Systematic differences in bucket sea surface temperatures caused by misclassification of engine room intake measurements

DUO CHAN [1] *   PETER HUYBERS [1]

[1] *Department of Earth and Planetary Sciences, Harvard University, Cambridge, USA*

**Corresponding author address:* Duo Chan, Department of Earth and Planetary Sciences, Harvard University, 20 Oxford St., Cambridge, MA 02138, USA.

E-mail: duochan@g.harvard.edu

Generated using v4.3.2 of the AMS LaTeX template

ABSTRACT

Differences in sea surface temperature (SST) biases among groups of bucket measurements in the International Comprehensive Ocean-Atmosphere Data Set 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. Analyzing ICOADS3.0 SST measurements inferred to come from buckets indicates that offsets after the 1930s primarily result from the misclassification of ERI measurements in points of four 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 subsequently stronger and generally negative. (2) The introduction of ERI measurements in the 1930s is accompanied by a wider range of offsets and diurnal amplitudes across groups, with 20% of estimated diurnal amplitudes being significantly smaller than buoy and drifter observations. (3) Regression of offsets versus diurnal amplitudes intersect independently determined end-member values of ERI measurements. Finally, (4) offset-amplitude slopes become less negative across all regions and seasons between 1960 to 1980, when ERI temperatures were independently determined to become less warmly biased. These results highlight the importance of accurately determining measurement procedures in order to correct for biases and reduce 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 the El Nino 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, multi-decadal 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) that can be divided into physical and non-physical categories. Physical processes are defined as those causing difference between the temperature of measured water and that at the surface of the ocean, and are generally related to solar heating and evaporative and sensible cooling. The temperature of the ’surface’ of the ocean is typically taken as the bulk average over the upper several meters (Kennedy 2014; Kennedy et al. 2019). The relative contributions to heating and cooling of a bucket will depend upon bucket characteristics, environmental conditions, and measurement protocols (Ashford 1948; Folland and Parker 1995). Non-physical 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 of 0.45°C (Chan et al. 2019).

There are widely-used methodologies 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 night-time 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 -1 to +0.1 °C and for the global average ranges from -0.05 to 0.45°C (99% uncertainty range in 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 foregoing methods for correcting bucket temperatures is difficulty in accounting for regional changes in measurement details. For example, during 1900-1913, most SST measurements in the South Pacific and South Atlantic come from German compilations, averaging 231,000 measurements per year. However, from 1914-1920, contributions from German compilations drop off to 38,000 measurements per year, and the U.K. 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-effects model to detect offsets among groups of SST observations. The range of corrections is -1.0 to 1.3°C at the level of individual grid boxes (Chan et al. 2019) and, 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 1908-1941 changes from 0.31 to 0.56 °C per 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 accounted 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 meta-data 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 nearly 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 engine room intakes. They inferred nearly 100% accuracy after the 1990s but that approximately 10-20% of the bucket measurements available between the 1930s to 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 offsets relative to the amplitude of diurnal cycle is developed for the purpose of further exploring the origins and implication of offsets among groups of SST observations. After introducing data and methodology, 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 then diagnose offset-amplitude relationships from ICOADS3.0 bucket measurements and consider physical and non-physical contributions to SST offsets.

2. Data and methods

In-situ SST observations used in this study are from ICOADS3.0 (Freeman et al. 2017). Initial quality control, identification of bucket measurements, and removal of long-term climatology all follow Chan and Huybers (2019) and Chan et al. (2019). Our analysis also makes use of recent estimates of individual ship tracks (Carella et al. 2017) that are available for 82% of bucket measurements between 1880-2009 in ICOADS3.0. We perform analyses for 20-year periods starting from 1880-1899 and being slid forward annually until 1990-2009. Ship tracks are not available after 2009 (Carella et al. 2017), but neither are bucket measurements as common (Kennedy et al. 2011, 2019).

To intercompare subsets of bucket measurements, we assign groups according to combinations of deck numbers and nations, and those not associated with a nation are combined into a separate group according to deck number (Chan and Huybers 2019; Chan et al. 2019). The ’deck number’ refer 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 decks contain significant offsets ($P<0.1$, Chan and Huybers 2019).

A linear-mixed-effects methodology is used to identify offsets amongst groups of SSTs, accounting for variations across region, season, and year. This method is describe in detail by Chan and Huybers (2019), but several changes are made here. Only observations having valid diurnal anomaly estimates are used in order to base the offset analysis on similar data with diurnal analysis, and groups contributing less than 6,000 pairs in a 20-year period are excluded because resulting offset estimates are nosier than for groups with more data. Decadal variations are not explicitly accounted for because, unlike our previous analyses where we obtained estimates from 1850, the analysis is performed over 20-year intervals.

To explore plausible seasonality in offset-diurnal relationships, we include seasonal effects for DJF, MAM, JJA, and SON over latitude bands between 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 half of a year to account for different seasons between hemispheres. Finally, ERI observations are included in order to empirically constrain offsets relative to bucket temperatures. Our analysis excludes hull sensor SSTs and treats all ERI SSTs as a single group.

Diurnal amplitudes are computed using ICOADS3.0 measurements coming from tracked ships (Carella et al. 2017). For each day of each ship, diurnal anomalies are calculated relative to daily-mean SST (Carella et al. 2017) if there is at least one SST observation in each of four 6-hourly bins starting from local midnight. Diurnal SST anomalies are aggregated by local hours for each nation-deck group as resolved by the linear-mixed-effect intercomparison and are averaged annually for the tropics (20°S-20°N) and seasonally outside the tropics (20-40°N and 40-60°N). Amplitudes of diurnal cycles are obtained by calculating the amplitude of the once-per-day sinusoid where the fitting is weighted by the sample size in each hourly bin. A comparable analysis is performed to estimate diurnal amplitudes of ERI measurements, though these are only available back to the 1930s.

Because both offsets and amplitudes are uncertain, a York fit is used for purposes of estimating trends in offsets as a function of 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 100 times. Although ERI measurements are incorporated in the analysis, only bucket SST groups are used in York regressions. The percentage of intergroup offsets explained by diurnal amplitudes is quantified as the square of their Pearson's correlation coefficient (r^2) and associated confidence intervals are estimated following Lane et al. (2013). Codes for reproducing all results are posted at `https://github.com/duochanatharvard/LME-Offsets-vs-Diurnal-Amplitudes`

3. Model simulations

To develop baseline expectations for variability in mean offsets relative to the amplitude of diurnal cycles in bucket SST measurements, we first examine these features with respect to a wooden bucket model. The model is extended from that of Folland and Parker (1995) to also include the diurnal cycle as described in Appendix A and is referred to as FP95d.

We assume a standard set of parameters that follow Folland and Parker (1995) and are listed in Table 1, but with two exceptions for processes not fully accounted for in their model. Buckets may not be fully equilibrated with SSTs before a water sample is measured, requiring that the percentage of air temperature be specified; and a bucket may be in the shadow of a ship or measured within a sheltered enclosure, requiring the percentage of solar radiation that is absorbed to also be specified (Carella et al. 2018; Kennedy et al. 2019). These two effects are included in FP95d and the associated parameters adjusted in order to minimize the root-mean-square error (RMSE) averaged over all combinations of regions and seasons in 1990-2009, with the best fit coming from initial bucket temperature at the time of collecting seawater representing 20% of on-deck air temperature and 70% of insolation being absorbed.

The FP95d model well-reproduces the observed amplitude, phase, and seasonality of diurnal cycles of bucket SST measurements in 1990-2009 (Fig. 1a-c). In the tropics, the correlation of ICOADS bucket measurements with SSTs from buoys and drifters is already 0.93 (r-value), but once additional information associated with solar heating and latent cooling are incorporated through the bucket model, correlations increase to 0.98. RMSE decreases from an average of 0.12 to 0.04°C (Fig. 1a). Similarly good fits are found for other regions and seasons during 1990-2009. Between 1970-1989, however, FP95d overestimates observed diurnal amplitude by approximately 25% for all combinations of regions and seasons (Fig. 1d-f). This model-data mismatch could be due to systematic changes in bucket types or measurement practices, and better agreement could be achieved by decreasing the contribution of air temperature to the initial bucket temperature to 10% and insolation absorption to 50%. It will be argued in Sec. 4, however, that the better explanation relates to misclassifications of ERI measurements.

Using the 1990-2009 fit to observations, we vary individual model parameters and explore the relationship between biases in bucket SSTs and changes in diurnal cycles. Folland and Parker (1995) highlight 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).

Exposure time: The lower bound on elapsed time between a bucket's extraction from the water and measurement is taken as one minute, consistent with the time needed for hauling a bucket on deck (Folland and Parker 1995), except for perhaps with respect to smaller nineteenth-century ships. Once buckets are brought on deck, FP95 assigned an average on-deck time of four minutes to 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 steady-state to be reached, e.g., Kobe Imperial Marine Observatory (1925), which perhaps ranges out to 10 minutes. We thus explore total exposure times ranging from 1 to 11 minutes.

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 insolation, 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 indicates that canvas model behaviors are roughly reproducible by assuming a 2mm thick wooden bucket of the same size. We therefore, explore a wooden bucket having wall thicknesses ranging between 0.2–2 cm.

Bucket size: Small buckets tend to have a larger surface-area-to-volume ratio and, therefore, exchange heat more efficiently than large buckets (Folland and Parker 1995; 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.

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, giving a mean apparent wind of approximately 5.5 m/s. For an upper bound, we assume a ship under power making 10 m/s into a prevailing wind of 5 m/s, where such ship speed is the approximate upper bound indicated in Fig. 11 of Carella et al. (2017). This upper bound is specified in FP95d by scaling the standard apparent wind by a factor of three. For the lower bound, we assume no wind for an entirely sheltered bucket.

Insolation: Folland and Parker (1995) note limited evidence that bucket measurements were exposed to direct solar radiation on ship decks, and that limited evidence mostly pertaining to 19th century reports. Carella et al. (2018), however, showed excessive diurnal cycles for bucket SSTs that they attributed to solar heating, and Kennedy et al. (2019) gives evidence for strong solar heating over the mid-latitude summer. We explore the full possible range of exposure to insolation from 0% to 100%.

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 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, 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. 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 complication are not accounted for.

Misclassification of ERI measurements: To the foregoing list of physical effects on buckets, we add the non-physical effects 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 non-physical because they stem from incorrectly identifying a data source. ERIs measure deeper and hence colder water than buckets, but warming of water within the engine room leads to temperatures that are generally biased between 0.1–0.3°C warm relative to true SSTs (Kennedy et al. 2011; Kent et al. 2017). The greater depth at which ERI measurements come from also implies a smaller amplitude diurnal cycle (Kawai and Wada 2007; Carella et al. 2018).

Measurement type is inferred for ICOADS3.0 data from an indicator in ship log books (Freeman et al. 2017) or, after 1960, from WMO publication 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 was 60%

between 1960-1980, 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 mis-identified, and we explore scenarios having between 0–100% 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, and the diurnal amplitude and mean biases of temperatures are examined for different combinations of latitude bands and seasons (Fig. 2a-c). Mean bias is computed as the daily-average difference between bucket water temperatures and true SSTs, whereas diurnal amplitude is obtained by fitting a once-per-day sinusoid.

Most parameter variations lead to an anticorrelation between mean temperature biases and diurnal amplitudes (Fig. 2a). 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, except in certain long-daylight, high-intensity cases found during summertime. Similar decreases in mean SST and increases in 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 air temperature onboard a ship responds more strongly than SSTs to the diurnal solar cycle (Berry et al. 2004) and is generally cooler than the sea surface. Misclassification of ERI measurements has the effect of introducing warmer average temperatures and smaller amplitude diurnal cycles into a group, thereby altering offsets and amplitudes along an axis similar to the foregoing properties. A mostly orthogonal response comes from increasing insolation absorbed by a bucket because it gives a larger diurnal amplitude and a higher diurnal average temperature through daytime warming. Finally, a nearly vertical offset-amplitude relationship is associated with varying apparent wind because wind-induced evaporation has almost no apparent diurnal cycles.

Summer and winter exhibit distinct offset-amplitude relationships (Fig. 2b-c). During winter there are weaker diurnal variation in solar 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. In summertime mid-latitude, 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 mid-latitude are generally warmer than SSTs.

4. Observational results

Observational results generally indicate that groups that are offset cold also have a larger diurnal amplitude (Fig. 3). In the tropics, a strong anti-correlation is found between the average offset and the diurnal amplitude among groups over 20-year periods between 1930-2009, with the mean r^2 being 0.51 (Fig. 5a). Predicted negative slopes of offsets as a function of amplitude range from -4.5 to -1.24 °C/°C, and observed slopes similarly range from -4.2 to -0.5 °C/°C (Fig. 5). The range of amplitude and offset values also generally accord with expectations except that the maximum predicted diurnal amplitude is about 0.3°C, whereas observed amplitudes range higher. Netherlands deck 926 has the largest diurnal amplitude, reaching 0.47±0.01°C between 1990-

2009, and is also the greatest outlier in other regions and time intervals, perhaps indicative of buckets residing on ship decks for extended periods prior to measurement.

Subtropical and mid-latitude regions also generally have a strong negative relationship between offset and amplitude after the 1930s. Furthermore, in these regions, it is possible to examine trends during different seasons. In the subtropics, offset-amplitude slopes are more negative in winter than summer (Figs. 4a-b, 5b), as predicted (Fig. 1b). In the extra tropics significantly positive trends are found in summer after the 1970s (Fig. 4d), also consistent with predictions (Fig. 1c). The lowest offset-amplitude correlations and most uncertain York fit slopes are found for the mid-latitude winter (green curves in Fig. 5a-b), which features the smallest diurnal signals.

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 several lines of additional evidence, however, that support misclassification of ERI measurements as the predominate source of inter-group 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. Subsequent to introduction of ERI measurements in the 1930s, offset-amplitude covariance is strong and generally negative (Fig. 3).

Second, both the spread in groupwise offsets and amplitudes are narrower prior to 1930 than after (Fig.5 c and d). The range of diurnal cycle amplitudes before 1930 is particularly small, suggesting that variations in wind-induced evaporation may be an important contribution to changing mean offsets (Fig. 2). Furthermore, diurnal amplitudes center on values that are significantly greater than buoy and drifter SSTs and are consistent with bucket measurements. In contrast, the estimated amplitude of the diurnal cycle is significantly smaller than reported by buoy and drifter observations for 20% of all nation-deck groups since the 1930s (Fig.5 c). None of the parameters explored with respect to our bucket model lead to a diurnal amplitude smaller than that of actual SSTs except for misclassification of ERI measurements. These groups are also associated with the warmest offset that in the Tropics average 0.15°C warmer than decks having diurnal amplitudes significantly greater than buoy and drifter observations. Most Russian decks and US deck 927 since 1950 appear especially likely to be composed predominantly of ERI measurements given their warmth and small amplitudes.

Third, the slopes inferred from York regressions intersect the offset and diurnal amplitude independently determined for ERI values (e.g., Figs. 2 and 3). These intersections are consistent within the 95% confidence intervals for 17 of the 20 combinations of regions, seasons, and independent 20-year intervals since the 1930s. Such consistency of slopes and ERI values suggests that, not only are certain groups likely composed predominantly of ERI values, but that the major axis of variation across all other groups is consistent with an admixture of varying amounts of ERI data.

Finally, between 1960 to 1980, the offset-amplitude slope gradually becomes less negative across all regions and seasons (Fig. 5b and Fig. A1b), shifting from approximately -2 to -0.5 °C/°C 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 inter-group offsets reflecting mixing with ERI measurements, less ERI warming is expected to make the offset-amplitude slope less negative (Fig. 5d). 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 is generally near the warmest offset, the lowest offsets could represent noise or outlier behavior. The 25th-100th range is 0.6°C in 1930-1949 and 1950-1969 and then decreases to 0.4°C and 0.3°C in 1970-1989 and 1990-2009, respectively. 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 other features of the data that require further comment. First, whereas misclassification of ERI measurements is generally expected to lead to offsets becoming more negative with increasing diurnal amplitude (Fig. 2c), this pattern appears to be contradicted by the positive scaling of mid-latitude data during summertime (Fig. 4d). 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 biassed, as anticipated during mid-latitude summer on account of increased air temperature, humidity, and insolation. Second, as noted in Sec. , the average diurnal cycle associated with bucket measurements is 20%-30% larger in 1990-2009 than in 1970-1989 across all regions and seasons (Fig. 1). Such an increase in the amplitude of the diurnal cycle is consistent with a smaller proportion of ERI measurements being misclassified during this interval (Carella et al. 2018; Kennedy et al. 2019).

5. Discussion and conclusions

It appears that the majority of inter-group 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 or combination thereof 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 suggest pervasive contamination. Misclassification of ERI measurements is thus offered as the simplest 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 afterwards, when ERI data becomes available. Positive covariance between offsets and amplitudes possibly result from variations in apparent wind or solar absorption (e.g., Folland and Parker 1995; 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 is consistently uncertain across all examined time periods but becomes additionally uncertain with the advent of the potential for misclassification of ERI data in the 1930s. Such speculation is supported by further analysis of variability in offsets. Prior to 1930, the mean standard deviation of tropical offsets is 0.09°C. After 1930, the mean standard deviation of offsets increases to 0.15°C, but if the component of offset variations that linearly depends on diurnal amplitude is first regressed out, residual standard deviation averages 0.10°C (Fig. 6). Assuming that the regression removes contributions from ERI misclassification, pre-1930 and post-1930 variations in offsets similarly correspond to bucket data and are of a consistent magnitude.

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 focussed 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 hours, possibly because of incorrectly recording Beijing time as Greenwich time. It may also be useful to also 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 final consideration for further analysis, there appears potential for better identifying misclassified ERI data using both offsets and diurnal amplitudes. By way of example, Carella et al. (2018) classify German deck 888 and Japanese deck 926 as containing 100% bucket measurements 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. 3b). 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, though such end-member values may be expected to vary across group 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. 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 are attributable to misclassification of ERI data. Other sources of variability prior to the introduction of ERI measurements in 1930, as well as post-1930 once offsets attributable to ERI misclassification are removed, appear consistent with physical contributions associated with difference in bucket design and measurement technique. Errors associated with data truncation (Chan et al. 2019) or other record keeping issues appear 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 supports the credibility of the linear-mixed-effects methodology for identifying offsets (Chan and Huybers 2019).

Acknowledgments

Elizabeth Kent (National Oceanography Centre) provided ship track data. Support was provided by the Harvard Global Institute.

APPENDIX

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.

Solar scheme: We model the total insolation absorbed by the top of a bucket as,

$$(1-a)(1-s)Q_g\pi r^2, \tag{A1}$$

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 reanalysis (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),

$$F = \begin{cases} 0 & \text{if } \frac{Q_g}{Q_0} \leq 0.35, \\ 2\frac{Q_g}{Q_0} - 0.7 & \text{if } \frac{Q_g}{Q_0} > 0.35, \end{cases} \tag{A2}$$

where Q_0 is incoming solar radiation at the top of the atmosphere. Values of $\frac{Q_g}{Q_0}$ below 0.35 are assumed to have complete cloud coverage. Incoming solar radiation is approximated as,

$$Q_0 = Q_{cs}[1 + 0.033\cos(2\pi\frac{t_d}{365})]\cos(\theta), \tag{A3}$$

where Q_{cs} is the solar constant (1370 $Jm^{-2}s^{-1}$), 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,

$$(1-a)(1-s)Q_g F\tan(\theta)2rh, \tag{A4}$$

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,

$$(1-a)(1-s)Q_g(1-F)\pi rh. \tag{A5}$$

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,

$$Q = (1-a)(1-s)Q_g[\pi r^2 + F\tan(\theta)2rh + (1-F)\pi rh]. \tag{A6}$$

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 ’true’ SST. Specifically, diurnal 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 ID indicator, source ID, platform, and deck information (Table A1). 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 cycle of air temperature, dew point 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 an NCDC-QC flag larger than five—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. Ship-board 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) corrects 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 to equal that of nighttime SSTs and then subtracting the mean difference between SST and shipboard temperatures. Note that the diurnal amplitude of shipboard 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. *Journal of Climate*, **26 (13)**, 4518–4534.

Ashford, O., 1948: A new bucket for measurement of sea surface temperature. *Quarterly Journal of the Royal Meteorological Society*, **74 (319)**, 99–104.

Berry, D. I., and E. C. Kent, 2009: A new air–sea interaction gridded dataset from ICOADS with uncertainty estimates. *Bulletin of the American Meteorological Society*, **90 (5)**, 645–656.

Berry, D. I., E. C. Kent, and P. K. Taylor, 2004: An analytical model of heating errors in marine air temperatures from ships. *Journal of Atmospheric and Oceanic Technology*, **21 (8)**, 1198–1215.

Brooks, C. F., 1926: Observing water-surface temperatures at sea. *Mon. Wea. Rev*, **54**, 241–253.

Carella, G., 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. *Geophysical Research Letters*, **45 (1)**, 363–371.

Carella, G., E. C. Kent, and D. I. Berry, 2017: A probabilistic approach to ship voyage reconstruction in ICOADS. *International Journal of Climatology*, **37 (5)**, 2233–2247.

Chan, D., and P. Huybers, 2019: Systematic differences in bucket sea surface temperature measurements amongst nations identified using a linear-mixed-effect method. *Journal of Climate*.

Chan, D., E. C. Kent, D. I. Berry, and P. Huybers, 2019: Correcting datasets leads to more homogeneous early-twentieth-century sea surface warming. *Nature*, **571 (7765)**, 393.

Cowtan, K., R. Rohde, and Z. Hausfather, 2018: Evaluating biases in sea surface temperature records using coastal weather stations. *Quarterly Journal of the Royal Meteorological Society*, **144 (712)**, 670–681.

Dee, D., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137 (656)**, 553–597.

Folland, C., and D. Parker, 1995: Correction of instrumental biases in historical sea surface temperature data. *Quarterly Journal of the Royal Meteorological Society*, **121 (522)**, 319–367.

Freeman, E., and Coauthors, 2017: ICOADS Release 3.0: a major update to the historical marine climate record. *International Journal of Climatology*, **37 (5)**, 2211–2232.

Hirahara, S., M. Ishii, and Y. Fukuda, 2014: Centennial-scale sea surface temperature analysis and its uncertainty. *Journal of Climate*, **27 (1)**, 57–75.

Huang, B., and Coauthors, 2017: Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *Journal of Climate*, **30 (20)**, 8179–8205.

Jones, P., 2016: The reliability of global and hemispheric surface temperature records. *Advances in Atmospheric Sciences*, **33 (3)**, 269–282.

Kawai, Y., and A. Wada, 2007: Diurnal sea surface temperature variation and its impact on the atmosphere and ocean: A review. *Journal of Oceanography*, **63 (5)**, 721–744.

Kennedy, J., P. Brohan, and S. Tett, 2007: A global climatology of the diurnal variations in sea-surface temperature and implications for MSU temperature trends. *Geophysical Research Letters*, **34 (5)**.

Kennedy, J., N. Rayner, 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. *Journal of Geophysical Research: Atmospheres*, **124 (14)**, 7719–7763.

Kennedy, J., N. Rayner, R. Smith, D. Parker, and M. Saunby, 2011: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. biases and homogenization. *Journal of Geophysical Research: Atmospheres*, **116 (D14)**.

Kennedy, J. J., 2014: A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Reviews of Geophysics*, **52 (1)**, 1–32.

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. *Journal of Atmospheric and Oceanic Technology*, **24 (2)**, 214–234.

Kent, E. C., and Coauthors, 2017: A call for new approaches to quantifying biases in observations of sea surface temperature. *Bulletin of the American Meteorological Society*, **98 (8)**, 1601–1616.

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.

Lane, D., D. Scott, M. Hebl, R. Guerra, D. Osherson, and H. Zimmer, 2013: *Introduction to Statistics: An Interactive e-Book*.

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. *Journal of Climate*, **19 (3)**, 446–469.

Reda, I., and A. Andreas, 2004: Solar position algorithm for solar radiation applications. *Solar energy*, **76 (5)**, 577–589.

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. *Agricultural and Forest Meteorology*, **38 (1-3)**, 217–229.

Ting, M., Y. Kushnir, and C. Li, 2014: North atlantic multidecadal SST oscillation: External forcing versus internal variability. *Journal of Marine Systems*, **133**, 27–38.

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. *Monthly Weather Review*, **139 (4)**, 1070–1082.

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

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 (7263)**, 511.

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. *American Journal of Physics*, **72 (3)**, 367–375.

LIST OF TABLES

TABLE 1. Parameters for the FP95d extended wooden bucket model. Values are assigned following Folland and Parker (1995). Two exceptions are insolation and percentage of air temperature in initial bucket temperature, which are determined by minimizing RMSE with ICOADS observations in 1990-2009 and are indicated by stars.

Parameter	Value
Exposure time (s)	300
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)	5.5
Ship-speed (m/s)	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 $(Jkg^{-1}K^{-1})$	1900
Albedo of bucket	0
Time of hauling (s)	60
Heat capacity of thermometer (gram of water)	35
Turbulence viscosity $(m^2\ s^{-1})$	$1.5e^{-5}$
Water thickness on wall (mm)	0.1
Relative humidity at water surface	0.98

Table A1. 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,

LIST OF FIGURES

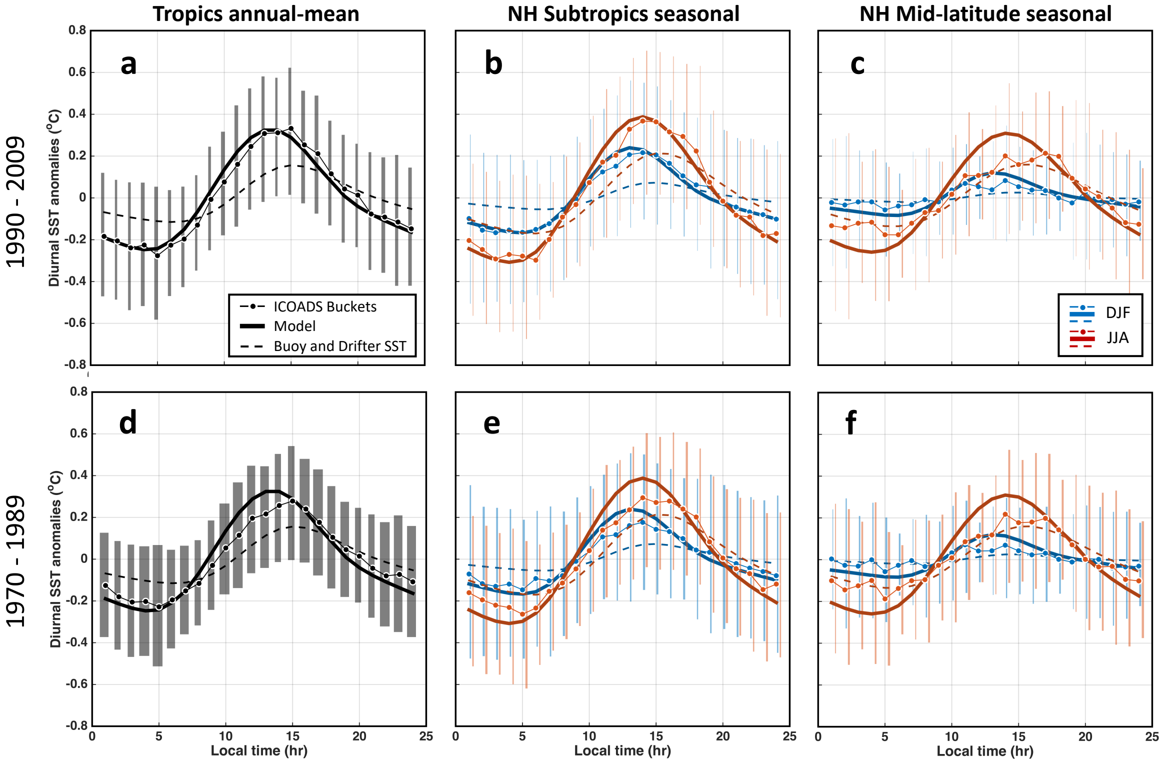

FIG. 1. **Observed and modeled diurnal cycles of bucket measurements.** The observed diurnal cycle of bucket temperatures (dotted lines) is in better agreement with diurnal variability simulated by the FP95d model (thick solid lines) than the diurnal cycle diagnosed from buoy and drifter measurements (dashed lines). The upper row shows the average diurnal cycle between 1990-2009 for (a) annual mean over the Tropics (20°S-20°N), (b) DJF (blue) and JJA (red) over Northern Hemisphere subtropics (20-40°N), and (c) DJF and JJA over NH mid-latitude (40-60°N). The lower row shows the same quantities but for 1970-1989. Model simulations are based on a same set of parameters for both rows (Table 1). The interquartile range of observations is indicated by bar lengths and their sample size is proportional to bar width.

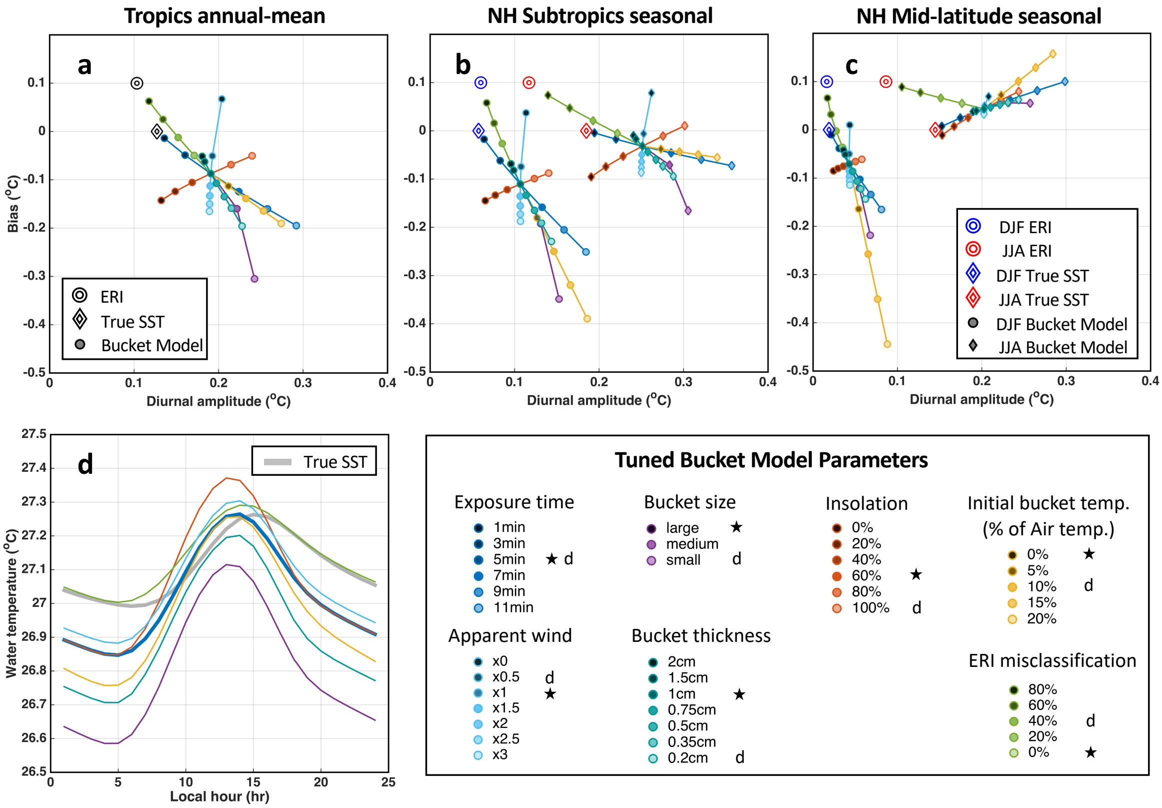

FIG. 2. **Simulated diurnal cycles and daily-mean SST biases using the FP95d model.** Changes in SST offsets and diurnal amplitudes in response to changes in model parameters are shown for the Tropics (a), Northern Hemisphere subtropics (b), and NH mid-latitudes (c). Extratropical results are for winter (circle) and summer (diamond). Reference parameters are indicated in the legend (black stars) and are listed in Table 1. Example diurnal cycles for the Tropics are shown as estimated from drifter and buoys (thick gray line), the reference simulation (thick blue line), and simulations varying individual parameters (thin lines and values indicated by ”d” in the legend).

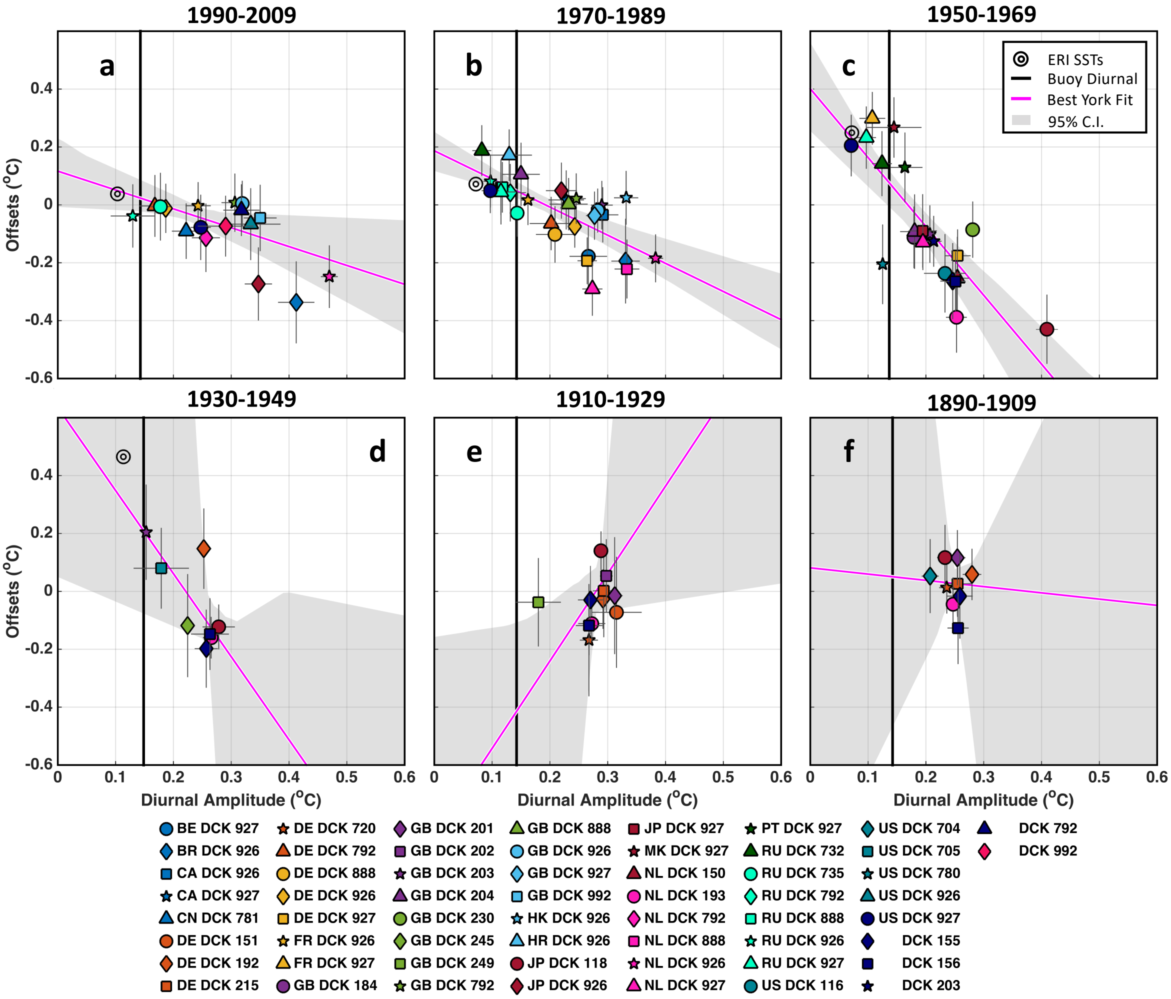

FIG. 3. **Groupwise bucket SST offsets and diurnal amplitudes in the Tropics.** Clear negative covariance exists between offsets and diurnal amplitudes across groups for 20 year periods between 1930 and 2009 (a-d), but covariance weakens and changes sign between 1910-1929 (e) and is essentially absent between 1890-1909 (f). Two standard deviation uncertainties are estimated from the linear-mixed-effects 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, and that numerous groups show a diurnal amplitude that is similar to or lower than that of buoy and drifter SSTs (vertical black lines).

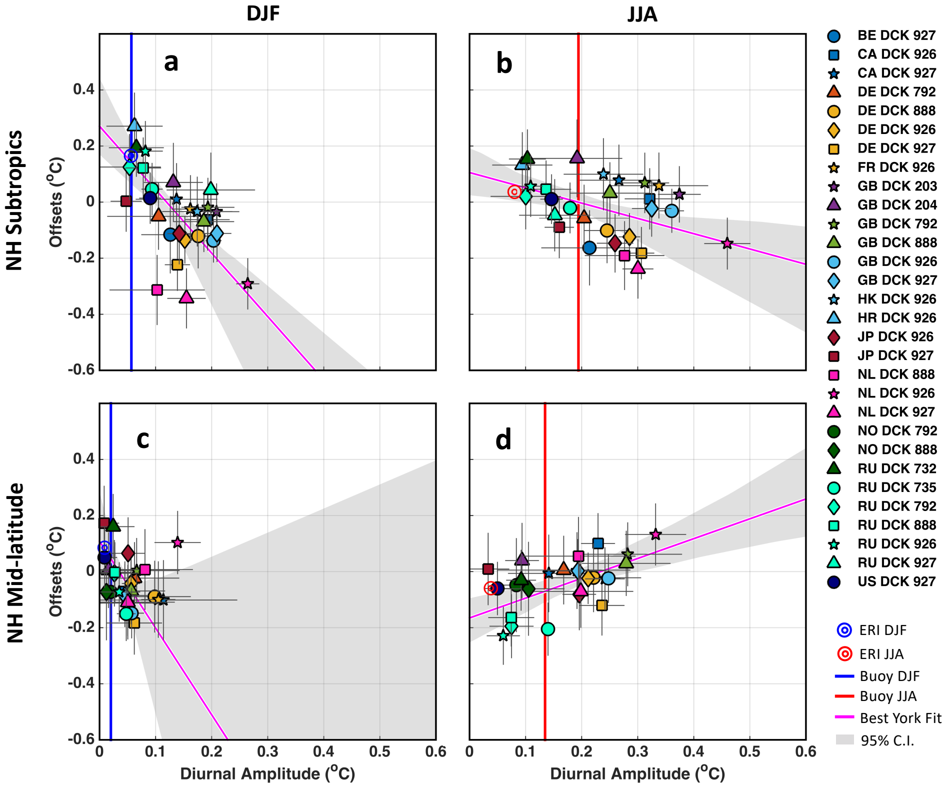

FIG. 4. **Diurnal cycles and groupwise SST offsets outside the Tropics.** The Northern Hemisphere subtropics show strong negative covariance in the winter (a, DJF) and a larger range of diurnal amplitudes but weaker covariance during summer (b, JJA). The Northern Hemisphere mid-latitudes show a similar pattern but also a smaller range of diurnal amplitudes during winter (c), consistent with weak diurnal variations in insolation, and a positive scaling during summer (d), indicative of greater solar heating during the day leading to warming and increased diurnal amplitudes (Fig. 2c). Results are for 1970-1989. Regression slopes intersect the offset and diurnal amplitude associated with ERI measurements (double circles) within uncertainties with the exception of mid-latitude summer. Approximately a third of the groups show diurnal amplitudes during summer that are smaller than found in buoy and drifter SST data (vertical lines).

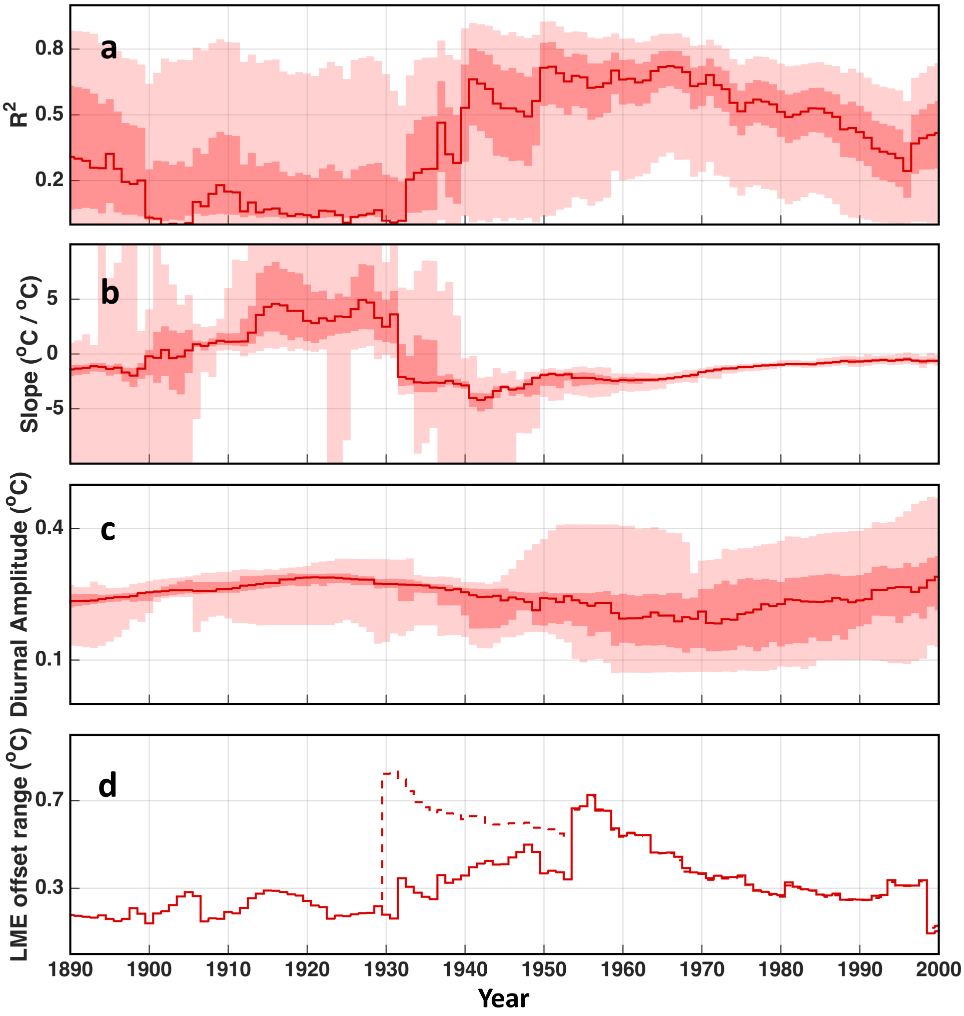

FIG. 5. **Evolution of groupwise offsets and diurnal amplitudes in the Tropics.** Major changes in bucket characteristics occur across 1930, with squared cross-correlation between diurnal amplitudes and offsets across bucket groups increasing (a) and the slope from a York fit switching from positive to negative values (b). Variations in diurnal amplitude change more smoothly (c), as do the 25% to 100% range of groupwise offsets (d, solid lines), unless the ERI end-member is included (d, dashed line). Each panel shows the median value (solid line) and (a-c) also include the interquartile range (dark shading) and 95% range (light shading). All analyses are from a 20-year sliding window with results plotted against the average year.

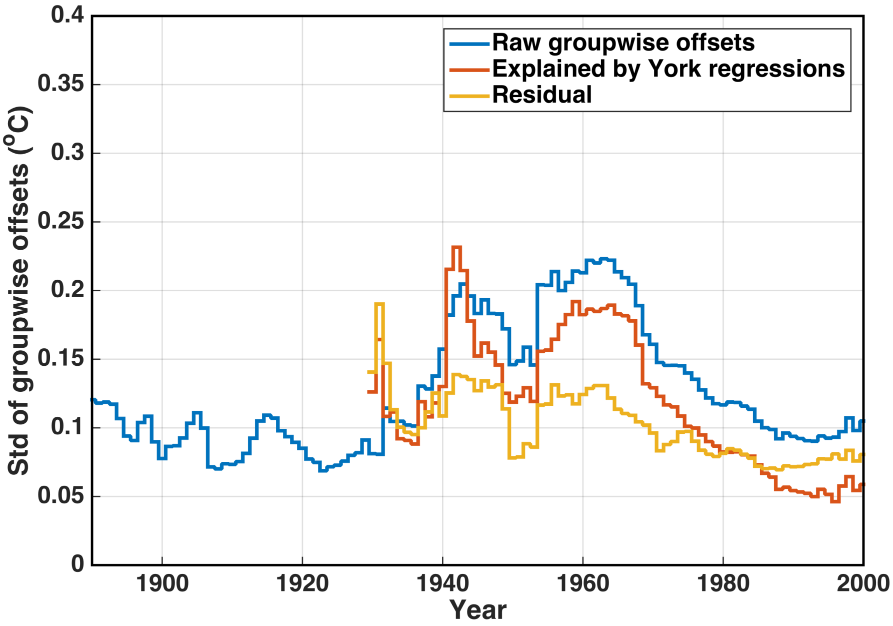

FIG. 6. **Standard deviations of groupwise offsets in the Tropics**. The standard deviation of offsets across groups increases after 1930 (blue curve; from markers in Fig. 3). If the component that linearly covaries with diurnal amplitude (red curve; c.f. magenta lines in Fig. 3) is first removed, however, the standard deviation of the residuals is more stable (yellow curve).

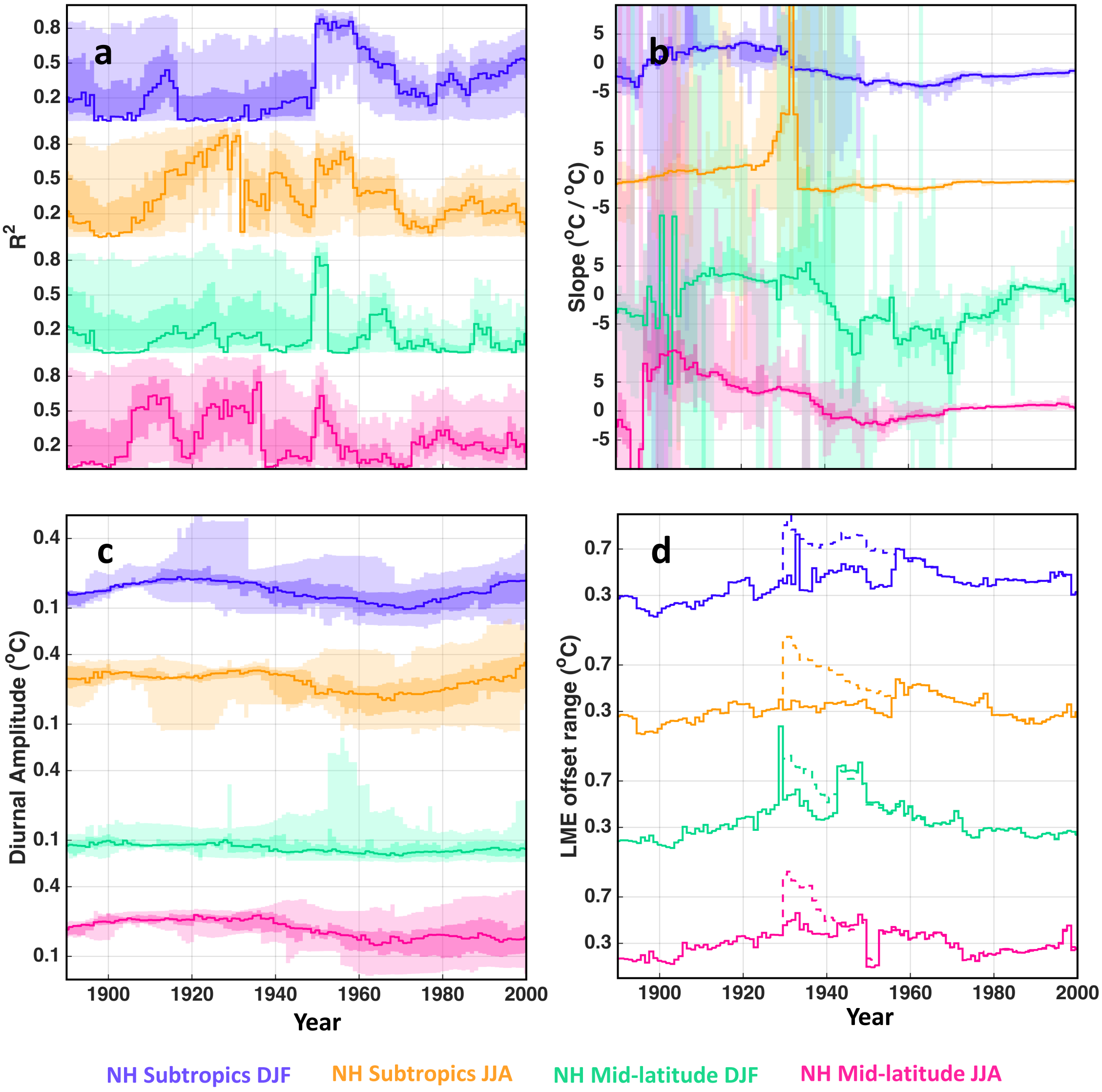

Fig. A1. **Evolution of groupwise diurnal amplitudes and offsets since 1890 outside the tropics.** Individual panels are as found in Fig. 5 but for different region and season combinations outside the Tropics.