To develop baseline expectations for variability in mean offsets and the amplitude of diurnal cycles, we first examine these properties in the context of a wooden bucket model, where the original model is that of Folland and Parker (1995, hereinafter FP95) and our update is referred to as FP95d to indicate that it represents diurnal variability. Appendix B describes our inclusion of solar heating. We use the same parameters as originally proposed for the FP95 model (Table 2), but with two exceptions for processes not otherwise fully accounted for. First, bucket temperatures may not be fully equilibrated with SST before a water sample is measured, suggesting that some percentage of air temperature may be retained. Second, a bucket may be in the shadow of a ship or measured within a sheltered enclosure, suggesting that the percentage of absorbed solar radiation should also be specified (Carella et al. 2018; Kennedy et al. 2019).

We use our model to simulate diurnal variations in bucket SSTs as a function of location and season. To initialize the FP95d model we use SSTs that are a combination of daily averages from National Oceanography Center (NOC), version 2.0 (Berry and Kent 2009), and diurnal anomalies from drifting buoys (Chan and Huybers 2019). For purposes of brevity, we refer to this combined estimate as drifter SSTs. Atmospheric conditions associated with air temperature, humidity, and surface wind are estimated using daily-mean values from NOCv2.0 and diurnally resolved values from ships that report bucket SSTs (see appendix B for more details). Downward insolation at the ocean surface is from 3-hourly ERA-interim reanalysis (Dee et al. 2011). Model fit is computed using the root-mean-square error (RMSE) between observed and modeled bucket water temperatures. The two added parameters are assigned values that minimize the RMSE averaged over all combinations of regions and seasons between 1990 and 2009. The best fit comes from an initial bucket temperature at the time of collecting seawater that represents a 20% mixture of on-deck air temperature and 80% actual sea surface temperature, and conditions wherein absorbed solar radiation is reduced to 70% of total available insolation.

For purposes of evaluating the skill of the FP95d model, it is useful to examine the difference between simulated bucket SSTs and the original SSTs used for driving the model. These simulated modification of SST by the bucket sampling procedure are then compared against differences between bucket observations of SSTs and SST values from drifters, where the latter are considered to more closely indicate actual SSTs. After averaging according to region and season, the FP95d model generally reproduces the difference in temperature between bucket and drifter observations in terms of amplitude, phase, and seasonality of diurnal cycles (Fig. 2). For the annual mean in the tropics, RMSE decreases from an average of 0.12°C between the observed bucket and drifter SSTs to 0.04°C between observed and modeled bucket water temperatures (Fig. 2a; $P < 0.05$ in standard F test; $N = 24$). Similarly good fits are found for other regions and seasons across 1990–2009.

An exception to the overall good fit of our bucket model to observations is that modeled bucket temperatures lead observed bucket temperatures by about an hour. We speculate that this lead comes from the fact that, following FP95, we assume water temperature in the bucket is homogeneous. In laboratory experiments, Carella et al. (2017b) found that unstirred buckets tend to cool more slowly than predicted from bucket models, and a slower rate of cooling is consistent with a greater lag. A more

TABLE 2. Parameters for the FP95d extended wooden bucket model. Values are assigned following Folland and Parker (1995). Two additions that were not parameterized in the original bucket model are insolation and a percentage of air temperature in initial bucket temperature.

Parameter	Value
Exposure time (s)	240
Bucket thickness (cm)	1
Bucket diameter (cm)	25
Bucket depth (cm)	20
Insolation (%)	70
Initial bucket temperature (% of air temperature)	20
ERI misclassification (%)	0
Mean apparent wind ($m\,s^{-1}$)	5.5
Ship speed ($m\,s^{-1}$)	7
Ambient wind exposure during hauling (%)	60
Ambient wind exposure on deck (%)	40
Ship speed exposure during hauling (%)	100
Ship speed exposure on deck (%)	67
Density of bucket ($kg\,m^{-3}$)	800
Specific heat of bucket ($J\,kg^{-1}\,K^{-1}$)	1900
Albedo of bucket	0
Time of hauling (s)	60
Heat capacity of thermometer (gram of water)	35
Turbulence viscosity ($m^2\,s^{-1}$)	1.5×10^{-5}
Water thickness on wall (mm)	0.1
Relative humidity at water surface	0.98

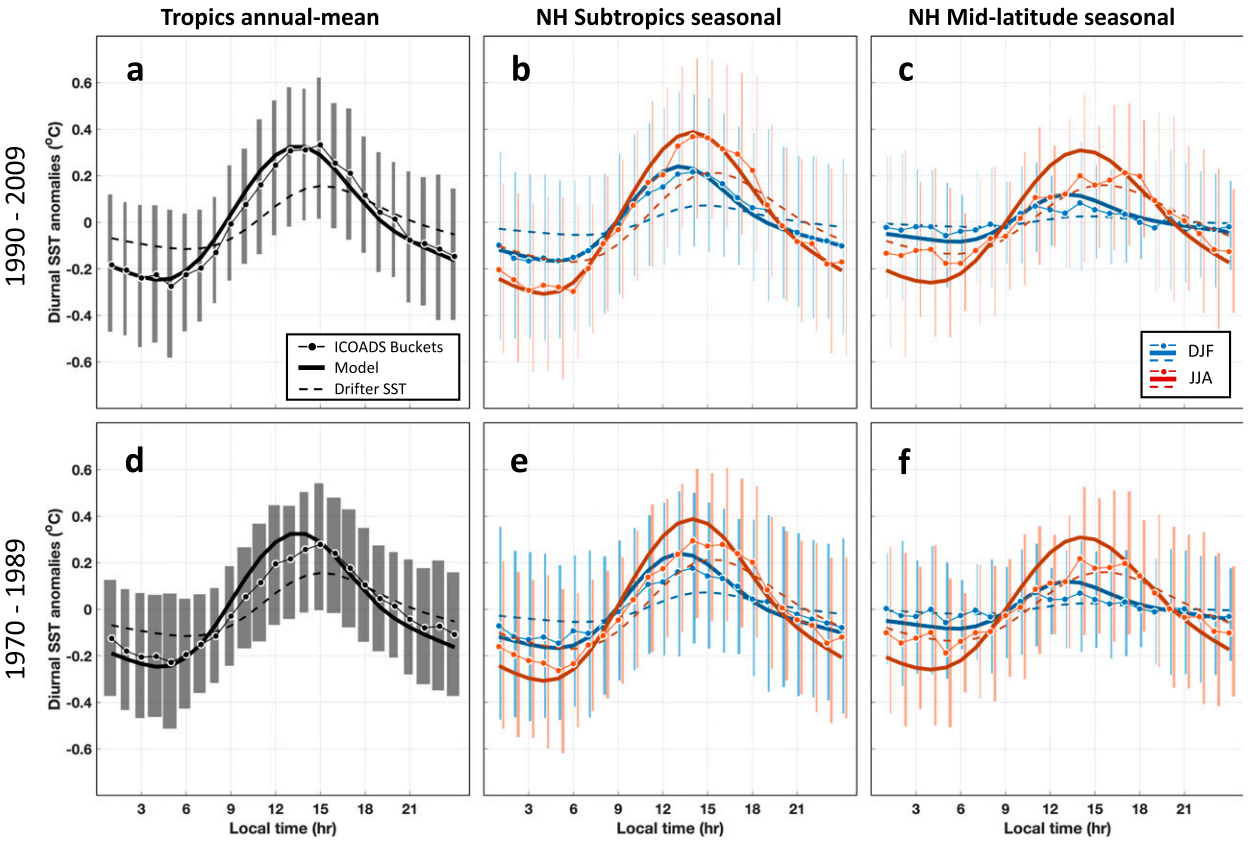

FIG. 2. Observed and modeled diurnal cycles of bucket measurements: The observed diurnal cycle of bucket temperatures (dotted lines) is in better agreement with diurnal variability simulated by the FP95d model (thick solid lines) than the diurnal cycle diagnosed from buoy and drifter measurements (dashed lines). Shown are the average diurnal cycle (top) between 1990 and 2009 and (bottom) between 1970 and 1989 for (a),(d) the annual mean over the tropics (20°S–20°N), (b),(e) DJF (blue) and JJA (red) over the Northern Hemisphere subtropics (20°–40°N), and (c),(f) DJF and JJA over the Northern Hemisphere midlatitudes (40°–60°N). Model parameters are prescribed according to Table 2 for all depicted simulations. The interquartile range of observations is indicated by bar lengths, and sample size is proportional to bar width.

complete model representing the dynamics of water in a bucket may be useful to explore in future work.

To explore the relationship between diurnal cycles and biases in bucket SSTs, we consider plausible perturbations to model parameters. FP95 highlighted four sources of physical uncertainty in parameterizing their bucket model: exposure time, bucket insulation, bucket size, and apparent wind. As noted above, recent findings also suggest consideration of insolation absorption, initial bucket temperature, and misclassified ERI data (Carella et al. 2018; Kennedy et al. 2019).

a. *Exposure time*

The lower bound on elapsed time between a bucket's extraction from the water and measurement is taken as 1 min, consistent with the time needed for hauling a bucket on deck (FP95), except for perhaps with respect to smaller nineteenth-century ships. Once buckets are brought on deck, FP95 assigned an average on-deck time of 4 min for wooden buckets, which was estimated to have a standard error of 13% (Rayner et al. 2006). We, however, expect the range of on-deck time for individual nations to be wider because documents indicate that the amount of time thermometers were left to equilibrate with water ranges from one minute or less (e.g., Wyman 1877; Ashford 1948), to waiting for a steady state to be reached (e.g., Kobe Imperial Marine Observatory 1925), which perhaps ranges out to 10 min. We thus explore total exposure times ranging from 1 to 11 min.

b. *Bucket insulation*

Different types of buckets may have distinct rates at which heat fluxes in or out of the water, which is mathematically similar in our model to differences in exposure time. To account for different bucket insulation, FP95 considered separate models for thin canvas buckets and 1-cm-thick wooden buckets. Although canvas buckets have water leakage, a higher albedo, and sometimes include a lid, FP95 indicate that canvas model results are generally reproducible by assuming a 2-mm-thick wooden bucket of the

same size. We, therefore, explore a wooden bucket having wall thicknesses ranging between 0.2 and 2 cm.

c. Bucket size

Small buckets tend to have a larger ratio of surface area to volume and, therefore, exchange heat more efficiently than large buckets (FP95; Ashford 1948). We adopt the three bucket sizes listed by FP95: a large bucket of 25-cm diameter and 20-cm depth, a medium bucket of 16.3-cm diameter and 14-cm depth, and a small bucket of 8-cm diameter and 12-cm depth.

d. Apparent wind

The wind experienced by a bucket is influenced by the wind speed, relative ship motion, and the degree of sheltering. FP95 took apparent wind to equal sheltered wind speed and ship speed summed in quadrature, assuming wind directions to be uniformly distributed across all angles and giving a mean apparent wind of approximately $5.5\,m\,s^{-1}$. For an upper bound, we assume a ship under power making $10\,m\,s^{-1}$ into a prevailing wind of $5\,m\,s^{-1}$, where such ship speed is the approximate upper bound indicated in Fig. 11 of Carella et al. (2017a). Although it is unadvisable for sailors to make bucket measurements on a high-speed ship ($>7.2\,m\,s^{-1}$) out of safety concerns (Met Office 1956), for the purpose of exploring possible ranges, we test this limit in the FP95d model. This upper bound is specified in FP95d by scaling the standard apparent wind by a factor of 3. For the lower bound, we assume no wind for an entirely sheltered bucket.

e. Insolation

FP95 noted limited evidence, mostly pertaining to nineteenth-century reports, that bucket measurements were exposed to direct solar radiation on ship decks. Carella et al. (2018), however, showed excessive diurnal cycles for bucket SSTs that they attribute to solar heating, and Kennedy et al. (2019) gives evidence for strong solar heating over the midlatitude summer. We explore the full possible range of exposure to insolation from 0% to 100%. Variations in insolation can arise from changes in either solar shading or bucket albedo, which we do not distinguish in this analysis.

f. Initial bucket temperature

If the wood in a bucket of 25-cm diameter and 20-cm depth is specified to be 2 cm thick, it accounts for approximately 16% of the total heat capacity when the bucket is filled with seawater. In an extreme case where the bucket has no time to equilibrate with seawater before hauling, approximately 16% of the water temperature measured in the bucket could instead reflect the initial bucket temperature. Taking into account uncertainties in bucket designs and uncertainties in air–sea temperature differences, we explore up to 20% of the initial bucket temperature in fact representing air temperature. Here, we assume that the initial temperatures of the bucket material and water in the bucket are in equilibrium. Also possible is for buckets to be cooler than on-deck air temperature if not kept dry and subject to evaporation (Brooks 1926) or warmer than on-deck air temperature if in direct sunlight, but these additional complications are not accounted for.

g. Misclassification of ERI measurements

To the foregoing list of physical effects on buckets, we add the nonphysical effect of incorrectly categorizing ERI measurements as coming from buckets. Although the reasons for ERI measurement bias are themselves physical, in the present context these are considered nonphysical because they are imposed on account of incorrectly identifying a data source. ERIs sample water coming from below the surface that are, hence, initially biased cold. Warming of water within the engine room, however, leads to temperatures that are generally biased between 0.1° and 0.3°C warm relative to actual SST (Kennedy et al. 2011; Kent et al. 2017). The greater depth at which ERI measurements come from also implies a diurnal cycle having a smaller amplitude (Kawai and Wada 2007; Carella et al. 2018).

As noted, measurement type is inferred for ICOADS3.0 data from an indicator in ship log books (Freeman et al. 2017) or, after 1960, from WMO No. 47 (Kent et al. 2007), but there is substantial uncertainty in the provenance of many measurements. For example, Kennedy et al. (2011) and Hirahara et al. (2014) estimate that the proportion of measurements coming from buckets is 60% between 1960 and 1980 but Carella et al. (2018) estimate that only 40% of observations come from buckets during this interval. There exists the potential for entire groups of data to be misidentified, and we explore scenarios having between 0% and 100% misclassification of ERI measurements. To represent ERI misclassification, we estimate the diurnal cycle of ERI SSTs from 1990–2009 ERI measurements in ICOADS3.0 as a function of region and season and assume that ERI SSTs are warmly biased by 0.1°C.

Individual parameters are varied in the bucket model across the above-indicated ranges sequentially. The full diurnal amplitude and the bias in daily-mean temperature are examined for different combinations of latitude bands and seasons (Figs. 3a–c). Mean bias is computed as the daily-average difference between modeled bucket water temperatures and drifter SSTs. Diurnal amplitude is obtained by fitting a once-per-day sinusoid to modeled water temperatures in buckets.

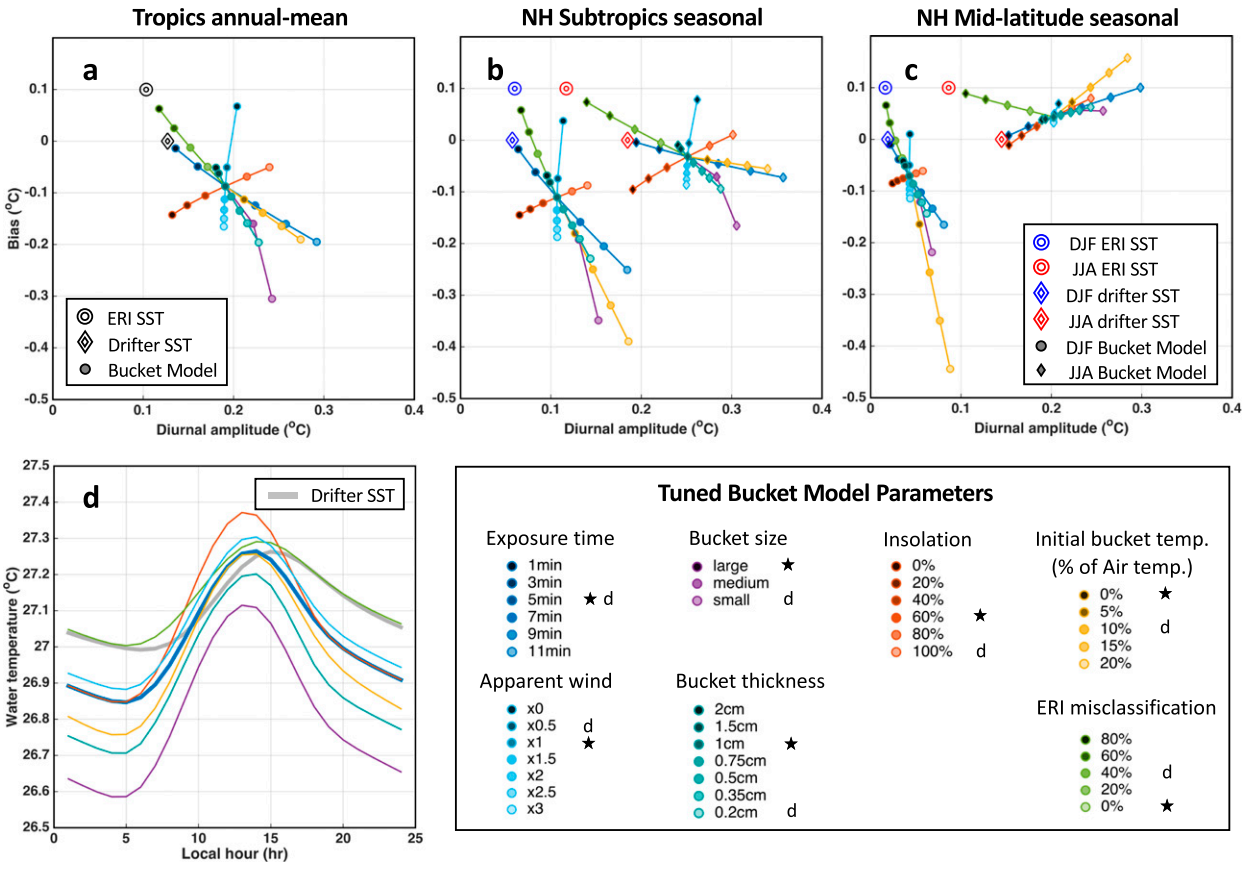

FIG. 3. Correspondence between diurnal amplitudes and daily-average biases as function of model parameters: Changes in SST offsets and diurnal amplitudes in response to changes in model parameters are shown for (a) the tropics, (b) Northern Hemisphere subtropics, and (c) Northern Hemisphere midlatitudes. Extratropical results are for winter (lines with circles) and summer (lines with diamonds). Reference parameters are indicated in the legend (black stars) and are listed in Table 2. (d) Example diurnal cycles for the tropics are shown as estimated from drifting buoys (thick gray line), the reference simulation (thick blue line), and simulations that vary individual parameters (thin lines and values indicated by "d" in the legend).

Most parameter variations lead to an anticorrelation between mean temperature biases and diurnal amplitudes (Fig. 3a). The longer a bucket is aerially exposed, the more evaporative cooling and daytime solar heating it experiences, leading to a larger diurnal amplitude. Furthermore, because net evaporative heat loss is generally greater than solar heating, longer aerial exposure generally also leads to colder mean temperatures. An exception to aerial exposure leading to daily-average cooling is in certain long-daylight, high-intensity cases found during summertime (Fig. 3c). Similar trends toward greater daily-average cooling and increased diurnal amplitude result from decreasing bucket insulation or bucket size, as well as for prescribing a greater influence of initial air temperature. The latter arises because the shipboard air temperature responds more strongly than SST to the diurnal cycle because of greater sensitivity to solar heating (Berry et al. 2004) but its daily mean is usually cooler than SSTs. An important further effect is that misclassification of ERI measurements introduces samples having, on average, warmer temperatures and smaller diurnal amplitudes into a group, thereby altering offsets and amplitudes along an axis similar to the foregoing properties.

There are, however, a few cases in which the scaling between mean temperature biases and diurnal amplitudes is nonnegative. For example, increasing the insolation absorbed by a bucket causes a largely orthogonal response because it gives a larger diurnal amplitude and a higher diurnal average temperature through daytime warming. In addition, increasing the apparent wind leads to mean cooling because of greater wind-induced evaporation but has almost no influence on simulated diurnal amplitudes.

Summer and winter exhibit distinct offset–amplitude relationships (Figs. 3b,c). During winter there are weaker

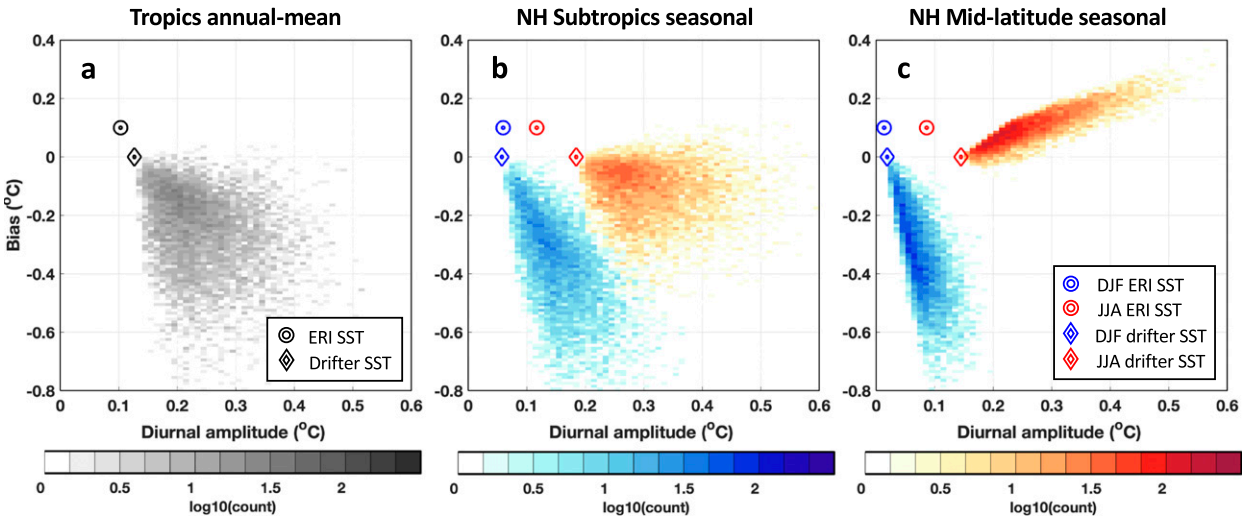

FIG. 4. Randomized bucket simulations: Diurnal amplitude and daily-mean biases are shown from 10 000 randomized bucket simulations for the (a) tropics, (b) Northern Hemisphere subtropics, and (c) Northern Hemisphere midlatitudes. Extratropical results are for winter (blue) and summer (red). For each run, tunable bucket parameters are drawn from uniform distributions whose ranges are indicated in Fig. 3, with the exception that misclassification of ERI is fixed at 0%.

diurnal variation in insolation and a generally deeper mixed layer, leading to smaller-amplitude SST diurnal cycles. Bucket temperature, however, cools faster in colder wintertime air through both evaporative and sensible heat fluxes accentuating cold offsets. During winter we, therefore, expect offsets to be colder and diurnal amplitudes to be smaller, leading to steeper slopes, and vice versa in summer. Such seasonality is stronger at higher latitudes. During summertime, midlatitude solar gain may outperform evaporative cooling, leading to a reversal in slope whereby greater exposure to insolation leads to increased diurnal amplitude and overall warmer temperatures. A positive slope may also be obtained on account of initial conditions because summertime on-deck air temperatures in the midlatitude are generally warmer than SSTs.

In addition to varying parameters sequentially, we also explore simultaneous parameter perturbations. We draw sets of tunable parameters 10 000 times and compute amplitude and mean offsets in each case (Fig. 4). The range of parameters is the same as that in the legend of Fig. 3, except that we fix the misclassification of ERI to be 0% in order to facilitate comparisons of bucket results against ERI values. These realizations of an ensemble of parameters have the advantage of capturing interactions between parameter changes and more fully describe the range of possible model behaviors (Fig. 4). For example, a small and thin bucket that stays on deck for an anomalously long time will have a larger diurnal cycle (up to 0.5°C in the tropics) and a colder daily mean bias (down to −0.6°C in the tropics) than would result from individual perturbations. It follows that the uncertainties of the offset–amplitude slopes are greater after jointly accounting for uncertainties of model parameters. Notable, however, is that all randomized simulations predict diurnal cycles that are no less than the original SST being sampled.

4. Results

Observational results generally indicate that groups that are offset cold also have a larger diurnal amplitude (Fig. 5). In the tropics, a strong anticorrelation is found between the average offset and the diurnal amplitude among groups over 20-yr periods between 1930 and 2009, with the mean r^2 being 0.51 (Fig. 6a). Predicted negative slopes of offsets as a function of amplitude range from −4.5° to −1.2°C °C^{-1} as a function of individual parameters (Fig. 3), and observed slopes similarly range from −4.2° to −0.5°C °C^{-1} (Fig. 6b). The range of predicted slopes when simultaneously changing model parameters is even larger as a consequence of involving combinations of maximal variations in parameters (Fig. 4). The range of observed amplitude and offset values also generally accord with those simulated by the bucket model, with the exception of small diurnal amplitudes associated with certain groups that we return to later.

Subtropical and midlatitude regions also generally have a strong negative relationship between offsets and amplitudes after the 1930s. Furthermore, in these regions, it is possible to examine trends during different

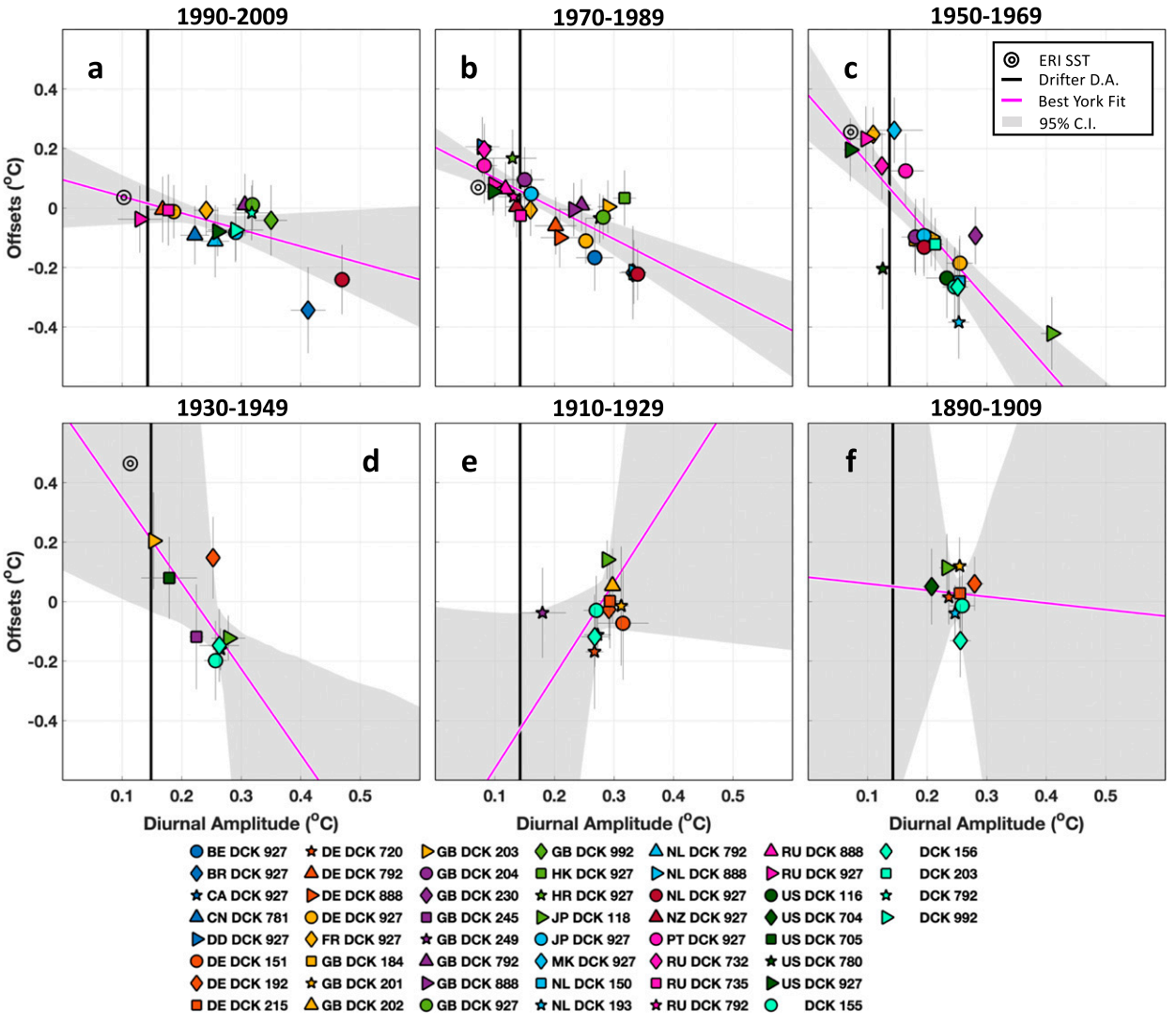

FIG. 5. Groupwise bucket SST offsets and diurnal amplitudes in the tropics: (a)–(d) Clear negative covariance exists between offsets and diurnal amplitudes across groups for 20-yr periods between 1930 and 2009, but (e) covariance weakens and changes sign between 1910 and 1929 and (f) is essentially absent between 1890 and 1909. Two-standard-deviation uncertainties are estimated from the LME analysis for each group (vertical bars on each marker) and for the least squares sinusoidal fit of amplitude (horizontal bars). The central estimate of a York regression (magenta line) is also shown in each panel along with its 95% coverage interval estimated by bootstrapping individual groups (gray shading). Note that regressions intersect the offset and diurnal amplitude of ERI measurements (double circles) since ERI becomes available in 1930; note also that numerous groups show a diurnal amplitude that is similar to or lower than that of drifter SSTs (vertical black lines).

seasons. In the subtropics, offset–amplitude slopes are more negative in winter than summer (Figs. 7a,b and 8b), as predicted (Fig. 3b). In the midlatitudes, significant positive trends are found in summer after the 1970s (Fig. 7d), also consistent with predictions (Fig. 3c). The lowest offset–amplitude correlations and most uncertain York fit slopes are found for the midlatitude winter (green curves in Figs. 8a,b), which features the smallest diurnal amplitude. Note that groupwise offsets estimated from the LME analysis (Figs. 5 and 7) are relative to the mean bias of all measurements used in that 20-yr analysis, which is distinct from the absolute biases obtained from our bucket simulations (Figs. 3 and 4).

Slopes between groupwise offsets and diurnal amplitudes do not, of themselves, allow for distinguishing between individual contributions coming from initial bucket temperature, exposure time, heat transfer rates, or misclassification of ERI measurements because each gives similar relationships. There are five lines of

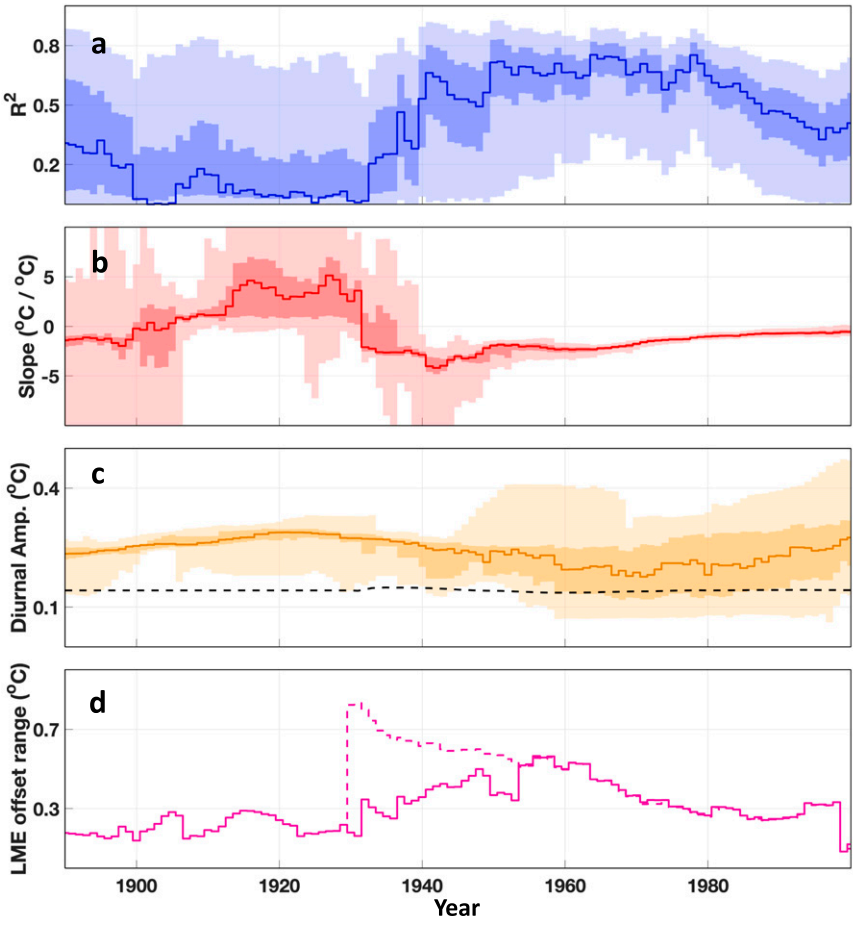

FIG. 6. Evolution of groupwise offsets and diurnal amplitudes in the tropics: (a) squared cross-correlation between diurnal amplitudes and groupwise offsets, (b) slope from a York fit, (c) diurnal amplitude, and (d) the 25%–100% range of offsets between bucket groups. Panels (a)–(d) show median values (solid lines), interquartile ranges (dark shading), and 95% ranges (light shading). Also shown is the climatological diurnal amplitude from drifting buoys averaged over locations where bucket SSTs are measured [dashed line in (c)] and the 25%–100% range of groupwise offsets including ERI data [dashed line in (d)]. All analyses are from a 20-yr sliding window, with results plotted against the average year.

additional evidence, however, that support misclassification of ERI measurements as the predominant source of intergroup variations in offsets and amplitudes since the 1930s.

First, before the 1930s, ERI measurements are not available (Carella et al. 2018) and there is weak covariance between offsets and amplitudes that is generally positive (Figs. 5e,f and 6a,b). After introduction of ERI measurements in the 1930s, offset–amplitude covariance is strong and generally negative (Figs. 5a–d and 6a,b). We are not aware of another large-scale change in observational characteristics that would so strongly alter the covariance between diurnal amplitudes and mean temperatures.

Second, the spread in both groupwise offsets and amplitudes is narrower prior to 1930 than after (Figs. 6c,d and 8c,d). In the tropics, the interquartile range of diurnal amplitudes averages 0.02°C prior to 1930 and 0.11°C afterward (Fig. 6c). Similarly, the 25th–100th range in offsets goes from 0.21°C before 1930 to 0.35°C afterward (Fig. 6d). The sense of amplitude–offset variation before the 1930s appears consistent with differences in wind exposure (Fig. 3), although it may also result from a combination of several factors involving bucket designs and measurement protocols (Fig. 4).

Third, diurnal amplitudes prior to 1930 center on values that are significantly greater than buoy and drifter SSTs and are consistent with bucket measurements. In

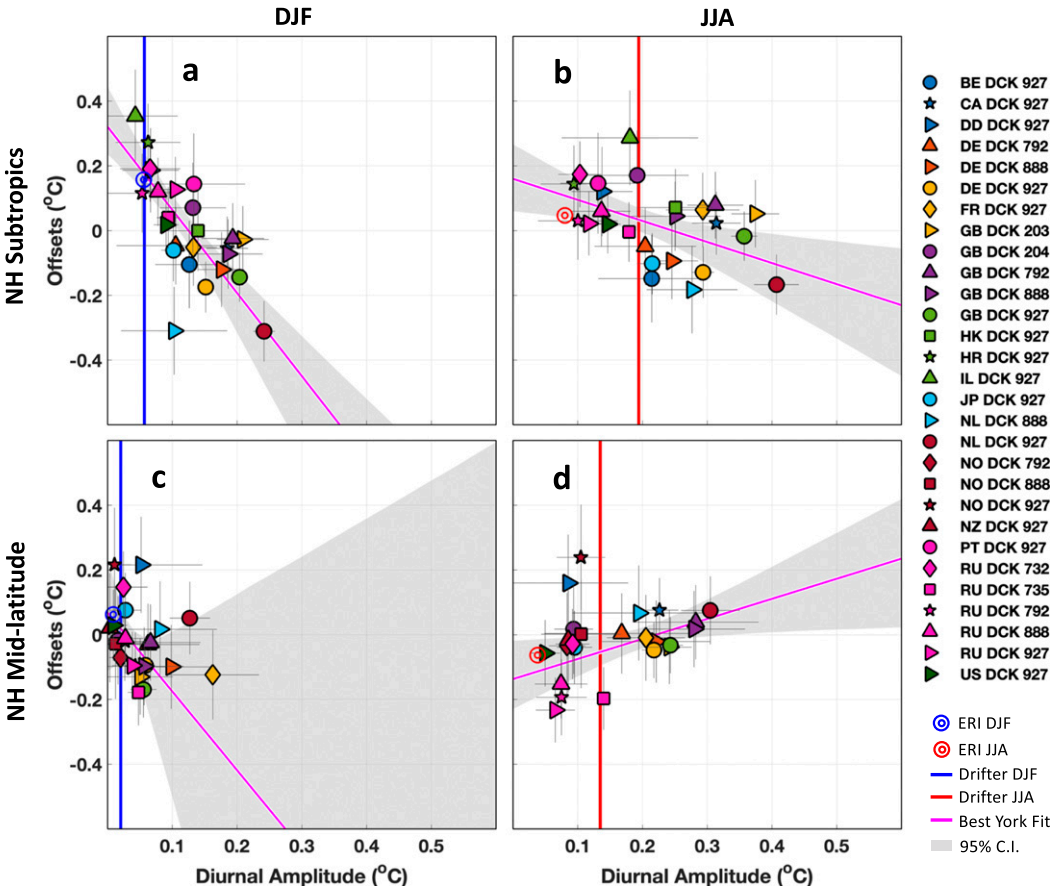

FIG. 7. Diurnal cycles and groupwise SST offsets outside the tropics: The Northern Hemisphere subtropics show (a) strong negative covariance in the winter (DJF) and (b) a larger range of diurnal amplitudes but weaker covariance during summer (JJA). The Northern Hemisphere midlatitudes show (c) a similar pattern but also a smaller range of diurnal amplitudes during winter, consistent with weak diurnal variations in insolation, and (d) a positive scaling during summer, indicative of greater solar heating during the day leading to warming and increased diurnal amplitudes (Fig. 3c). Results are for 1970–89. Regression slopes intersect the offset and diurnal amplitude associated with ERI measurements (double circles) within uncertainties. Approximately one-third of the groups show diurnal amplitudes during summer that are smaller than are found in drifter SST data (vertical lines).

contrast, the estimated amplitude of the diurnal cycle is significantly smaller than reported by buoy and drifter observations for approximately 20% of all nation-deck groups since 1930 (Figs. 6c and 8c). Significance is defined such that the 95% confidence interval of amplitude estimates does not contain the climatological estimates of drifting buoys. At the same time, none of the parameters explored with respect to our bucket model lead to a diurnal amplitude smaller than that of SSTs from drifters, whether considered individually or in combination, except for misclassification of ERI measurements (Figs. 3 and 4). Groups having the smallest diurnal amplitudes are also associated with the warmest offsets. In the tropics, groups having a diurnal amplitude that is significantly ($P < 0.05$) smaller than that of drifter observations are, on average, 0.15°C warmer than groups that have a diurnal amplitude that is significantly larger. Since 1950, most Russian decks and U.S. deck 927 appear especially likely to be composed predominantly of ERI measurements as judged from their anomalous warmth and small amplitudes.

Fourth, York regressions of offset versus amplitude across groups of bucket measurements generally intersect the offset and diurnal amplitude independently determined for ERI values (e.g., Figs. 5 and 7). These intersections are consistent within the 95% confidence intervals for 17 of the 20 combinations of regions, seasons, and independent 20-yr intervals since the 1930s. Such consistency of slopes and ERI values suggests that the major axis of variation across all other groups is consistent with an admixture of varying amounts of ERI data. As noted, negative covariance between amplitudes

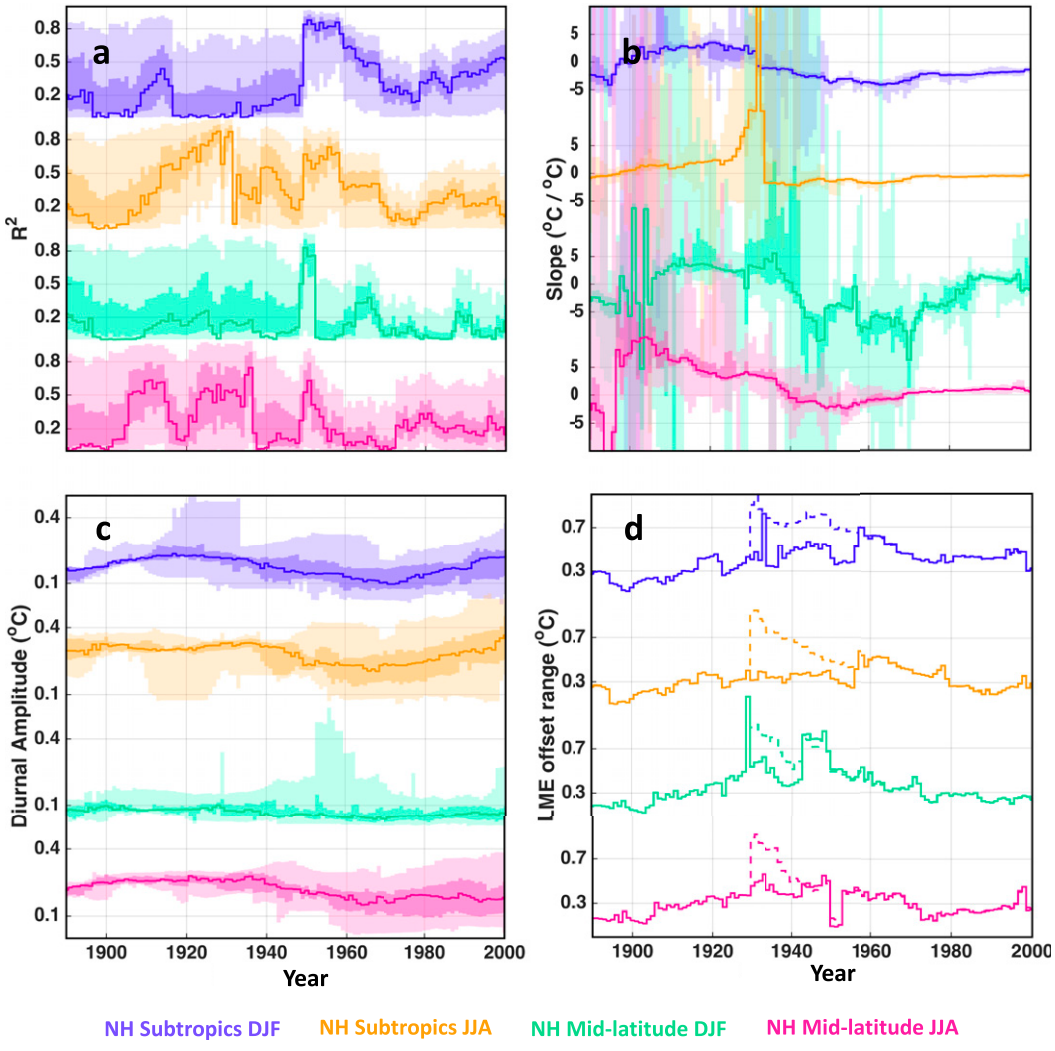

FIG. 8. Evolution of groupwise offsets and diurnal amplitudes outside the tropics: Individual panels are as in Fig. 6, but for different region and season combinations outside the tropics.

and offsets can be associated with multiple different parameters in our bucket model, but given that ERI measurements seems necessary to explain the very small diurnal amplitudes of certain groups, parsimony suggests invoking variable mixtures of ERI data to explain overall regression behavior.

Last, between 1960 and 1980, the offset–amplitude slope gradually becomes less negative across all regions and seasons (Figs. 6b and 8b), shifting from approximately −2° to −0.5°C °C^{-1} in the tropics. Kennedy et al. (2019) identified a gradual decrease in ERI biases over this interval by comparing with the uppermost temperature measurements from XBT and CDT profiles. Under our hypothesis of major intergroup offsets reflecting mixing with ERI measurements, less ERI warming is expected to make the offset–amplitude slope less negative (Figs. 6b and 8b). A related prediction associated with a diminishing ERI bias is for the range of mean offsets to decrease. We examine the 25th to 100th percentile range of offsets because, whereas ERI data are generally near the warmest offset, the lowest offsets could represent noise or outlier behavior. In the tropics, the 25th–100th range is 0.50°C in 1950–69 and then decreases to 0.27°C and to 0.10°C in 1970–89 and 1990–2009, respectively (Figs. 5a–c and 6d). Increases in bucket insulation associated with switching from canvas to rubber buckets may also contribute to the smaller range during more recent intervals (Kennedy et al. 2011).

There are two further features of the data that merit further comment. Whereas misclassification of ERI measurements is generally expected to lead to offsets becoming more negative with increasing diurnal

amplitude (Fig. 3), this pattern appears to be contradicted by the positive scaling of midlatitude data during summertime (Fig. 7d). A reversal in slope can occur, however, if ERI measurements have a smaller bias, as anticipated if seawater temperature is already closer to engine room temperature (Kent et al. 2017), and if bucket measurements are more warmly biased, as anticipated during midlatitude summer on account of increased air temperature, humidity, and insolation.

The second issue is that the average diurnal cycle associated with bucket measurements is approximately 20% smaller in 1970–89 than in 1990–2009 for nearly all regions and seasons (Fig. 2). Such a change in the amplitude of the diurnal cycle could reflect some combination of changes in the physical environment and changes in measurement practices. We first consider, and discard, several potential physical effects. Wind stilling could decrease vertical mixing and increase diurnal amplitudes, but trends in ICOADS2 wind data show a steady increase of approximately $0.2\,\mathrm{m\,s^{-1}}$ per decade between 1982–2002 (Thomas et al. 2008), opposite to the trend needed to explain changes in diurnal amplitude. Changes in surface winds using reanalysis (Thomas et al. 2008; Vautard et al. 2010) and island stations (Vautard et al. 2010) are heterogeneous. Similarly, global warming has been suggested to result in greater nighttime than daytime temperature warming because of increased clouds and water vapor (Easterling et al. 1997), and this would also give a trend in the opposite direction of that needed to explain the observed increase in amplitude. Solar brightening since the 1990s, following a dimming period from the 1950s to 1980s (Wild 2009), is qualitatively of the correct sign but the trend over the ocean is estimated to only be 1%–2% per decade from the 1980s to the 2000s (Pinker et al. 2005; Hatzianastassiou et al. 2005), too small to account for the observed 20% change in diurnal amplitude.

An alternative set of explanations for changes in diurnal amplitudes comes from potential changes in measurement practices. One possibility is that the difference in diurnal amplitudes between 1970–89 and 1990–2009 results from differential sampling of spatially variable diurnal amplitudes (Kennedy et al. 2007), but a parallel analysis using anomalies in diurnal amplitude relative to climatologies developed from drifting buoys (Morak-Bozzo et al. 2016; Chan and Huybers 2019) leads to a similar interdecadal discrepancy in amplitude. We favor an explanation that relates to the misclassification of ERI measurements. Whereas our quantitative FP95d bucket model overestimates diurnal amplitudes by approximately 20% between 1970–89 when averaged over all combinations of regions and seasons, it produces amplitudes that agree with observations during 1990–2009, consistent with approximately 30% of measurements identified as coming from buckets actually being misclassified ERI measurements during the earlier interval. Our inferred 30% misclassification is higher than the 10% estimated by Carella et al. (2018) but consistent with the estimate by Kennedy et al. (2019). Last, diurnal amplitudes from drifting buoys show insignificant trends since the 1980s, supporting the inference that recent changes in diurnal amplitudes represent improvements in the cataloging of ERI versus bucket measurements.

5. Further discussion and conclusions

It appears that the majority of intergroup variability after the 1930s can be explained as arising from varying proportions of ERI data being mixed into groups otherwise considered as coming from buckets. Although some of the covariance between offsets and amplitudes almost certainly arises because of intergroup variations in bucket measurement characteristics, we are not aware of any bucket parameter (Fig. 3), or combination thereof (Fig. 4), that under plausible modification would explain so much of the intergroup variability. In particular, the lower-end range of diurnal amplitude and upper-end range of offsets strongly suggest ERI measurements, and the fact that slopes intersect this end-member since the 1930s suggests pervasive contamination of observations previously thought to come from buckets instead being ERI measurements (Figs. 5 and 6). Misclassification of ERI measurements is thus offered as the primary explanation for intergroup offsets after the 1930s.

In addition to misclassification of ERI data, additional intergroup variations from bucket design or measurement protocols are almost certainly present. Prior to 1930, the offset–amplitude relationship appears largely orthogonal to that found afterward, when ERI data become available. Positive covariance between offsets and amplitudes possibly results from variations in apparent wind or solar absorption (e.g., FP95; Kent et al. 2017), and variations in offsets that occur without changes in amplitude may result from data management errors, such as the truncation of Japanese Kobe collections (Chan et al. 2019). We speculate that bucket data are consistently uncertain across all examined time periods but becomes additionally uncertain with the advent of ERI data in the 1930s and its potential misclassification (Fig. 6d).

To further evaluate if the introduction of ERI data is a sufficient explanation for increased variation across groups, we examine the interquartile range (IQR) of intergroup offsets in relation to the diurnal amplitude. Prior to 1930, the mean IQR of tropical offsets is 0.11°C, and after 1930 the mean IQR of offsets nearly doubles to 0.21°C (Fig. 9). Furthermore, IQR variations are large, peaking at values near 0.4°C in the 1940s and 1960s.

Under the assumption that variations in the diurnal amplitude across groups reflect the proportion of ERI measurements, we regress out the diurnal amplitude component of intergroup offsets to obtain an estimate of intergroup offsets absent ERI influences. The residual IQR before 1930 is essentially unchanged because of low covariance between amplitude and offsets, but after 1930 it drops to 0.13°C (Fig. 9) and stabilizes such that values in the 1940s and 1960s are consistent with the long-term average. The general stability of residual IQR supports the presence of ERI measurements being a sufficient explanation for the excess variability in offsets after 1930. Similar results are obtained if root-mean-square variability is instead used to quantify intergroup variations.

There are several potential extensions of the analysis and results presented here. First, useful information might also be extracted from the phase of the diurnal cycle. An examination of phase information for each group, however, shows close correspondence with amplitudes such that, beyond offering a check on our inferences, little additional information appears available. We have therefore focused exclusively on amplitude in this study, but note that China deck 781 has a reasonable diurnal shape and amplitude but a phase that is evidently shifted by 8 h, possibly because of incorrectly recording Beijing time as Greenwich time. It may also be useful to examine whether offsets exist among groups of ERI measurements, potentially because of misclassification of bucket measurements. Data indicated as coming from ERI, however, appear to be more accurately determined (Carella et al. 2018).

As a final consideration for further analysis, there appears potential for better identifying misclassified ERI data using both offsets and diurnal amplitudes. By way of example, if the criteria in Carella et al. (2018) are applied to subsets of measurements that are indicated as coming from buckets in ICOADS-SI or WMO No. 47, some SST groups (e.g., from German deck 888 and Japanese deck 927) would be classified as 100% bucket measurements in certain decades on the basis of diurnal amplitudes being insufficiently small to conclusively indicate ERIs, but our results help confirm the presence of ERI data because these groups are also offset toward warmer temperatures (Fig. 5b). Quantitative estimates of the fraction of ERI data misclassified within a group would benefit from ascertaining the offset and amplitude associated with a purely bucket end-member, although such end-member values may be expected to vary across groups because of differences in bucket and measurement characteristics.

Alternatively, it may be possible to examine the distribution of offsets and diurnal characteristics within individual groups to better ascertain its composition.

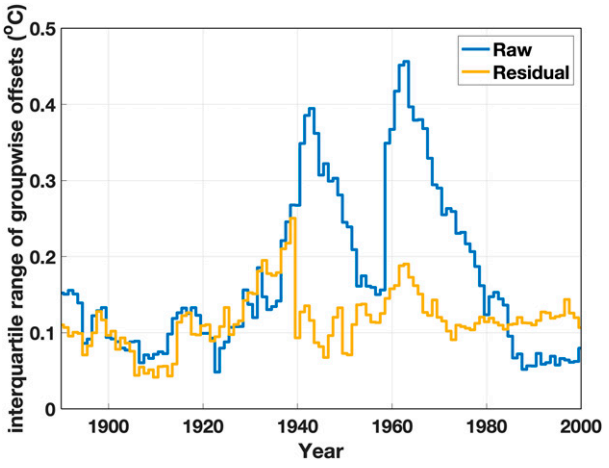

FIG. 9. Interquartile range of groupwise offsets in the tropics: The interquartile range of offsets across groups increases after 1930 (blue curve; from symbols in Fig. 5). If the component that linearly covaries with diurnal amplitude (cf. magenta lines in Fig. 5) is first removed, the interquartile range of the resulting residuals is more stable (yellow curve).

For example, negative skewness of the distribution of amplitudes among individual ships is expected if there is a minority of ERI measurements in the group, and increased kurtosis is expected if the group is equally composed of ERI and bucket measurements. Such an undertaking, however, awaits a better developed model of noise characteristics associated with individual measurements and ship tracks.

Our primary finding is that previously identified offsets among groups of SST data (Chan and Huybers 2019) are attributable to misclassification of ERI data. Other sources of variability prior to the introduction of ERI measurements in 1930, as well as after 1930 once offsets that are attributable to ERI misclassification are removed, appear to be consistent with physical contributions associated with differences in bucket design and measurement protocol. Errors associated with data truncation (Chan et al. 2019) or other record-keeping issues appear to be the exception, as opposed to a predominant source of intergroup offsets. Covariance between amplitudes and offsets and its systematic change in accord with historical variations in measurement techniques also support the credibility of the linear-mixed-effect method for identifying average offsets between groups of SST data. There remain major uncertainties associated with SST reconstructions that potentially have first-order implications for our understanding of internal and externally forced variations in climate (Davis et al. 2019). Continued work to identify the sources of these uncertainties is important for improving the accuracy and interpretation of historical SST estimates.

TABLE A1. Sensitivity of tropical regressions to alternative analysis configurations. Regression are from a York fit of offsets vs diurnal amplitude in the tropics (see Fig. 5). The 95% confidence intervals are from bootstrapping individual groups and are given in parentheses. Columns show configurations using the standard analysis described in section 2, WMO No. 47 in preference to ICOADS-SI, no measurements inferred to come from bucket SSTs, grouping data only according to country information, and using anomalies in diurnal amplitudes relative to a 1990–2014 climatology from drifting buoys.

	Standard	Prefer WMO No. 47	Exclude inferred buckets	Nation-level	Amplitude anomaly
1990–2009	−0.56 (−0.99, 0.11)	−0.40 (−0.81, 0.14)	−0.53 (−0.96, 0.13)	−0.54 (−1.24, 0.21)	−0.56 (−0.99, 0.10)
1970–89	−1.03 (−1.38, −0.64)	−0.94 (−1.29, −0.51)	−0.95 (−1.30, −0.55)	−1.04 (−1.44, −0.57)	−1.07 (−1.40, −0.70)
1950–69	−2.29 (−3.30, −1.65)	−2.30 (−3.30, −1.65)	−2.61 (−4.27, −1.59)	−1.87 (−3.25, −1.44)	−2.31 (−3.25, −1.72)
1930–49	−2.88 (−23.1, −0.80)	−2.88 (−23.6, −0.80)	−3.20 (−23.4, −0.87)	−2.55 (−34.0, 0.11)	−2.77 (−24.7, 0.60)
1910–29	3.12 (−0.23, 15.5)	3.15 (−0.23, 15.5)	3.13 (−0.23, 15.5)	17.9 (−91.6, 95.8)	1.11 (−31.2, 12.1)
1890–1909	−0.22 (−13.3, 6.25)	−0.22 (−13.3, 6.28)	−0.22 (−13.4, 6.27)	−1.55 (−33.1, 47.2)	−0.29 (−3.82, 2.54)

Acknowledgments. We thank Elizabeth Kent (National Oceanography Centre) for providing ship-track data and useful discussion, as well as two anonymous reviewers for comments. Support was provided by the Harvard Global Institute.

APPENDIX A

Sensitivity Analysis

The analysis choices that we make are appropriate and plausible but not always unique, such that it is useful to summarize the sensitivity of our results to alternative, plausible formulations. Specifically, we explore how identifying and grouping of bucket SSTs and the computation of diurnal amplitudes influence our results. All comparisons are relative to the approach described in section 2.

a. Identifying bucket SSTs

WMO No. 47 metadata from 1965 onward disagree with ICOADS-SI for 8.2% of bucket measurements indicated by ICOADS-SI. In the standard analysis, we use ICOADS-SI when the two metadata sets disagree. To examine the implication of this choice, we rerun the analysis prioritizing WMO No. 47 metadata and obtain results that are indistinguishable within uncertainties. For example, the slope is −1.03 ([−1.38, −0.64], 95% confidence interval) in the tropics in 1970–89 in the standard analysis and is −0.94 ([−1.29, −0.51]) when WMO No. 47 metadata are preferred (Table A1).

A second choice involves the inclusion of measurements that are inferred to come from buckets, as opposed to a direct indication from ICOADS-SI or WMO No. 47. The inference method follows Kennedy et al. (2011) and leads primarily to the inclusion of Russian decks. Specifically, 75% of inferred bucket SSTs come from Russian decks 732, 888, and 927, but whose status as coming from buckets is questionable on the basis of mean offsets and diurnal amplitudes that are more consistent with ERI observations (Fig. 5). Excluding inferred bucket SSTs reduces the percentage of Russian measurements that are otherwise inferred to be bucket measurements from 49% to 16%. Results of this altered analysis are, again, statistically indistinguishable from our main analysis. We note that 16% of all Russian SST measurements are identified as coming from buckets using ICOADS and WMO metadata, and that all of these appear consistent with ERI measurements. These putative Russian bucket measurements come from decks 732, 735, 792, 888, and 927.

b. Grouping by country

We divide groups according to both deck and country information. Although decks do not necessarily have physical implications, we resolve our groupings according to decks in the main line of analysis because Chan and Huybers (2019) found that decks coming from the same nations often exhibited statistically significant offsets from one another. Furthermore, Chan et al. (2019) found a truncation bias that happened only to decks 118 and 119 among Japanese ships. Deck divisions can, however, be arbitrary. For example, data from a single ship track can be divided into several decks (Carella et al. 2017a). We therefore rerun our analysis grouping only according to country information. After the 1930s, results from grouping only according to country are statistically consistent with our main analysis. Statistical significance does not decrease because York fit regressions (Fig. 5) account for the greater confidence we have in averaged data. Before the 1930s, negative slopes disappear as in the standard analysis, but the small number of national groups before the 1930s and smaller range of available data impinge upon the ability to estimate slopes, such that uncertainties increase.

c. Diurnal amplitude

Changes in offsets are compared with diurnal amplitudes, as opposed to anomalies in diurnal amplitudes relative to a local climatology, leading to the concern that groupwise differences in diurnal amplitude could reflect differences in geographic distribution, as opposed to measurement

characteristics. We examine the sensitivity of our results to using amplitude anomalies relative to a climatological amplitude estimated from drifting buoys. We test this sensitivity using buoy estimates from both Morak-Bozzo et al. (2016) and Chan and Huybers (2019). Again, results are found to be statistically indistinguishable (Table A1). In principle, removing spatial heterogeneity in diurnal amplitudes is expected to decrease noise, but, in practice, the average climatological diurnal cycles for individual groups are very similar to one another, such that overall changes are small.

Sampling frequency gradually increases from once per 6 h to as much as hourly after the Second World War, which in principle could influence estimates of diurnal amplitudes, but our methods are robust to such effects on two accounts. First, we pool SST anomalies across different longitude bands such that, in any given year between 1850 and 2014, each of the 24 local hours has samples for most ship groups. We also weigh harmonic fits by numbers of measurements in each bin to account for heteroscedastic errors. Second, if we sample every 6 h from an hourly-resolved diurnal cycle (e.g., 1990–2009 climatology of all bucket SSTs in the tropics; Fig. 2a), the standard error in the amplitude of first-order harmonics is only 0.01°C, which is small relative to 0.3°C range of differences among groups. We, therefore, conclude that our results are robust to increased sampling frequency.

APPENDIX B

Extended Folland and Parker (1995) Bucket Model

The standard FP95 bucket model represents daily mean quantities. We extend FP95 to include diurnal effects associated with insolation, SST, winds, and relative humidity.

a. Solar scheme

We model the total insolation absorbed by the top of a bucket as

$$Q^{\text{top}} = (1-a)(1-s)Q_g \pi r^2, \tag{B1}$$

where a is the albedo of bucket materials, s is the percentage of shaded insolation, and r is bucket radius; Q_g is the sum of direct and diffuse radiation at the ocean's surface after accounting for scattering and reflection and is diagnosed as a function of location, month, and local hour from ERA-interim reanalysis. Specifically, Q_g is computed from 1985–2014 3-hourly ERA-Interim data (Dee et al. 2011) and interpolated to hourly resolution.

Direct and diffuse insolation are modeled separately for bucket walls because of differential absorption. Because a partition between direct and diffuse radiation is not available from ERA-interim reanalysis (Dee et al. 2011), a segmented linear model is used to estimate the fraction of direct radiation F (Spitters et al. 1986),

$$F = \begin{cases} 0 & \text{if } \dfrac{Q_g}{Q_0} \le 0.35 \\ 2\dfrac{Q_g}{Q_0} - 0.7 & \text{if } \dfrac{Q_g}{Q_0} > 0.35 \end{cases}, \tag{B2}$$

where Q_0 is incoming solar radiation at the top of the atmosphere. Values of Q_g/Q_0 below 0.35 are assumed to have complete cloud coverage. Incoming solar radiation is approximated as

$$Q_0 = Q_s \left[1 + 0.033 \cos\left(2\pi \frac{t_d}{365}\right)\right] \cos(\theta), \tag{B3}$$

where Q_s is the solar constant (1370 J m^{-2} s^{-1}), t_d is day of the year, and the first cosine function accounts for Earth's eccentric orbit. Sun zenith θ is computed following Reda and Andreas (2004).

Heating on bucket walls from direct insolation is

$$Q^{\text{wall\_dir}} = (1-a)(1-s)Q_g F \tan(\theta) 2rh, \tag{B4}$$

where h is bucket height. The term $\tan(\theta)$ gives the horizontal component from downward insolation, and $2rh$ is the area of the vertical cross section of a bucket. Diffuse insolation is assumed to come equally from the overhead hemisphere:

$$Q^{\text{wall\_diff}} = (1-a)(1-s)Q_g (1-F)\pi rh. \tag{B5}$$

Note that the area of bucket walls absorbing diffuse insolation is $2\pi rh$ but, given the assumed hemispheric radiation, the diffuse energy flux onto a vertical surface is only half that onto a horizontal surface.

Summing direct and diffuse components at the top and sides gives total absorbed radiation:

$$\begin{aligned} Q_{\text{tot}} &= Q^{\text{top}} + Q^{\text{wall\_diff}} + Q^{\text{wall\_dir}} \\ &= (1-a)(1-s)Q_g[\pi r^2 + F\tan(\theta)2rh + (1-F)\pi rh]. \end{aligned} \tag{B6}$$

b. Other environmental forcing

Hourly-resolved environmental fields are incorporated as a function of 5° grid boxes and month. SSTs are initialized using buoy and drifter measurements that are assumed as unbiased actual SSTs. Specifically, diurnal

TABLE B1. ICOADS metadata for identifying buoy and drifter measurements.

ICOADS metadata name	Metadata values
ID indicator	3, 4, 11
Source ID	24, 55, 50, 61, 62, 63, 66, 86, 87, 117, 118, 120, 121, 122, 139, 147, 169, 170
Deck	143, 144, 146, 714, 734, 793, 794, 876, 877, 878, 879, 880, 881, 882, 883, 893, 894, 993, 994, 235
Platform	6, 7, 8

anomalies are diagnosed from the 1990–2014 quality-controlled buoy and drifter observations (Chan and Huybers 2019) assuming that they are bias-free with respect to diurnal cycles of SSTs. Buoy and drifter observations are identified using the ICOADS identifier (ID) indicator, source ID, platform, and deck information (see Table B1). For each buoy in each day, SST anomalies relative to the daily mean are computed and binned by 5° latitude bands and seasons for shapes of diurnal cycles, which are normalized to have a mean of zero and range of one. The amplitude of the predetermined diurnal shapes is evaluated for each buoy in each day using least squares and averaged to 5° grids (Chan and Huybers 2019).

To represent the environment in which bucket SSTs are measured, the diurnal cycles of air temperature, dewpoint temperature, and wind are calculated using measurements from ships taking bucket SSTs between 1970 and 2009. Measurements that are considered low quality (i.e., having a National Climatic Data Center quality control flag larger than 5) are excluded. Unlike for SST estimates, both tracked and untracked ships are used to estimate the diurnal cycles of environmental forcing because ship reports are too sparse to map reliable and spatially complete forcing fields. For each month, all data are first averaged to hourly-resolved 5° grids and then fit with predetermined diurnal shapes using least squares, similar to the approach of Kennedy et al. (2007). Diurnal cycles shapes are determined for each month and 5° latitude band by averaging diurnal anomalies from tracked ships taking bucket measurements, and fits are weighted by sample sizes in individual bins.

Diurnal variations are summed with the 1973–2002 climatology diagnosed from the NOCSv2.0 monthly dataset (Berry and Kent 2009) to provide a diurnally resolved climatology. Shipboard air temperatures are treated specially, however, because daytime heating of ship decks causes air temperature to have larger diurnal variations than either SSTs or ambient marine surface air temperatures (Berry et al. 2004). Berry and Kent (2009) correct for excessive daytime heating of shipboard air temperatures by assuming that differences in the diurnal variation of ambient marine air temperature and SST are negligible. Our interest is in the conditions aboard a ship, however, as opposed to ambient marine air temperatures. Thus, following Berry and Kent (2009), we assume that ambient air temperature and shipboard temperatures are equivalent during nighttime, and that ambient air temperature is equivalent to SST but with a mean offset given by NOCSv2.0. Under these assumptions, we are able to specify a mean value for shipboard diurnal variations in air temperature by shifting average nighttime air temperature anomalies to equal that of nighttime SST anomalies and then subtracting the daily-mean difference between SST and ambient air temperatures. Note that the diurnal amplitude of shipboard air temperatures generally exceeds that of SSTs but that shipboard air temperatures are generally cooler than SSTs during nighttime, making whether shipboard air temperatures are greater than SSTs during daytime a function of region and season.

REFERENCES

Armour, K. C., C. M. Bitz, and G. H. Roe, 2013: Time-varying climate sensitivity from regional feedbacks. *J. Climate*, **26**, 4518–4534, https://doi.org/10.1175/JCLI-D-12-00544.1.

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.

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

——, ——, and P. K. Taylor, 2004: An analytical model of heating errors in marine air temperatures from ships. *J. Atmos. Oceanic Technol.*, **21**, 1198–1215, https://doi.org/10.1175/1520-0426(2004)021<1198:AAMOHE>2.0.CO;2.

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

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

——, A. Morris, R. Pascal, M. Yelland, D. Berry, S. Morak-Bozzo, C. J. Merchant, and E. Kent, 2017b: Measurements and models of the temperature change of water samples in sea-surface temperature buckets. *Quart. J. Roy. Meteor. Soc.*, **143**, 2198–2209, https://doi.org/10.1002/qj.3078.

——, 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.

Chan, D., and P. Huybers, 2019: Systematic differences in bucket sea surface temperature measurements among nations identified

using a linear-mixed-effect method. *J. Climate*, **32**, 2569–2589, https://doi.org/10.1175/JCLI-D-18-0562.1.

——, E. C. Kent, D. I. Berry, and P. Huybers, 2019: Correcting datasets leads to more homogeneous early-twentieth-century sea surface warming. *Nature*, **571**, 393–397, https://doi.org/10.1038/s41586-019-1349-2.

Cowtan, K., R. Rohde, and Z. Hausfather, 2018: Evaluating biases in sea surface temperature records using coastal weather stations. *Quart. J. Roy. Meteor. Soc.*, **144**, 670–681, https://doi.org/10.1002/qj.3235.

Davis, L. L., D. W. Thompson, J. J. Kennedy, and E. C. Kent, 2019: The importance of unresolved biases in twentieth-century sea surface temperature observations. *Bull. Amer. Meteor. Soc.*, **100**, 621–629, https://doi.org/10.1175/BAMS-D-18-0104.1.

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.

Easterling, D. R., and Coauthors, 1997: Maximum and minimum temperature trends for the globe. *Science*, **277**, 364–367, https://doi.org/10.1126/science.277.5324.364.

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.

Hatzianastassiou, N., C. Matsoukas, A. Fotiadi, K. Pavlakis, E. Drakakis, D. Hatzidimitriou, and I. Vardavas, 2005: Global distribution of Earth's surface shortwave radiation budget. *Atmos. Chem. Phys.*, **5**, 2847–2867, https://doi.org/10.5194/ACP-5-2847-2005.

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, 2017: Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Climate*, **30**, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1.

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

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

Kennedy, J. J., P. Brohan, and S. F. B. 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. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011: 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.

——, ——, C. Atkinson, and R. Killick, 2019: An ensemble data set of sea surface temperature change from 1850: The Met Office Hadley Centre HadSST. 4.0.0.0 data set. *J. Geophys. Res. Atmos.*, **124**, 7719–7763, https://doi.org/10.1029/2018JD029867.

Kent, E. C., 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.

——, and Coauthors, 2017: A call for new approaches to quantifying biases in observations of sea surface temperature. *Bull. Amer. Meteor. Soc.*, **98**, 1601–1616, https://doi.org/10.1175/BAMS-D-15-00251.1.

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 Rep., https://babel.hathitrust.org/cgi/pt?id=uc1.$c188170&view=1up&seq=7.

Lane, D., D. Scott, M. Hebl, R. Guerra, D. Osherson, and H. Zimmer, 2013: *Introduction to Statistics: An Interactive e-Book*. Self-published, https://books.apple.com/us/book/introduction-to-statistics/id684001500.

Met Office, 1956: *Handbook of Meteorological Instruments Part I: Instruments for Surface Observations*. Cambridge University Press, 458 pp.

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.

Pinker, R., B. Zhang, and E. Dutton, 2005: Do satellites detect trends in surface solar radiation? *Science*, **308**, 850–854, https://doi.org/10.1126/science.1103159.

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.

Reda, I., and A. Andreas, 2004: Solar position algorithm for solar radiation applications. *Sol. Energy*, **76**, 577–589, https://doi.org/10.1016/j.solener.2003.12.003.

Spitters, C., H. Toussaint, and J. Goudriaan, 1986: Separating the diffuse and direct component of global radiation and its implications for modeling canopy photosynthesis Part I. Components of incoming radiation. *Agric. For. Meteor.*, **38**, 217–229, https://doi.org/10.1016/0168-1923(86)90060-2.

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

Ting, M., Y. Kushnir, and C. Li, 2014: North Atlantic multidecadal SST oscillation: External forcing versus internal variability. *J. Mar. Syst.*, **133**, 27–38, https://doi.org/10.1016/j.jmarsys.2013.07.006.

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.

Vecchi, G. A., M. Zhao, H. Wang, G. Villarini, A. Rosati, A. Kumar, I. M. Held, and R. Gudgel, 2011: Statistical–dynamical predictions of seasonal North Atlantic hurricane activity. *Mon. Wea. Rev.*, **139**, 1070–1082, https://doi.org/10.1175/2010MWR3499.1.

Wild, M., 2009: Global dimming and brightening: A review. *J. Geophys. Res.*, **114**, D00D16, https://doi.org/10.1029/2008JD011470.

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

Yeh, S.-W., J.-S. Kug, B. Dewitte, M.-H. Kwon, B. P. Kirtman, and F.-F. Jin, 2009: El Niño in a changing climate. *Nature*, **461**, 511–514, https://doi.org/10.1038/nature08316.

York, D., N. M. Evensen, M. L. Martınez, and J. De Basabe Delgado, 2004: Unified equations for the slope, intercept, and standard errors of the best straight line. *Amer. J. Phys.*, **72**, 367–375, https://doi.org/10.1119/1.1632486.