Correcting Observational Biases in Sea Surface Temperature Observations Removes Anomalous Warmth during World War II

DUO CHAN[a] AND PETER HUYBERS[a]

[a] *Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts*

(Manuscript received 25 November 2020, in final form 18 February 2021)

ABSTRACT: Most historical sea surface temperature (SST) estimates indicate warmer World War II SSTs than expected from forcing and internal climate variability. If real, this World War II warm anomaly (WW2WA) has important implications for decadal variability, but the WW2WA may also arise from incomplete corrections of biases associated with bucket and engine room intake (ERI) measurements. To better assess the origins of the WW2WA, we develop five different historical SST estimates (reconstructions R1–R5). Using uncorrected SST measurements from the International Comprehensive Ocean–Atmosphere Data Set (ICOADS) version 3.0 (R1) gives a WW2WA of 0.41°C. In contrast, using only buckets (R2) or ERI observations (R3) gives WW2WAs of 0.18° and 0.08°C, respectively, implying that uncorrected biases are the primary source of the WW2WA. We then use an extended linear-mixed-effect method to quantify systematic differences between subsets of SSTs and develop groupwise SST adjustments based on differences between pairs of nearby SST measurements. Using all measurements after applying groupwise adjustments (R4) gives a WW2WA of 0.13°C [95% confidence interval (c.i.): 0.01°–0.26°C] and indicates that U.S. and U.K. naval observations are the primary cause of the WW2WA. Finally, nighttime bucket SSTs are found to be warmer than their daytime counterparts during WW2, prompting a daytime-only reconstruction using groupwise adjustments (R5) that has a WW2WA of 0.09°C (95% c.i.: −0.01° to 0.18°C). R5 is consistent with the range of internal variability found in either the CMIP5 (95% c.i.: −0.10° to 0.10°C) or CMIP6 ensembles (95% c.i.: −0.11° to 0.10°C). These results support the hypothesis that the WW2WA is an artifact of observational biases, although further data and metadata analyses will be important for confirmation.

SIGNIFICANCE STATEMENT: Major observational sea surface temperature (SST) estimates show a warm anomaly during World War II (WW2) that exceeds the warming expected from internal variability and known climate forcing. We systematically intercompare different groups of SST observations and trace the origin of the WW2 warmth foremost to anomalously warm U.S. and U.K. naval observations. We also find that nighttime bucket SSTs are anomalously warm, likely because of being measured inboard to avoid light pollution. SST estimates adjusted for these systematic biases give a more stable and smoothly evolving record of historical warming with a WW2 SST anomaly within the 95% range of internal variability found in an ensemble of general circulation model simulations.

KEYWORDS: Sea surface temperature; In situ oceanic observations; Ship observations; Bias; Climate variability

1. Introduction

The two most recent versions of the extended-reconstructed SST datasets, ERSST4 (Huang et al. 2015) and ERSST5 (Huang et al. 2017), both show anomalous warmth in global-mean SSTs that are, respectively, 0.30°C [95% confidence interval (c.i.): 0.17° to 0.41°C] and 0.29°C (0.23° to 0.37°C) during World War II (WW2) (Fig. 1a and Table 1). SST anomalies are calculated as the global, annual average between 1941 and 1945 relative to the average over 1936–40 and 1946–50, and are referred to as the World War II warm anomaly (WW2WA). All uncertainties are reported as 95% coverage intervals unless otherwise noted. Version 4 of the Hadley Center SST (HadSST4) shows a similar WW2WA of 0.19°C (Kennedy et al. 2019), although with a much large uncertainty estimate ranging from −0.09° to 0.45°C (Table 1).

If the WW2WA reflects physical changes in climate, it would have important implications for understanding the magnitude of decadal climate variability (Hansen et al. 2010; Morice et al. 2012; Vose et al. 2012), constraining uncertain external forcing (Stevens 2015), and partitioning relative contributions of anthropogenic forcing and internal variability in driving historical climate change (Jones et al. 2013; Bindoff et al. 2013; Maher et al. 2014; Hegerl et al. 2018). For example, such an anomaly could indicate the ability of El Niño–Southern Oscillation (ENSO) to lead to larger and more persistent warming than is otherwise understood (Thompson et al. 2009).

A number of other data-based analyses and simulations suggest that the physicality of the WW2WA is questionable. The WW2WA is essentially absent in HadSST3 (Kennedy et al. 2011b). SSTs referenced to air temperatures from nearshore weather stations (Cowtan et al. 2018) and temperature proxies derived from isotopes in tropical coral reefs (Pfeiffer et al. 2017) also show a negligible WW2WA. Furthermore, the WW2WA found in ERSST and

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-20-0907.s1.

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

DOI: 10.1175/JCLI-D-20-0907.1

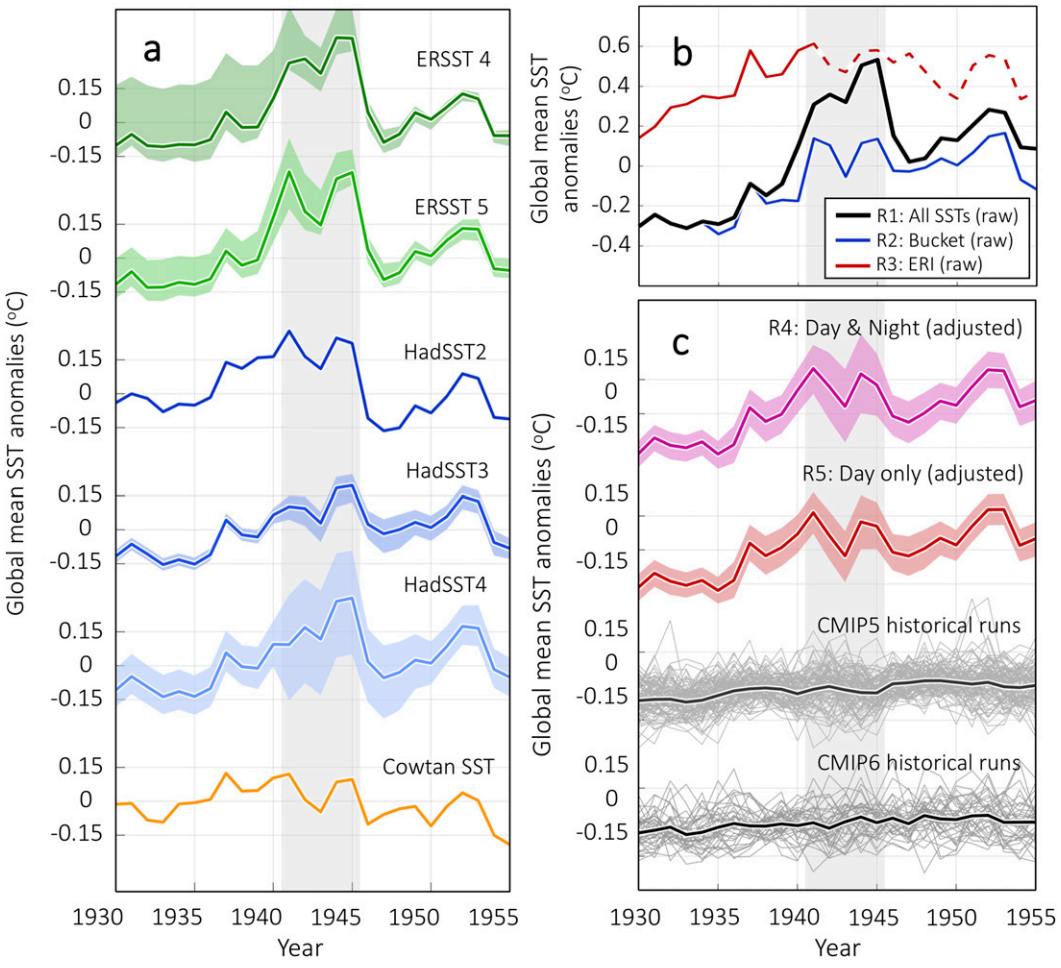

FIG. 1. SSTs across WW2 from various estimates. (a). Estimates from previously published SST datasets. ERSST4 (dark green), ERSST5 (light green), and HadSST4 (light blue) show a high World War 2 warm anomaly (WW2WA), whereas HadSST3 (medium blue) and Cowtan SST (orange) show no apparent WW2WA. (b) SSTs in raw ICOADS (R1–R3). Combining all available measurements (black; R1) results in a 0.41°C WW2WA that greatly exceeds that in either bucket-only (blue; R2) or ERI-only estimates (red; R3). U.S. SSTs without method information (dashed red) are assumed to be ERI SSTs because of their small-amplitude diurnal cycle (Carella et al. 2018). (c) SSTs after accounting for groupwise offsets using a linear-mixed-effect (LME) intercomparison method for daytime and nighttime SSTs (magenta; R4) and for daytime-only SSTs (red; R5). Also shown are ensemble averages over 94 CMIP5 (gray) and 38 CMIP6 (black) historical simulations, together with individual ensemble members (light gray). All SSTs, including observational estimates and simulations, are masked by the least-common coverage between HadSST4 and ICOADS daytime estimates on a month-by-month basis. SSTs are annual and global averages and are plotted relative to the average over 1936–40 and 1946–50 for each SST estimate, except that anomalies in R2 and R3 [in (b)] are relative to the 1936–40 and 1946–50 mean of R1 to show offsets between bucket and ERI SSTs. Shading (green, blue, magenta, or red) denotes the 95% confidence interval insomuch as an ensemble of adjustments or corrections is available.

HadSST4 estimates greatly exceeds that reproduced by any of the CMIP5 (Taylor et al. 2012) or CMIP6 (Eyring et al. 2016) historical simulations available over this interval (gray curves in Fig. 1c). Neither can statistical models explain this warm anomaly using known climate forcing and internal variability (Folland et al. 2018).

A leading hypothesis for the WW2WA relates to switching from predominantly bucket measurements of SST before and immediately after WW2 to engine-room-intake (ERI) measurements during WW2 (Thompson et al. 2008). A typical measurement from a U.K. canvas bucket has been estimated to be, on average, 0.4°C cooler than actual SSTs because of latent cooling before measurement (Folland and Parker 1995). Conversely, although ERI measurements are typically sampled at 5–15 m below the surface and should be consequently cooler than SSTs, given that SSTs are typically defined as coming from depths between 20 and 30 cm (Kennedy et al. 2019), ERI SSTs have an average warm bias ranging from 0.1° to 0.5°C because of absorbing heat released from ship engines (Kennedy et al. 2011b, 2019). A 58% reduction in the number of SST measurements during the WW2 interval (Freeman et al. 2017) could also make errors in a subset of the data more likely to lead to seemingly global anomalies.

A second hypothesis involves changes in protocols for taking measurements at night. Nighttime marine air temperature

TABLE 1. WW2WA and SST variability. The WW2WA is the average over 1941–45 relative to average over 1936–40 and 1946–50. The standard deviation of the global-mean SST (second column) is computed from annual averages between 1936–50. The average regional standard deviation (third column) is the square root of the global-average of variance at 5° × 5° grids. ERSSTs have lower regional variability because their mapping technique truncates small-scale variability. The Cowtan SST is only available as global averages. CMIP5 models do not contain sampling and random observational errors and, therefore, show lower regional variability. Sampling and random errors cancel under averaging and become negligible at global and decadal scales (Kennedy et al. 2011a; Chan et al. 2019). The 95% confidence intervals are given in square brackets and are estimated using the following ensembles: 1000 random adjustment members for groupwise-adjusted SSTs (R4 and R5), ERSST4, and ERSST5; 100 members for HadSST3; 200 members for HadSST4; 94 simulation members for CMIP5 historical runs; 38 members for CMIP6 historical runs; 1662 15-yr segments for CMIP5 preindustrial control simulations; and 1020 segments for CMIP6 control simulations.

	WW2WA (°C)	Global-mean s.d. (°C)	Average regional s.d. (°C)
R1, all SSTs (raw)	0.41	0.23	0.51
R2, bucket (raw)	0.18	0.13	0.47
R3, ERI (raw)	0.08	0.09	0.52
R4, day and night (adjusted)	0.13 [0.01, 0.26]	0.09 [0.07, 0.15]	0.45 [0.44, 0.46]
R5, daytime only (adjusted)	0.09 [−0.01, 0.18]	0.07 [0.06, 0.11]	0.45 [0.45, 0.46]
ERSST4	0.30 [0.17, 0.41]	0.16 [0.12, 0.22]	0.38 [0.33, 0.44]
ERSST5	0.29 [0.23, 0.37]	0.17 [0.14, 0.20]	0.41 [0.32, 0.44]
HadSST2	0.21	0.15	0.48
HadSST3	0.12 [0.03, 0.18]	0.08 [0.06, 0.11]	0.45 [0.45, 0.46]
HadSST4	0.19 [−0.09, 0.45]	0.12 [0.07, 0.23]	0.46 [0.45, 0.49]
Cowtan SST	0.05	0.08	—
CMIP5 historical	−0.02 [−0.14, 0.07]	0.07 [0.03, 0.11]	0.36 [0.27, 0.46]
CMIP6 historical	−0.00 [−0.10, 0.09]	0.07 [0.04, 0.12]	0.39 [0.30, 0.49]
CMIP5 control	−0.00 [−0.10, 0.10]	0.06 [0.03, 0.11]	0.36 [0.26, 0.45]
CMIP6 control	−0.00 [−0.11, 0.10]	0.07 [0.04, 0.12]	0.38 [0.30, 0.47]

readings are thought to have been taken inboard to avoid detection and, consequently, to be warmly biased by approximately 0.8°C (Folland et al. 1984). Bucket SSTs may have also been read inboard during WW2. We also note that the proportion of SST readings during the day, as opposed to the night, shifts from 55% in the surrounding 10 years of the war to 61% between 1941 and 1945 (Freeman et al. 2017), suggesting a preference for taking measurements during the daytime.

To further assess disagreement in existing estimates, we evaluate contributions to the WW2WA from specific groups of measurements and differences between day and nighttime measurements. In section 2, we investigate the evolution of bucket (R2) and ERI-only SSTs (R3) and show that, although SSTs from the two methods have systematic offsets, neither estimate has a strong WW2WA compared to when all data are used together (R1). In section 3, we re-examine the WW2WA after removing systematic offsets using an extended version of a linear-mixed-effect methodology (Chan and Huybers 2019) (R4). We also test the hypothesis of problematic nighttime bucket measurements using a daytime-only SST reconstruction (R5) in section 4. Finally, in section 5, we compare our results with estimates from previous studies and general circulation model simulations and discuss the implication of our updated estimate of the WW2WA.

2. R1–R3: Uncorrected reconstructions using all measurements or only buckets or ERIs

Six major SST estimates that cover the WW2 period (Fig. 1a) give distinct estimates for the WW2WA. All six estimates rely upon data coming from the International Comprehensive Ocean–Atmosphere Data Set (ICOADS; Woodruff et al. 2011; Freeman et al. 2017). Differences among estimates largely reflect differences in bias corrections, although use of different mapping procedures and inclusion criteria may also contribute. In one type of correction, bucket and ERI measurements are not distinguished, and global-mean SSTs are corrected to follow independent estimates of temperatures. Results depend on the choice of reference temperature. For example, ERSST5 (Huang et al. 2017) is referred to Hadley nighttime marine air temperatures (NMAT; Kent et al. 2013), from which the global average inherits a WW2WA of 0.22°C in NMAT estimates. Like SSTs, ship-based air temperatures are potentially subject to their own biases on account of changes in measurement protocols (Folland et al. 1984). Alternatively, referencing against air temperatures from coastal and island weather stations leads to removal of the WW2WA (Cowtan et al. 2018), an estimate that we refer to as Cowtan SST.

A second approach to correcting SST biases involves distinguishing between bucket and ERI measurements and attempts to account for their respective biases (Kennedy et al. 2011b; Hausfather et al. 2017; Kennedy et al. 2019), potentially giving a more detailed correction than available from a bulk correction of all SST data. A major impediment to such corrections, however, is that method information is poorly documented for most measurements during WW2, with only 6% of observations explicitly indicated as coming from buckets, 11% explicitly indicated as coming from ERIs, and 83% whose method requires some degree of inference (Freeman et al. 2017; Kent et al. 2017). Magnitudes of measurement biases are also uncertain and may have changed during WW2 (Folland

et al. 1984; Kent et al. 2013). The lack of information regarding measurements has been addressed through plausible but potentially insufficient assumptions. In constructing HadSST3, for example, Kennedy et al. (2011b) assumed that U.S. and U.K. naval ships with missing method information take ERI measurements of SST that are, on average, warmly biased by 0.2°C. HadSST4 randomly designates measurements with missing method information during WW2 to be either bucket or ERI SSTs, with the portion of bucket measurements ranging from 0% to 25%. Wartime ERI measurements in HadSST4 are assumed to be biased warm, on average, by 0.25°C, whereas bucket SSTs are assumed to be biased cold, on average, by −0.2°C (Kennedy et al. 2019).

The implications of these corrections for the WW2WA are not obvious, and it is useful to make raw estimates—neither including corrections nor infilling regions that lack data—to better quantify the magnitude of the WW2WA in the underlying data. We, therefore, first reconstruct SST using all quality-controlled raw ship-based measurements (R1) available from ICOADS3.0 (Freeman et al. 2017). We also examine SSTs estimated using data thought to come only from buckets (R2) or ERIs (R3).

Quality control procedures for SST measurements are the same as those in Chan and Huybers (2019). We identify ship-based SSTs using the ICOADS platform metadata (PT from 0 to 5). Method information is identified from ICOADS SST measurement method (SI) metadata. If the SI metadata are not available, the measurement method is assigned to be unknown. Following Kennedy et al. (2011b), an exception is made for SST measurements from U.S. ships, which are assumed to be ERI measurements. This assumption is supported by the fact that U.S. measurements have a diurnal cycle that is smaller than that expected from bucket measurements (Carella et al. 2018). A small diurnal cycle is consistent with ERI measurements that are typically sampled at a depth of 5–15 m that is less affected by the diurnal cycle of insolation (Carella et al. 2018). We identify nations first using the ICOADS country code (C1). If C1 is not available, nations are inferred from ship call signs (Chan et al. 2019) or deck information (Kennedy et al. 2011b; Chan and Huybers 2019).

Global-average estimates of raw SSTs using only observations thought to come from buckets (R2) gives a WW2WA of 0.18°C, and a similar estimate for ERI-only SSTs (R3) gives a WW2WA of 0.08°C (Fig. 1b). Both estimates are far more stable than the 0.41°C WW2WA obtained if all available raw ICOADS SSTs are evaluated (R1; Fig. 1b). The fact that R3 is, on average, 0.52°C warmer than R2 highlights the potential for misidentification of measurements methods or insufficient corrections leading to biases remaining in existing SST estimates. Note that quantifying the WW2WA as the difference between the average over 1941–45 and the average over the 10 surrounding years neutralizes the effect of a constant SST bias and also accounts for the potential for an underlying linear trend between 1936 and 1950.

In addition to a reduced WW2WA in SST estimates stratified by instruments, R1 follows the ERI-only estimate (R3) more closely during the war and the bucket-only estimate (R2) before and after the war (Fig. 1b). The proportion of SSTs we identify to come from buckets decreases from 44% before and after the war to 6% during the war, whereas the proportion identified to come from ERIs, including both explicitly indicated and inferred U.S. measurements, increases from 25% to 50%. The remaining 44% of observations during WW2 have unassigned measurement types. This initial investigation of raw ICOADS indicates that the WW2WA mainly reflects instrumental changes at the start and the end of the war (Thompson et al. 2008).

3. R4: Accounting for groupwise offsets

To better account for limitations in previous corrections, we use a linear-mixed-effect (LME) intercomparison framework (Chan and Huybers 2019) to quantify systematic offsets associated with distinct groups of SSTs. We use our LME model to compare nearby measurements and, thereby, obtain estimates of SST offsets among different groups regardless of whether the method of measurement is known. Moreover, by diagnosing data offsets associated with individual nations, groups of SSTs, and available measurement types, the LME method allows for inference of more-detailed SST adjustments than previous estimates (e.g., Kennedy et al. 2019). Details include different magnitudes of offsets potentially contributed by different buckets or ERI designs, distinct protocols, or separate postprocessing effects, as well as the temporal and spatial patterns of observational biases associated with each group.

We applied a similar LME method to only bucket SST observations and showed that it accurately identifies offsets between nation and deck groups (Chan and Huybers 2019). The skill of the LME method is also supported by negative correlations between offsets and the amplitude of diurnal cycles in SSTs (Chan and Huybers 2020), identification of offsets later found to come from data truncation (Chan et al. 2019), and improved agreement between adjusted SSTs and air temperatures from nearby coastal weather stations (Chan et al. 2019). In this study, we assess all available ship-based SSTs that come from buckets, ERIs, or hull sensors, or with missing method information.

a. *Linear-mixed-effect method*

SST observations are grouped according to nation, deck, and method of measurement. Nation and method information is identified following the same approach as in the last section, but we no longer assume that U.S. SSTs with missing method information are ERI measurements, as done in obtaining R3. Rather, we define "missing method" as a category and allow the LME method to determine any required adjustment for these U.S. measurements. We include deck numbers in defining different groups because these indicate information regarding ICOADS data collectors and processors (Freeman et al. 2017), and processing has been found to be a potentially important source of bias (Chan et al. 2019).

To intercompare different groups of SSTs and estimate systematic offsets, we first pair SSTs if they come from distinct grouping according to nation, deck, and method and are within 300 km and 2 days of one another. We use each measurement at most once to prevent error covariance between pairs, with

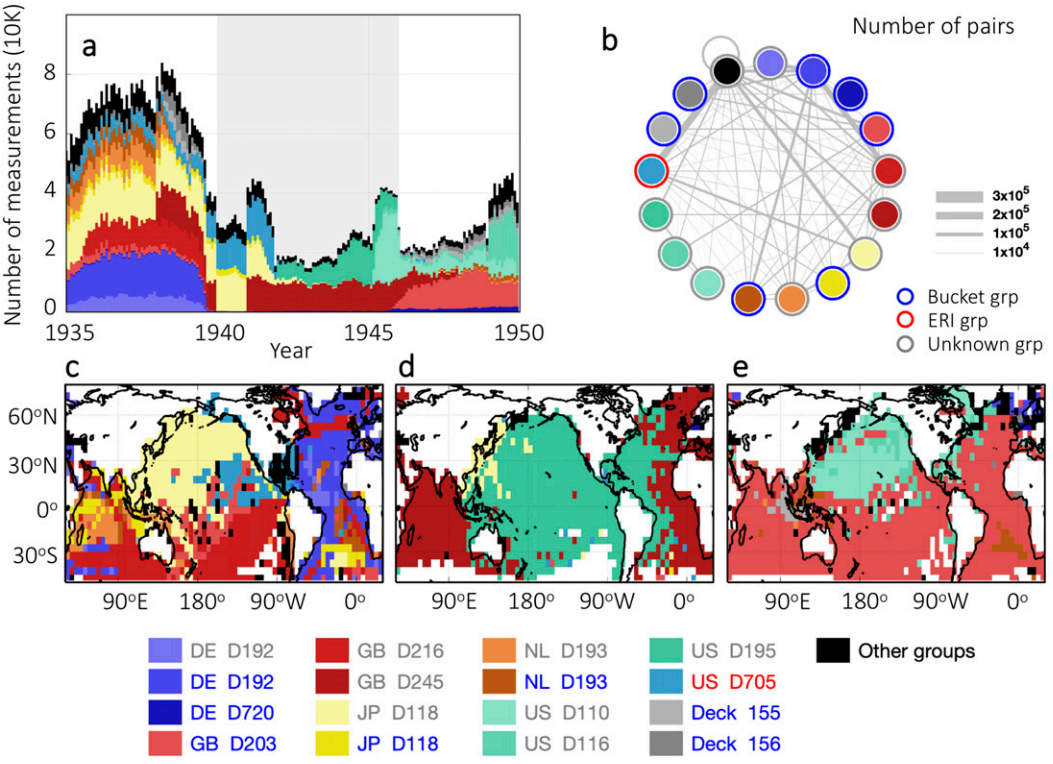

FIG. 2. Distribution of SST measurements. (a) The number of SST measurements according to individual groups from 1935–49. Groups are identified by country, deck, and instrument. Nation abbreviations are DE: Germany; GB: Great Britain; JP: Japan; NL: the Netherlands; and U.S.: the United States. Groups having fewer than 100 000 measurements in 1935–49 are labeled as "other groups." Instruments are indicated by the color of group names for bucket (blue), ERI (red), and unknown methods (gray). (b) Numbers of SST pairs (width of connections) between individual groups (filled circles) during 1935–49. The color of outer circles denotes measurement methods. Also shown are maps indicating groups that contribute the most observations within 5° grid boxes, for (c) 1936–40, (d) 1941–45, and (e) 1946–50. White grid boxes have fewer than 3 years of data within corresponding 5-yr intervals.

the pairing algorithm prioritizing SST pairs that are closest in space (Chan and Huybers 2019). Compared with Chan and Huybers (2019), who only intercompared SSTs thought to come from buckets, the inclusion of SSTs from ERIs, Hull sensors, or with missing method information increases the total number of pairs from 17.8 to 45.8 million throughout 1850–2014. These pairs come from 492 groups that each contribute at least 5000 pairs of SSTs. Our focus is on the years 1935–49 that contains a subset of 1.8 million SST differences from 66 groups (Fig. 2b and Table 2), but we analyze all 45.8 million SST pairs for purposes of more fully accounting for covariance across groupings. To account for physical separation between paired measurements, we first remove climatological differences expected from geographical and temporal displacement. The expected differences are estimated from NOAA optimal interpolated SSTs (Reynolds et al. 2007) and drifter observations in ICOADS3.0 (Chan and Huybers 2019).

After removing expected offsets, the remaining differences in paired temperatures $\delta\mathbf{T}$ are represented as arising from offsets among groups:

$$\delta\mathbf{T} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}_y\boldsymbol{\beta}_y + \mathbf{Z}_r\boldsymbol{\beta}_r + \boldsymbol{\epsilon}. \quad (1)$$

The fixed-effect term $\boldsymbol{\alpha}$ describes offsets between groups, and random effects terms describe 5-yr and regional variations, $\boldsymbol{\beta}_y$ and $\boldsymbol{\beta}_r$ respectively, around mean offsets. Fixed effects are constrained so that the offset is zero when averaged across all pairs over all years. Design matrices $\mathbf{X}$, $\mathbf{Z}_y$, and $\mathbf{Z}_r$ have entries of 0, 1, and -1 that specify which observations are intercompared. Figure 3 illustrates the element-wise specification given in Eq. (1). The designation of fixed and random effects is identical to that in Chan and Huybers (2019) but with allowance for groupings that involve different measurement types and a related larger number of observations.

In practice, to reduce the computational cost, we aggregate data by averaging SST differences according to combinations of pairs of groups, regions, and years before estimating offsets. Uncertainties associated with aggregated SST differences are budgeted to account for observational error, physical SST variability, and heteroscedasticity associated with distinct group size, and are used to weigh aggregated pairs in the LME analysis. The error estimate resulting from the LME analysis is a multivariate Gaussian that accounts for covariance. We represent the uncertainties of groupwise adjustments using a 1000-member ensemble of random adjustments having

TABLE 2. Measurement groups containing SSTs during 1935–49. Among 66 groups that have groupwise offset estimates, 33 have valid estimates for the amplitude of diurnal cycles. Shown LME offsets are annual mean offsets averaged over 1935–49 for the analysis that uses both daytime and nighttime SSTs (''Mean offsets'' column) and the analysis that only uses daytime SSTs (''Day offsets'' column). Diurnal amplitudes are anomalies relative to collocated 1990–2014 climatology estimated from drifting buoys (''Excess DA'' column) . One asterisk (*) indicates groupwise offsets differing from zero ($P < 0.05$) or diurnal amplitudes differing that of drifting buoys ($P < 0.05$). Two asterisks (**) indicate significance after Bonferroni corrections ($P < 0.05/66$ for groupwise offsets and $P < 0.05/33$ for diurnal amplitudes). Checkmarks highlight U.S. groups with unknown method information, which are assumed to contain ERI SSTs when computing ERI-only estimates (R3).

Group	Nation	Deck	Method by ICOADS SI	Thousands of measurements	Mean offsets	Day offsets	Excess DA	
DE DCK 151	Germany	Pacific HUSST German receipts	Bucket	3	−0.10	−0.07	0.08*	
DE DCK 192	Germany	Deutsche Seewarte Marine	Unknown	259	0.12	0.13	0.13**	
DE DCK 192	Germany	Deutsche Seewarte Marine	Bucket	707	0.12	0.10	0.11**	
DE DCK 215	Germany	German Marine	Bucket	35	0.10	0.13	0.09**	
DE DCK 720	Germany	Deutscher Wetterdienst Marine Met. Archive	Bucket	103	−0.06	−0.12	0.05**	
GB DCK 184	Great Britain	Great Britain Marine	Unknown	10	−0.01	−0.02	0.07	
GB DCK 184	Great Britain	Great Britain Marine	Bucket	62	−0.03	−0.02	−0.00	
GB DCK 203	Great Britain	Selected U.K. ships	Bucket	568	−0.02	0.01	0.02**	
GB DCK 204	Great Britain	British Navy (HM) ships	Bucket	71	−0.06	−0.09	0.01	
GB DCK 216	Great Britain	U.K. Merchant ship logbooks	Unknown	415	−0.06	−0.07	0.08**	
GB DCK 245	Great Britain	Royal Navy ship's logs	Unknown	901	0.06	0.04	0.05**	
HO DCK 705	—	U.S. Merchant Marine Collection (series 500)	Bucket	13	−0.15	−0.17	−0.02	
HO DCK 705	—	U.S. Merchant Marine Collection (series 500)	ERI	21	0.54**	0.47**	−0.10**	
JP DCK 118	Japan	Kobe Collection data	Unknown	1 056	−0.32**	−0.28**	0.20**	
JP DCK 118	Japan	Kobe Collection data	Bucket	178	−0.37*	−0.31*	0.12**	
NL DCK 150	Netherland	Pacific HUSST Netherlands receipts	Bucket	9	−0.23*	−0.23	0.11**	
NL DCK 193	Netherland	Netherlands Marine	Unknown	296	−0.23*	−0.17*	0.11**	
NL DCK 193	Netherland	Netherlands Marine	Bucket	227	−0.22*	−0.20*	0.12**	
PM DCK 705	—	U.S. Merchant Marine Collection (series 500)	ERI	27	0.64**	0.55**	−0.09**	
RU DCK 732	Russian	Russian Marine Met Dataset	Unknown	52	0.09	0.02	0.02	
RU DCK 735	Russian	Russian Research Vessel	Hull sensor	1	−0.10	−0.11	0.07*	
U.S. DCK 110	United States	U.S. Navy Marine	Unknown	402	0.51**	0.45**	−0.08**	✓
U.S. DCK 116	United States	U.S. Merchant Marine	Unknown	234	0.25*	0.19*	−0.04**	✓
U.S. DCK 281	United States	U.S. Navy Monthly Aerological Record	Unknown	90	0.47**	0.41**	−0.08**	✓
U.S. DCK 703	United States	U.S. Lightship Collections	Unknown	17	−1.08**	−1.02**	0.03	✓
U.S. DCK 705	United States	U.S. Merchant Marine Collection (series 500)	Unknown	26	0.42**	0.39**	−0.05	✓
U.S. DCK 705	United States	U.S. Merchant Marine Collection (series 500)	Bucket	99	0.01	−0.03	0.00	
U.S. DCK 705	United States	U.S. Merchant Marine Collection (series 500)	ERI	542	0.48**	0.45**	−0.09**	
U.S. DCK 710	United States	U.S. Arctic logbooks	Unknown	2	−0.17	0.02	0.04*	✓
U.S. DCK 780	United States	NOAA World Ocean Database	Bucket	9	0.16	0.11	−0.00	
ZA DCK 899	South Africa	South Africa whaling	Unknown	13	−0.01	−0.12	−0.03**	
DCK 155	—	Indian HSST	Bucket	119	−0.32**	−0.23	0.12**	
DCK 156	—	Atlantic HSST	Bucket	189	−0.17*	−0.15*	0.08**	
BX DCK 706	—	U.S. Merchant Marine Collection (series 600)	Unknown	5	−0.10	−0.15	—	
CN DCK 706	China	U.S. Merchant Marine Collection (series 600)	Unknown	1	−0.38*	−0.43**	—	
DL DCK 706	—	U.S. Merchant Marine Collection (series 600)	Unknown	2	0.07	0.04	—	

TABLE 2. (*Continued*)

Group	Nation	Deck	Method by ICOADS SI	Thousands of measurements	Mean offsets	Day offsets	Excess DA	
DN DCK 706	—	U.S. Merchant Marine Collection (series 600)	Unknown	3	−0.12	−0.04	—	
FR DCK 706	France	U.S. Merchant Marine Collection (series 600)	Unknown	7	0.42**	0.39**	—	
GB DCK 152	Great Britain	Pacific HUSST U.K. receipts	Bucket	<1	−0.04	−0.05	—	
GB DCK 202	Great Britain	All ships (U.K. Met Office MDB)	Bucket	<1	−0.06	−0.03	—	
GB DCK 205	Great Britain	Scottish fishery cruisers	Bucket	4	−0.43*	−0.39*	—	
GB DCK 705	Great Britain	U.S. Merchant Marine Collection (series 500)	Bucket	18	−0.12	−0.07	—	
GB DCK 705	Great Britain	U.S. Merchant Marine Collection (series 500)	ERI	11	0.13	0.08	—	
GB DCK 706	Great Britain	U.S. Merchant Marine Collection (series 600)	Unknown	23	−0.03	−0.09	—	
GB DCK 707	Great Britain	U.S. Merchant Marine Collection (series 700)	Unknown	<1	−0.19	−0.26*	—	
HO DCK 706	—	U.S. Merchant Marine Collection (series 600)	Unknown	<1	0.05	0.00	—	
HO DCK 707	—	U.S. Merchant Marine Collection (series 700)	Unknown	<1	0.09	0.05	—	
IY DCK 706	—	U.S. Merchant Marine Collection (series 600)	Unknown	6	0.19	0.20*	—	
JP DCK 705	Japan	U.S. Merchant Marine Collection (series 500)	Unknown	10	−0.31*	−0.27*	—	
JP DCK 705	Japan	U.S. Merchant Marine Collection (series 500)	Bucket	26	−0.38*	−0.35*	—	
JP DCK 705	Japan	U.S. Merchant Marine Collection (series 500)	ERI	5	−0.31*	−0.31*	—	
JP DCK 706	Japan	U.S. Merchant Marine Collection (series 600)	Unknown	8	−0.26*	−0.28**	—	
NL DCK 705	Netherland	U.S. Merchant Marine Collection (series 500)	Bucket	10	−0.21	−0.19	—	
NL DCK 706	Netherland	U.S. Merchant Marine Collection (series 600)	Unknown	17	−0.18*	−0.14*	—	
NO DCK 706	Norway	U.S. Merchant Marine Collection (series 600)	Unknown	3	0.34*	0.31**	—	
PM DCK 707	—	U.S. Merchant Marine Collection (series 700)	Unknown	<1	0.59**	0.46*	—	
RU DCK 735	Russian	Russian research vessel	Bucket	<1	−0.05	—	—	
SP DCK 706	—	U.S. Merchant Marine Collection (series 600)	Unknown	<1	0.01	−0.01	—	
U.S. DCK 116	United States	U.S. Merchant Marine	Bucket	13	−0.12	−0.14	—	
U.S. DCK 195	United States	U.S. Navy ships logs	Unknown	383	0.43**	0.39**	—	✓
U.S. DCK 706	United States	U.S. Merchant Marine Collection (series 600)	Unknown	51	0.28*	0.29**	—	✓
U.S. DCK 707	United States	U.S. Merchant Marine Collection (series 700)	Unknown	2	0.30*	0.27*	—	✓
U.S. DCK 780	United States	NOAA World Ocean Database	Unknown	1	0.22*	0.20*	—	✓
ZA DCK 927	South Africa	International Marine	Unknown	17	0.01	0.01	—	
DCK 197	—	Danish Marine (Polar)	Unknown	3	−0.12	−0.06	—	
DCK 255	—	Undocumented TDF-11 decks or MDB series	Bucket	<1	0.10	0.13	—	

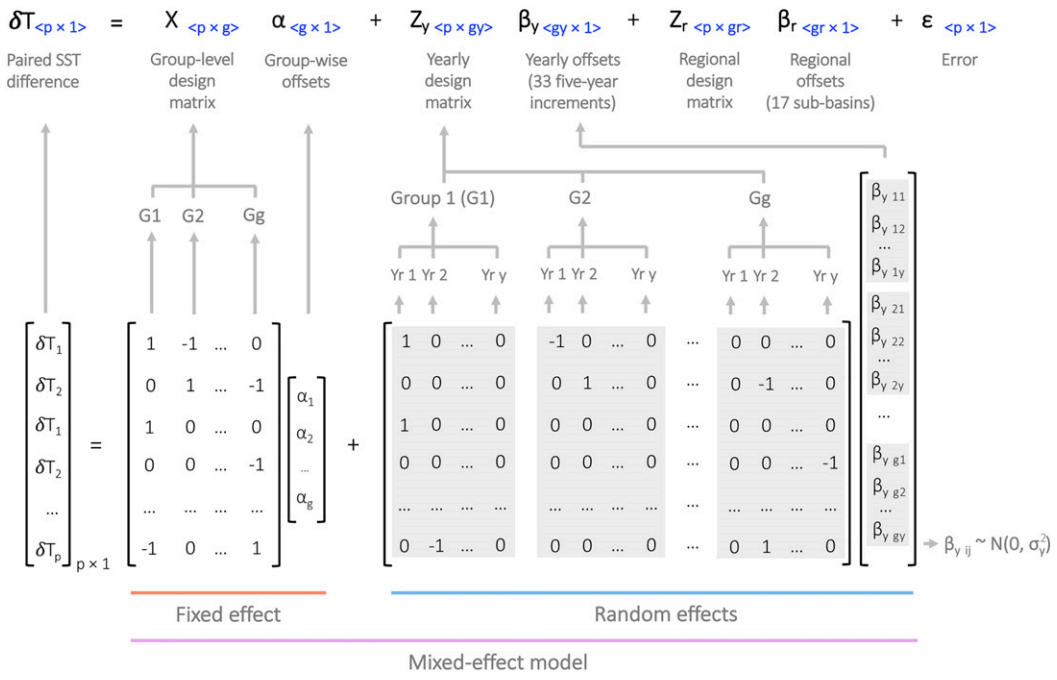

FIG. 3. An element-wise illustration of the LME model in Eq. (1). Equation (1) is given along with the dimensionality of matrices and vectors (blue), where p, g, y, and r are, respectively, numbers of pairs, groups, 5-yr increments, and regions, respectively. Three terms are illustrated in detail: 1) paired SST differences $\delta\mathbf{T}$; 2) $\mathbf{X}$ is a design matrix that specifies groupwise interactions between paired observations and $\boldsymbol{\alpha}$ are the fixed effects; and 3) $\mathbf{Z}_y$ is a design matrix expanded to specify 5-yr bins in which groupwise interactions take place and $\boldsymbol{\beta}_y$ represents 5-yr random effects that are assumed to follow a Gaussian distribution, $\beta_{y_{ij}} \sim N(0, \sigma_y^2)$. Regional effects $\mathbf{Z}_r\boldsymbol{\beta}_r$ are also estimated for individual groups and have a similar structure to $\mathbf{Z}_y\boldsymbol{\beta}_y$.

groupwise offsets that are perturbed according to the estimated multivariate Gaussian. The ensemble captures error covariance among fixed and random effects as well as covariance introduced by changes in the spatial coverage of individual groups. Further details regarding the LME implementation are in appendix A.

One limitation of the LME method is that it informs regarding relative offsets and does not account for biases common to all groups. Common SST biases, however, may vary with time. For example, bucket biases toward being cold are generally thought to diminish with systematic changes from less-insulated canvas to more-insulated rubber buckets (Folland and Parker 1995; Kennedy et al. 2011b, 2019). Existing estimates represent this bucket change as occurring gradually from the 1930s to the 1970s (Kennedy et al. 2019) or as being confined to after the 1950s (Kennedy et al. 2011b). Systematic changes in biases for other ship-based measurements have also been identified for more recent years by comparison against marine profile measurements (e.g., Kennedy et al. 2019). We cannot rule out systematic changes between all types of measurements during WW2 but proceed with an examination of identifiable offsets.

b. *Groupwise offsets and the diurnal cycle*

We first apply the LME methodology to intercompare groups of SST measurements using all available SST data. Of the 66 groups present between 1935 and 1949, 29 have significant offsets ($P < 0.05$; Table 2). Significance is assessed relative to a null hypothesis of zero-mean offset relative to the average across all groups (Chan and Huybers 2019). In addition, 12 groups still show significant offsets after a Bonferroni correction. The Bonferroni correction compensates for the increased chance of false positives when conducting n tests by evaluating each at $P < \alpha/n$. In our case, we have n equal to 66 groups between 1935 and 1949 and α equal to 0.05. There are five positively identified ERI groups that are each found to be warmer than the 24 bucket groups, on average, by 0.53°C (0.25° to 0.72°C; Table 2). Offsets of groups having missing method information range from −0.4° to 0.6°C. This range is similar to that spanned by the entire population of the bucket and ERI groups, suggesting that at least some of these groups are distinctly from bucket or ERI measurements.

Chan and Huybers (2020) demonstrated the utility of using the diurnal cycle in combination with offsets to infer the composition of measurements within a group. A negative correlation is generally found between diurnal amplitudes and offsets across groups that have variable compositions of ERI and bucket data because ERIs are generally warmer and have a smaller diurnal cycle than bucket measurements. To estimate diurnal cycles, we make use of tracked ships (Carella et al. 2017) and only evaluate ships making measurements at least four times per day using a least squares fit

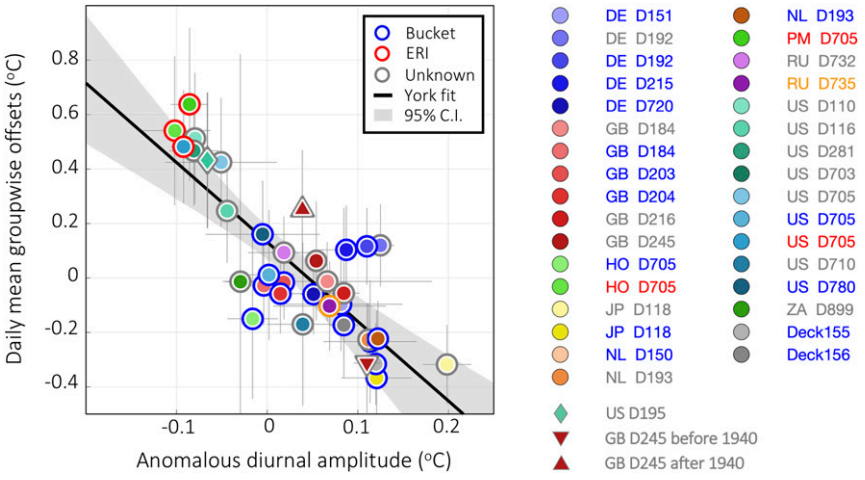

FIG. 4. Groupwise offsets are negatively correlated with diurnal-cycle amplitudes. Offsets are averaged over 1935–49 and plotted against diurnal amplitude anomalies. Diurnal anomalies are relative to a 1990–2014 climatology from drifters. Colors of inner circles denote nation and deck combinations, and colors of outer circles denote instruments. The 95% confidence intervals (c.i.) are from the linear-mixed-effect analysis (vertical bars on each marker) and least squares sinusoidal fits of amplitudes (horizontal bars). A linear trend of offset with amplitude (black line) is from a York regression with the 95% c.i. (gray shading) estimated by bootstrapping individual groups. U.S. deck 195 (diamond) samples at three local hours in ICOADS3.0 and has its diurnal amplitude estimated using a different method (see appendix B). Although GB deck 245 always has diurnal amplitudes significantly ($P < 0.05$) higher than that from drifters, SSTs before WW2 (1935–39; downward-pointing triangle) are coldly offset by –0.31°C, and SSTs afterWW2 begins (1941–47; upward-pointing triangle) are warmer and have an offset of 0.25°C.

of a once-per-day sinusoid. Although the Nyquist cutoff is two measurements per day, in practice non-sinusoidal components of the diurnal cycle make using higher-resolution data useful. U.S. deck 195 presents a special case, however, because it contributes 47% of U.S. wartime measurements (Fig. 2a) but has a sampling frequency of only three times per day. To fit the amplitude of this deck, we estimate a linear combination of diurnal cycles from two basis functions based upon known bucket and ERI measurements (see appendix B). The best estimate indicates that U.S. deck 195 is consistent with being purely composed of ERI measurements.

Half of the 66 groups present between 1935 and 1949, along with U.S. deck 195, have tracked ships with sufficient resolution of the diurnal cycle (Table 2). To account for distinct spatial and seasonal coverage of individual groups, we report diurnal amplitudes as anomalies relative to a 1990–2014 climatology of diurnal amplitudes estimated from drifting buoys (Chan and Huybers 2019). As expected, diurnal amplitude is strongly anticorrelated with groupwise offsets (Fig. 4). The relationship between diurnal amplitudes and groupwise offsets is estimated using a York regression (York et al. 2004) and associated uncertainties are estimated by bootstrapping individual groups 10 000 times with replacement. The three known ERI groups are associated with relatively warm and small-amplitude diurnal cycles, whereas all known bucket groups are colder and have a higher diurnal amplitude. In general, bucket groups have higher intergroup variability in terms of both diurnal amplitudes and groupwise offsets, consistent with the fact that a variety of bucket designs and measurement protocols were used to collect SSTs (Folland and Parker 1995; Kent and Taylor 2006). The large spread across bucket groups may also involve misclassification of ERI SSTs as coming from buckets (Carella et al. 2018; Chan and Huybers 2020).

U.S. decks 110, 116, 195, 281, and 705 account for 88% of all U.S. measurements during 1935–49 and each is significantly warmer than the average across all groups and exhibits a diurnal amplitude that is significantly smaller than a climatology derived from drifting buoys ($P < 0.05$; Fig. 4, Table 2). The combination of warm offsets and small diurnal amplitudes supports the assumption made in HadSST3 (Kennedy et al. 2011b) and the findings of Carella et al. (2018) that U.S. measurements with missing method information during WW2 are ERI measurements. Confirmation of U.S. decks being composed of ERI measurements also supports the offset between R2 and R3 reflecting biases between ERI and bucket measurements.

c. *Reduced WW2WA after removing groupwise offsets*

The R4 reconstruction of historical SSTs during WW2 comes from combining all groups of SSTs after adjusting for groupwise offsets (Fig. 1c) and gives a WW2WA of 0.13°C (0.01° to 0.26°C). R4 can be contrasted with the nonadjusted R1 reconstructions having a WW2WA of 0.41°C (Fig. 1b).

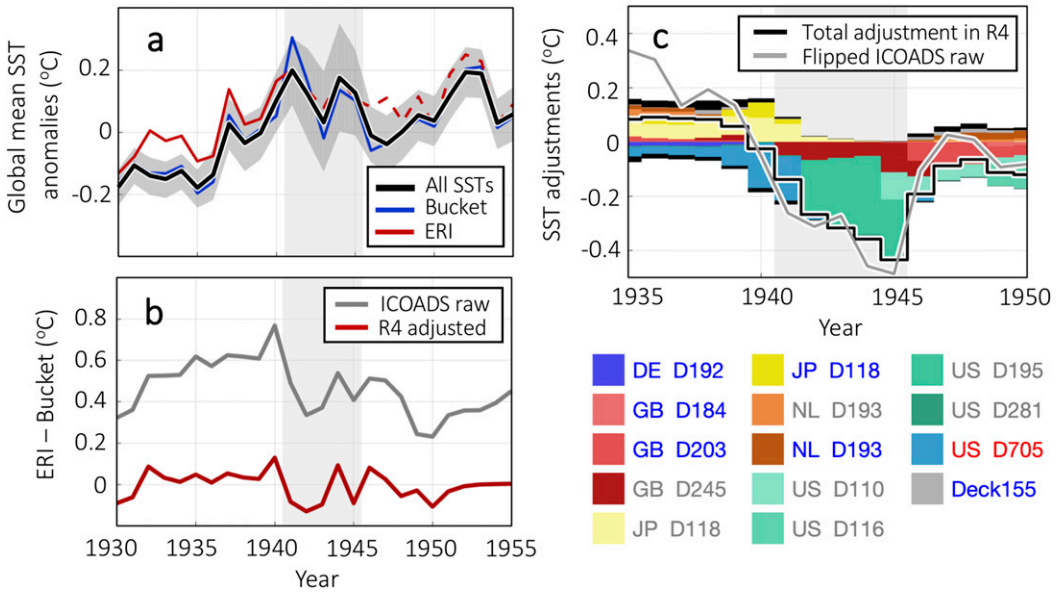

FIG. 5. Groupwise SST adjustments. (a) As in Fig. 1b, but after adjusting for groupwise offsets. The WW2WA reduces to 0.13°C (black; R4) with a 95% c.i. of 0.01 to 0.26°C (gray shading). Remaining offsets between groupwise adjusted bucket and ERI SSTs arise from different spatial coverage of the two methods. (b) The difference between collocated ERI minus bucket SSTs drops from approximately 0.5°C in raw ICOADS (gray; R1) to being centered on zero after groupwise adjustments (red; R4). (c) Groupwise decomposition of SST adjustments (stacked bars; R4). Adjustments during WW2 foremost relate to U.S. Navy ship logs (deck 195) and G.B. Royal Navy ship's logs (deck 245). Groups having fewer than 100 000 measurements in 1935–49 are clustered and shown in black for the purpose of this visualization.

Groupwise adjusted SSTs also show a smaller WW2WA in bucket-only and ERI-only estimates of the WW2WA (Fig. 5a) that are, respectively, 0.15°C (0.06° to 0.24°C) and 0.07°C (−0.08° to 0.24°C). As expected, collocated ERI minus bucket difference decreases from an average of 0.48°C over 1936–50 in raw ICOADS to being centered on zero after groupwise adjustments (Fig. 5b).

The diminished WW2WA in adjusted SSTs largely reflects adjustments of U.S. deck 195 that features a warm offset of 0.43°C (0.17° to 0.68°C) and whose adjustment alone revises the WW2WA from 0.41°C in raw ICOADS to 0.22°C (Fig. 5c). Also of note is U.K. deck 245, which has offsets of −0.31°C (−0.53° to −0.09°C) before 1940 and 0.25°C (0.03° to 0.47°C) between 1941 and 1947 (Fig. 4). The adjustment of deck 245 further diminishes the WW2WA by 0.06°C (Fig. 5c) with local decreases of more than 0.4°C over the Indian Ocean and the Pacific warm pool.

4. R5: Reconstruction using daytime-only measurements

A second effect that we examine stems from recognition by Folland et al. (1984) that nighttime marine air temperatures (MAT) were likely measured inboard during WW2. Folland et al. (1984) state that "the reason is thought to be that it was forbidden, at least on UK ships, to shine a torch in an exposed place, so night MAT was observed well inboard, with consequential larger heating errors" (p. 672). Inboard measurements may have been operationally required to minimize light pollution and potential detection by enemy ships or submarines. For the same reason, water temperatures inside buckets were likely to have been read inboard, with warmer indoor air temperatures and lower wind speeds expected to lead to less sensible and evaporative cooling.

There are five additional lines of evidence that point to nighttime SSTs measured using buckets being anomalously warm and taken inboard during WW2. First, we examine nighttime and daytime-only SSTs coming from bucket and ERI measurements. Day and nighttime observations are identified using the ICOADS night–day flag (ND). Whereas daytime bucket SSTs show a WW2WA of 0.09°C, the nighttime estimate indicates the WW2WA being 0.32°C (Fig. 6a), indicating nighttime estimates as the source of a larger anomaly.

Second, nighttime bucket SSTs reverse from being colder than collocated daytime temperatures by −0.20°C during the five years before and after WW2, as expected regardless of bucket design (Chan and Huybers 2020), to being 0.02°C warmer during WW2. The inversion of the day–night difference in bucket SSTs during WW2 is mainly attributable to British Navy ships from deck 204 that contribute more than 75% of open-ocean bucket SSTs from 1942 to 1945. SST observations from buckets that are concentrated near shore, such as deck 720 (Deutscher Wetterdienst Marine Meteorological Archive), have little overall influence on global SST estimates after gridding. Accordingly, the warmest anomalies in WW2 nighttime bucket SSTs are found over the Indian Ocean and the extratropical Atlantic (Fig. 6c).

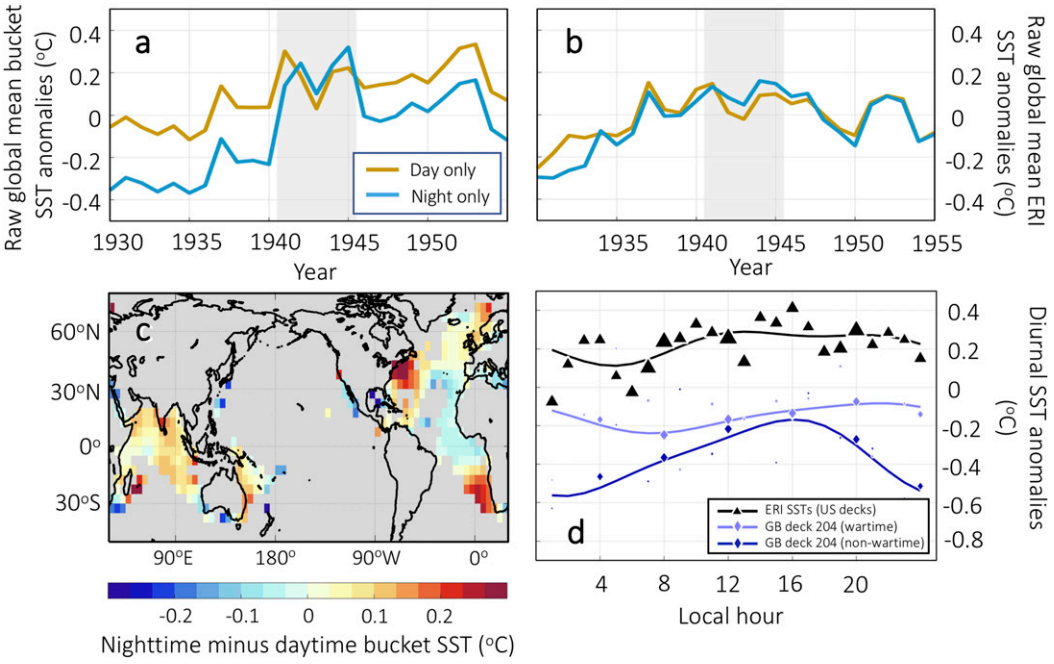

FIG. 6. Nighttime vs daytime bucket SSTs. (a) During WW2, nighttime bucket SSTs in raw ICOADS 3.0 (light blue) are 0.32°C warmer than in the surrounding 10 years and are 0.02°C warmer than daytime bucket SSTs (orange). (b) As in (a), but for ERI SSTs, which show no apparent WW2WA. (c) Spatially, nighttime minus daytime bucket SSTs (shading) are positive over the Indian Ocean and the extratropical Atlantic during WW2 (1942–45). Grid boxes having less than 3 months of data are displayed in gray. For visualization purposes, results are spatially smoothed using a nine-grid 2D convolutional smoother. (d) Nighttime SSTs from British deck 204 are anomalously warm during the war (1941–45; light blue) relative to those measured during 1936–40 and 1946–50 (dark blue). To obtain diurnal anomalies in (d), all measurements from deck 204 are binned and averaged according to local hour. Smoothed diurnal cycles are determined with once- and twice-per-day sinusoidal harmonics using a least squares fitting weighted by numbers of measurements in each hourly bin. Also shown are diurnal anomalies of ERI SSTs estimated from explicitly indicated ERI measurements and U.S. measurements with unknown methodology (black).

Third, and more specifically, the diurnal cycle of SSTs in deck 204 shifts from having peak temperatures at 1600 local time (LT) in the 5 years prior and after WW2 to 2000 LT during WW2, and the overall amplitude of the diurnal cycle decreases (Fig. 6d). Here, the diurnal cycle of deck 204 is estimated by directly binning all available SST anomalies by local hours. This approach allows for using all data but is more susceptible to noise contributions from changes in systematic offsets associated with individual ships relative to our typical approach of assessing the diurnal cycle (e.g., in Chan and Huybers 2020). If we compare the diurnal amplitude of deck 204 relative to collocated climatological amplitudes estimated from drifting buoys (Chan and Huybers 2019), the averaged anomalous amplitude of deck 204 decreases from being 0.03°C larger than drifters in the 5 years before and after WW2 to being 0.03°C smaller during WW2.

Fourth, it is possible to rule out other instrumental or physical causes for anomalously warm nighttime bucket SSTs. The smaller diurnal amplitude found in deck 204 is unlikely to be related to switching to ERI measurements because the average temperature of daytime measurements remains consistent with bucket measurements taken before and after WW2 and remains cooler than known ERI measurements (Fig. 6d). Furthermore, we are unaware of a physical mechanism that would cause nighttime SSTs to be routinely warmer, when averaged over a year, than daytime SSTs. A climatological cause of the WW2WA in nighttime bucket measurements is also contradicted by the lack of a warm anomaly in nighttime ERI measurements (Fig. 6b).

Finally, the inference that observation protocols were changed to avoid light pollution suggests that sailors would favor daytime over nighttime measurements. Indeed, the percentage of daytime bucket SSTs relative to all available bucket SST observations increases from 52% in the five years before and after WW2 to 62% during WW2. These additional lines of evidence provide a strong indication that nighttime bucket SST during WW2 were measured inboard. We also note that the shift from taking 55% of all available SST observations, including bucket, ERI, and unknown types, during the daytime in the five years before and after WW2 to 61% during WW2 makes only a minor contribution to the WW2WA. Sampling hourly-resolved climatological diurnal cycles from drifters indicates that such a shift contributes only 0.005°C to the observed warm anomaly.

Although the LME method could be further extended to temporally resolve anomalies in nighttime biases, building in such flexibility would force nighttime temperatures to be consistent with daytime temperatures. This approach would make nighttime temperatures effectively uninformative because daytime and nighttime measurements have approximately the same spatial and temporal coverage (Fig. S1 in the online supplemental material). Instead, we simply repeat our analysis excluding nighttime measurements. Specifically, we use 24.1 million pairings of daytime SST measurements between 1850 and 2014 with 1.0 million of these pairs available between 1935 and 1949. Whereas using both day and night measurements gives a 0.13°C (0.01 to 0.26°C) WW2WA (R4), the daytime-only analysis gives a WW2WA of 0.09°C (−0.01° to 0.18°C, R5; Table 1, Fig. 1c).

We compare the WW2WA in our various observational estimates against the variability found in CMIP5 models. To compare the WW2WA against simulated internal variability, we regrid a total of 25 236 years of simulated preindustrial SSTs from 42 available CMIP5 models (Taylor et al. 2012) to a common 5° resolution (see Table S1 for a list of models used). Simulations are then divided into 15-yr segments, with each segment masked using the 1936–50 least common coverage between HadSST4 and R5 on a month-by-month basis. The difference between the central five years and the surrounding 10 years is calculated for individual preindustrial segments to estimate the range of internal variability. The CMIP5 runs indicate a 95% range of internal variability being −0.10° to 0.10°C. A similar analysis based on 15 453 years of preindustrial runs from 27 available CMIP6 models gives a 95% range of internal variability of −0.11° to 0.10°C, where both the CMIP5 and CMIP6 results are consistent with R5 (Fig. 7b and Table 1). Also consistent with simulated internal variability are bucket-only estimates of daytime SST that have a WW2WA of 0.08°C (−0.02° to 0.17°C) and ERI-only estimates that have an anomaly of 0.03°C (−0.07° to 0.14°C). Time series of groupwise-adjusted daytime SSTs using only bucket or ERI measurements are shown in Fig. S2 in the supplemental material.

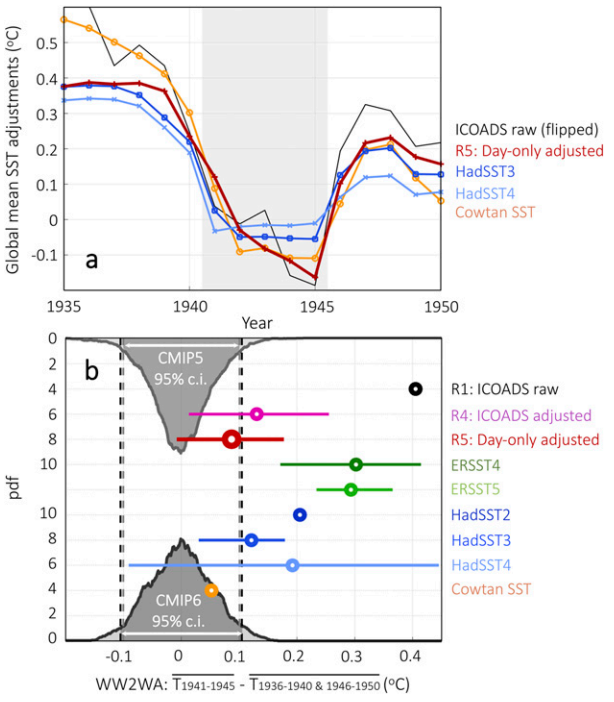

FIG. 7. Comparison with other SST estimates and internal variability in simulations. (a) Global and annual average SST adjustments from our linear-mixed-effect method for daytime observations (R5; red) compared against those from HadSST3 (dark blue), HadSST4 (light blue), and Cowtan SST (orange). Groupwise adjustments are calculated as daytime adjusted SSTs (R5) minus raw SSTs (R1), where the latter uses both daytime and nighttime measurements, and the difference is shifted positive by 0.22°C for purpose of comparison. Also shown are raw SST anomalies (R1, black) flipped for purposes of comparison. (b) WW2WA estimates (markers) and 95% c.i. (bars) for different SST estimates. The WW2WA is calculated as the mean difference between the average over 1941–45 and the average over 1936–40 and 1946–50. Best estimates from groupwise adjusted daytime-only SSTs (R5; red) and Cowtan SST (orange; referenced to coastal air temperatures) are within the 95% c.i. of internal variability (grayscale distributions) estimated from both CMIP5 and CMIP6 preindustrial control simulations.

5. Further discussion and conclusions

Our groupwise intercomparison indicates that ERI and bucket groups have an average offset of 0.53°C (0.25° to 0.72°C) during 1935–49. Such a difference is nearly 0.1°C higher than the wartime difference that averages 0.45°C across ensemble members as implemented in HadSST4 (Kennedy et al. 2019). Furthermore, our analysis indicates that unknown U.S. and U.K. measurements from 1942 to1945, which account for 98% of unknown wartime measurements, are offset warm. In HadSST4, an average of 12.5% of SSTs with missing method information were assumed to come from buckets and adjusted positively. The smaller offset assumed between the bucket and ERI SSTs and a higher percentage of observations assumed to come from buckets explains the re-emergence of the WW2WA from HadSST3 to HadSST4. Our findings indicate that HadSST3 provides a more accurate assessment of the WW2WA. In addition, whereas HadSST4 specifies large uncertainties during WW2, our result reduces the standard error of WW2WA from 0.14°C in HadSST4 to 0.05°C in R5 on account of attributing more variance in the raw data to systematic offsets that can be corrected (Fig. 7b). Our LME results also indicate that R5 has slightly lower uncertainty than R4, even though using daytime-only measurements approximately halves the sample size. Sampling and random errors are negligible at global and decadal scales, and we infer that the greater uncertainty in R4 is driven by systematic errors associated with nighttime measurements, possibly associated with variable inboard offsets.

The SST adjustments obtained through our LME approach agree with an independent estimate arrived at using nearshore, land-station data (Cowtan et al. 2018; Fig. 7a). Whereas the approach of Cowtan et al. (2018) requires average SSTs to agree with land-station data, our analysis shows that WW2WA is an artifact arising from specific groups and features of SST measurements. Once these artifacts are accounted for, both

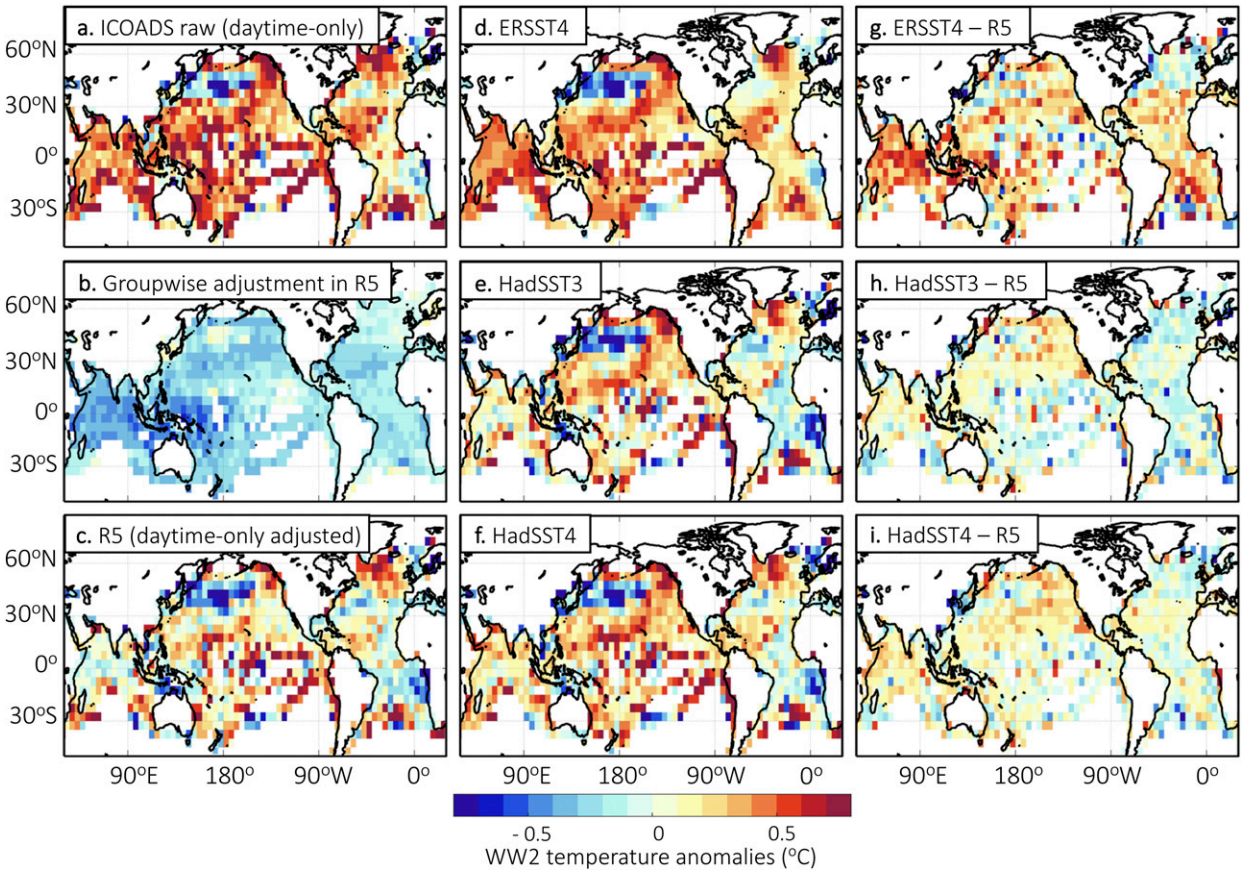

FIG. 8. Patterns of WW2WA. (left),(center) Maps of temperature differences between 1941 and 1945 and the mean over 1936–40 and 1946–50 for (a) raw daytime SSTs in ICOADS 3.0, (b) groupwise adjustments of daytime SSTs, (c) adjusted daytime SSTs [R5; values in (a) and (b) summed], (d) ERSST4, where ERSST5 has a similar pattern, (e) HadSST3, and (f) HadSST4. A grid box is shown if it has at least 2 years of data, each with at least 6 months of observations, with 1 year during 1941–45 and another in the 10 surrounding years. All SST estimates are masked by the least common coverage between HadSST4 and daytime-only estimates (R5) on a month-by-month basis. (right) Difference between daytime groupwise adjusted estimates (R5) and other datasets for (g) ERSST4, (h) HadSST3, and (i) HadSST4.

nearshore land temperatures and SSTs are in good agreement. Our results thus help confirm the global-average results reported by Cowtan et al. (2018). On the other hand, compared with ERSST estimates that are referenced to MATs, our results cast doubt on the reliability of using nighttime or daytime MATs for adjusting SSTs during WW2. Another discrepancy is reported in a more modern context whereby the global mean of Hadley Centre nighttime marine air temperatures (HadNMAT2.0.1.0) appears to be significantly colder ($P < 0.05$) than SSTs by more than 0.08°C after the 1990s (Kennedy et al. 2019).

An important attribute of groupwise adjustments is the ability to resolve regional biases arising from spatially heterogeneous distributions of distinct groups (Fig. 8). Removing groupwise offsets leads to a greater decrease in the WW2WA over the Indian Ocean and Pacific warm pool and smaller decreases over the tropical eastern Pacific and the South Atlantic (Fig. 8b). The spatial correlation of the WW2WA between our adjustments and R5 is $r_s = 0.04$, where the small correlation indicates that the magnitude of the pattern that we remove is appropriate because R5 is nearly free of the estimated pattern of bias. HadSST estimates partially account for biases associated with shifting instruments and have a similar pattern of correction, albeit one that is less complete such that $r_s = -0.15$ (Fig. 8e) for HadSST3 and -0.18 for HadSST4 (Fig. 8f). In contrast, ERSST estimates use a fixed spatial pattern (Huang et al. 2015, 2017) that does not account for patterns associated with groupwise offsets, giving an $r_s = -0.28$ for ERSST4 (Fig. 8d) and $r_s = -0.25$ for ERSST5. The zonally symmetric corrections from Cowtan et al. (2018) also do not capture the patterns of WW2 offsets.

An implication of the removal of WW2WA in our analysis is a more stable and smoothly evolving SST estimate (Table 1). The 1936–50 standard deviation of global-average, annual SST anomalies decreases from 0.24°C in raw ICOADS (R1) to 0.07°C (0.06° to 0.11°C) in the adjusted daytime-only estimates (R5). Such subdecadal variability is consistent with estimates from HadSST3 and Cowtan SST and lies within the 95%

confidence interval of CMIP5 and CMIP6 historical simulations that we analyze. In contrast, the HadSST4 median estimate has a standard deviation of 0.12°C that is larger than 93 out of 94 CMIP5 historical simulations and 37 out of 38 CMIP6 historical simulations, and ERSST median estimates have standard deviations of 0.16°–0.17°C that are higher than all CMIP5 and CMIP6 historical simulations (Table 1). On regional scales, the effect of groupwise adjustments are smaller than other sources of variability—including from physical changes, sampling uncertainty, and random measurement errors—such that the 1936–50 variance on 5° × 5° grids decreases, on average, by only 12%.

In sum, our results help confirm that the WW2WA in instrumental SST estimates is a data artifact that arises from instrumental changes (Thompson et al. 2008). We identify U.S. and U.K. ships as the primary origin of the WW2WA. Warm biases in WW2 nighttime bucket SSTs are also identified. Adjusting for these offsets removes the WW2WA and leads to a more homogeneous trend in SSTs. Our results highlight the importance of resolving systematic errors in SSTs and reconcile the largest existing discrepancy between historical surface temperatures and model estimates (Folland et al. 2018). The fact that our independently derived adjustments to the SST record leads to consistency with model simulations of SST variations during WW2 gives greater confidence in predictions based on such models.

Ongoing work to recover historic SST will make more wartime data available for U.S. deck 195, which currently only has measurements that were collected at 0800, 1200, and 2000 LT included in ICOADS. Metadata that allow for distinguishing between types of naval ships, such as destroyers and destroyer escorts, are also being recovered (Hawkins et al. 2020). Incorporation of these additional SST observations and metadata in future work should permit for a more accurate and detailed adjustments of systematic offsets and more accurate estimates of historical SST.

Acknowledgments. We thank three anonymous reviewers for their detailed and thoughtful feedback. Conversations with Carl Wunsch and Elizabeth Kent also improved the content of this manuscript. This study was supported by the Harvard Global Institute.

Data availability statement. All datasets used in this study are available as follows: ERSST4 and a 1000-member ensemble (https://psl.noaa.gov/data/gridded/data.noaa.ersst.v4.html; last access: 18 April 2020; doi:10.7289/V5KD1VVF), ERSST5 and a 1000-member ensemble (https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.ersst.v5.html; last access: 5 April 2020; doi:10.7289/V5T72FNM), Cowtan SST (https://www-users.york.ac.uk/~kdc3/papers/evaluating2017/methods.html; last access: 7 April 2020), HadSST2 (https://www.metoffice.gov.uk/hadobs/hadsst2/data/download.html; last access, 18 April 2020), HadSST3.1.1.0 and a 100-member ensemble (https://www.metoffice.gov.uk/hadobs/hadsst3/data/download.html; last access: 13 February 2020), and HadSST4.0.0.0 and a 200-member ensemble (https://www.metoffice.gov.uk/hadobs/hadsst4/data/download.html; last access: 5 April 2020). HadSST.4.0.0.0 data are ©British Crown Copyright, Met Office 2021, provided under an Open Government License, http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/. Monthly CMIP5 SST (tos) outputs regridded to 2.5° resolution are available from the ETH repository (contact Jan.Sedlacek@env.ethz.ch or cmip5-archive@env.ethz.ch for access; last access: 21 February 2019). Monthly CMIP6 outputs are from the ESGF portal (https://esgf-node.llnl.gov/search/cmip6/; last access: 8 January 2021). Raw and groupwise adjusted SSTs (R1–5), as well as key results in this manuscript, are available from the Harvard Dataverse repository, https://doi.org/10.7910/DVN/RJLBOQ. The full reference for ICOADS3.0 follows: Research Data Archive/Computational and Information Systems Laboratory/National Center for Atmospheric Research/University Corporation for Atmospheric Research, Physical Sciences Laboratory/Earth System Research Laboratory/OAR/NOAA/U.S. Department of Commerce, Cooperative Institute for Research in Environmental Sciences/University of Colorado, National Oceanography Centre/University of Southampton, Met Office/Ministry of Defence/United Kingdom, Deutscher Wetterdienst (German Meteorological Service)/Germany, Department of Atmospheric Science/University of Washington, Center for Ocean–Atmospheric Prediction Studies/Florida State University, and National Centers for Environmental Information/NESDIS/NOAA/U.S. Department of Commerce. 2016, updated monthly. International Comprehensive Ocean–Atmosphere Data Set (ICOADS) Release 3, Individual Observations. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory (https://doi.org/10.5065/D6ZS2TR3, accessed 3 October 2018).
Codes required to reproduce the full analysis and display items are available from a Github repository, https://github.com/duochanatharvard/World-War-II-Warm-Anomaly.

APPENDIX A

Setup and Implementation of the LME Methodology

The linear-mixed-effect (LME) method compares nearby measurements from different groups delineated according to method, nation, and deck numbers and estimates systematic offsets among these groups relative to the mean of all paired measurements. We pair SSTs that are within 300 km and 2 days of one another, but results are not qualitatively sensitive if smaller thresholds are used in the pairing process. Expected offsets associated with spatial and temporal displacement are removed using a climatology derived from high-resolution satellite and drifter measurements (see section 2c in Chan and Huybers 2019). The resulting residual SST differences $\delta\mathbf{T}$ are modeled using Eq. (1) in the main text, which is repeated here for ease of reference:

$$\delta\mathbf{T} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}_y\boldsymbol{\beta}_y + \mathbf{Z}_r\boldsymbol{\beta}_r + \boldsymbol{\epsilon}.$$

Matrix $\mathbf{X}$ is a design matrix for $\boldsymbol{\alpha}$, or the fixed-effect offsets among groups. Entries of $\mathbf{X}$ are 1, −1, and 0, specifying which

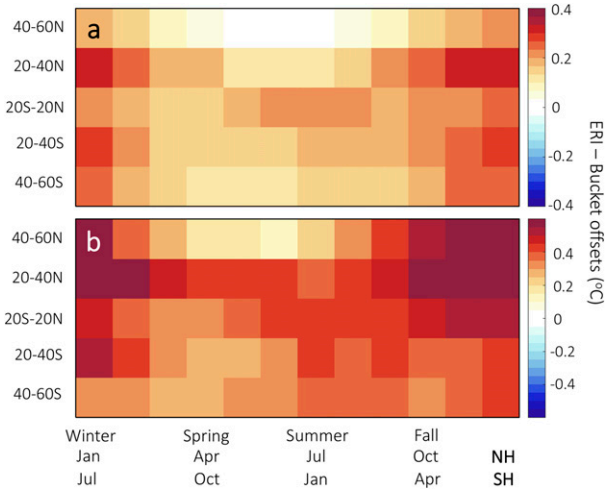

FIG. A1. Differences between bucket and ERI offsets. (a) Differences in ERI and bucket groupwise offsets as a function of season and latitude. In general, ERI–bucket differences are higher in winter, consistent with the expectation of greater wintertime heat loss from buckets. During summer, ERI–bucket differences are less positive in the extratropics than the tropics, consistent with reduced bucket heat loss in warm and humid air and active heating by incident solar radiation (Folland and Parker 1995). This latitude dependence is also consistent with warm high-latitude summertime bucket SSTs as reported in Kennedy et al. (2019) and Chan and Huybers (2020). The differences shown in (a) rely on estimates for all available groups during 1850–2014. Regional effects are first averaged within latitude bands for individual groups to obtain an estimate of latitude effects. Fixed plus latitude effects are then averaged over all bucket and ERI groups that are explicitly determined from ICOADS or WMO No. 47 metadata. Across-group averages are weighted by the total number of measurements in individual groups. (b) As in (a), but only for groups that contribute during 1935–49. A similar seasonality and latitude dependence is again found in this interval, but ERI–bucket differences are larger, possibly because of using less insulated buckets or ERI design.

data are intercompared (Fig. 3). The term $\boldsymbol{\beta}_y$ represents how offsets in each group vary around its mean offset across 5-yr increments since 1850, and $\boldsymbol{\beta}_r$ indicates how offsets specified for 17 different regions vary about the spatial mean (see Fig. 3 in Chan and Huybers 2019). Individual elements in $\boldsymbol{\beta}_y$ and $\boldsymbol{\beta}_r$ are assumed to follow normal distributions such that the model relaxes solutions toward zero when fewer data are available for estimating individual yearly effects. Matrix $\mathbf{Z}_y$ is a design matrix similar to $\mathbf{X}$ but has more columns because it indicates not only which groups are intercompared but also when comparisons are made. Matrix $\mathbf{Z}_r$ is similar but includes where comparisons are made.

Our setup of the LME method permits partly resolving variations in offsets due to geographically varying measurement environment (Fig. A1). Higher-order interactions that involve group, year, and region are not accounted for in this model to limit the number of free parameters. Note that the model does not explicitly resolve seasonal variations in offsets. To estimate seasonality, we fit the model on subsets of three consecutive months and combine 12 successive analyses to cover the full year. Southern Hemisphere SSTs are shifted by six months to account for different seasons between hemispheres (Chan and Huybers 2020).

There are 45.8 million SST pairs in our analysis using all daytime and nighttime measurements over the years 1850–2014. Inversion of error matrices having this number of dimensionality (i.e., 45.8M × 45.8M) is prohibitive despite use of state-of-the-art, large-memory systems. To increase computational efficiency, pairs are averaged according to combinations of groups, regions, and years. The residual error of the kth averaged pair $\boldsymbol{\varepsilon}_k$ are assumed to follow $N(0, \overline{\sigma}_k^2)$, where variance arises from three sources:

$$\overline{\sigma}_k^2 = \frac{2\sigma_o^2}{n_k} + \frac{2\sigma_s^2}{n_k^x} + \frac{\sum \sigma_c^2}{n_k^2}. \tag{A1}$$

The first term on the right-hand side denotes random observational errors, where n_k is the number of pairs in the kth average. Random observational error for individual ship-based SSTs σ_o^2 is estimated to be 0.93°C^2 using all bucket and ERI measurements following the method in Chan and Huybers (2019, see their Fig. 10). In our earlier estimate based only on bucket measurements (Chan et al. 2019) σ_o^2 was 0.86°C^2. The second term, $2\sigma_s^2/n_k^x$, denotes partially correlated observational errors, including systematic biases associated with individual ships. Because ship information is not always available, n_k^x is used to approximate effective numbers of ships within the kth average. Here, σ_s^2 based on all bucket and ERI measurements is 0.50°C^2 and x is 0.57, whereas in Chan et al. (2019) we obtained values of 0.38°C^2 and 0.57 using bucket SSTs. Finally, σ_c^2 denotes uncertainties associated with physical SST variations. Estimates of σ_c^2 account for interannual variance and covariance of physical SSTs as a function of location, month, and displacement. More details of the error structure can be found in section 5a of Chan and Huybers (2019).

After averaging, we use an iterative numerical algorithm (Harville 1977), as implemented in the Matlab function "fitlmematrix", to estimate the LME model. In particular, we estimate three hyperparameters: the variance of all yearly effects, σ_y^2; the variance of all regional effects, σ_r^2; and a scaling parameter applied to all $\overline{\sigma}_k^2$ to capture uncertainties in our error model [Eq. (A1)]. The algorithm iterates between these hyperparameters and the fixed effect $\boldsymbol{\alpha}$ until reaching a maximum likelihood estimate. The uncertainty of the fixed effect is estimated accounting for residual errors and covariance between random effects, whereas the mean and uncertainty of random effects are estimated using multivariate Gaussian distributions conditional on fixed-effect estimates. Associated equations are all detailed in the appendix of Chan and Huybers (2019).

The significance of groupwise offsets is estimated using a two-sided Z test from 1000 sets of groupwise randomized offset estimates. Offsets are realized according to the mean and uncertainties associated with fixed and random effects. A Bonferroni correction is also applied to account for the increased probability of incorrectly rejecting true null hypotheses in multiple hypothesis testing, which is carried out by lowering the threshold of the p value to be 0.05/66, where 66 is

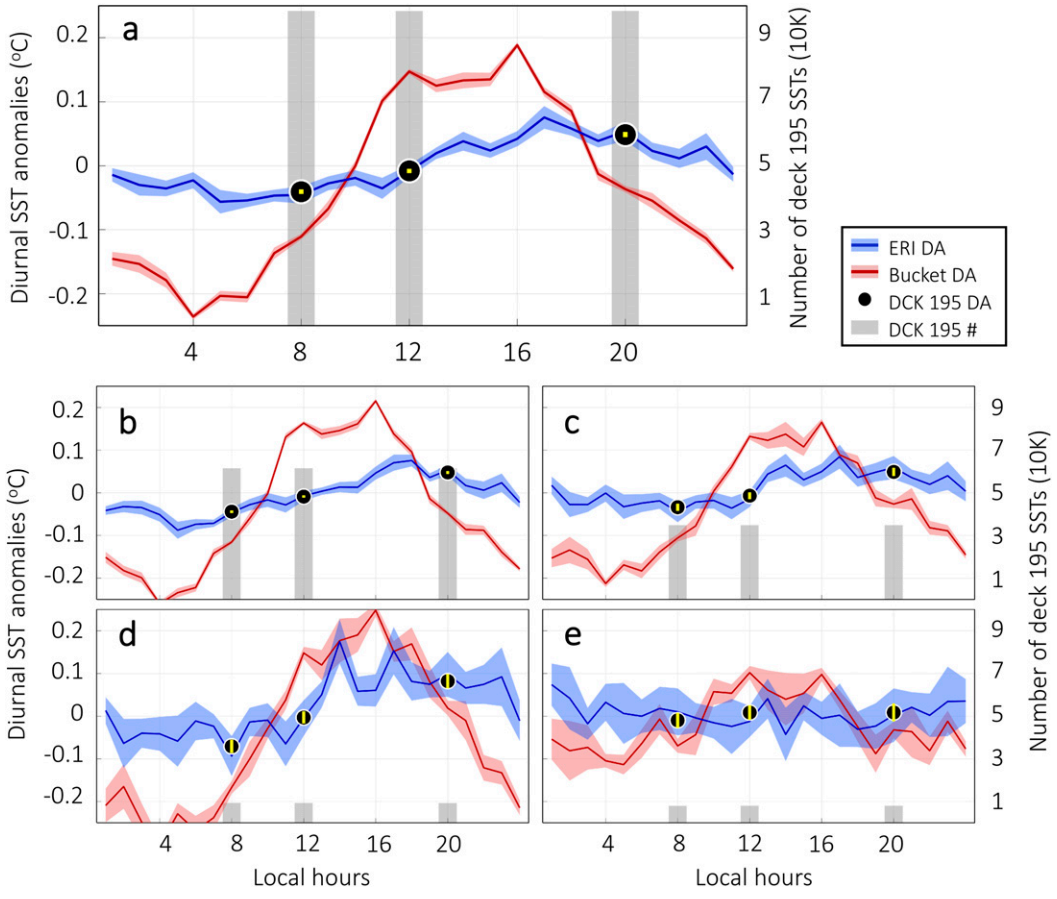

FIG. B1. Diurnal amplitude of U.S. deck 195. Diurnal anomalies for deck 195 (black circles; yellow bar shows the 95% c.i.), bucket (red; shading shows the 95% c.i.), and ERI (blue; shading shows the 95% c.i.) measurements. All shown diurnal anomalies are relative to the respective average over each of the three times in a day that deck 195 reports measurements, i.e., 0800, 1200, and 2000 LT. Diurnal cycles of bucket and ERI SSTs are based on tracked ships that have at least one measurement in each 6-hourly bin of a day over 1935–49. Unknown U.S. measurements from decks other than 195 are assumed to be ERI measurements because of their small diurnal amplitudes and warm groupwise offsets (Fig. 4 in the main text). Individual panels are for (a) all available measurements, (b) the tropics (30°S–30°N), (c) the NH extratropics (30°–60°N), (d) NH extratropical summer (JJA), and (e) NH extratropical winter (DJF). In all cases, the diurnal cycle of deck 195 follows ERI SSTs more closely than bucket SSTs.

the number of groups from 1935 to 1949. We also use these 1000 sets of randomized offset estimates to generate a 1000-member ensemble of adjusted monthly gridded SSTs, which permits accounting for error covariance reflecting the spatial and temporal coverage of individual groups.

APPENDIX B

Using Diurnal Variations to Infer Measurement Type for Deck 195 Measurements

All SST measurements from U.S. deck 195 are sampled at 0800, 1200, and 2000 LT, whereas elsewhere we have required that each 6-hourly bin of a day has at least one measurement (Carella et al. 2018; Chan and Huybers 2020). Because deck 195 contains 24% of observations within the WW2 interval, however, we take an alternative approach that uses diurnal cycles from bucket and ERI measurements as basis functions (Fig. B1). Specifically, the diurnal cycle of deck 195 is represented as a linear combination of bucket and ERI cycles, with the mixture determined using least squares fitting. When using all available SST measurements, the best fit yields 100% ERI and 0% bucket (Fig. B1a), equivalent to a diurnal amplitude of 0.05°C that is 0.07°C smaller than that of drifting buoys. Our inference that deck 195 is consistent with purely ERI SSTs is robust to dividing the analysis to focus on individual regions and seasons, where in each case the observed diurnal variations are consistent with other ERI data and highly inconsistent with bucket observations (Figs. B1b–e).

REFERENCES

Bindoff, N. L., and Coauthors, 2013: Detection and attribution of climate change: From global to regional. *Climate Change 2013: The Physical Science Basis*, Cambridge University Press, 867–952.

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

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

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.

——, and ——, 2020: Systematic differences in bucket sea surface temperatures caused by misclassification of engine room intake measurements. *J. Climate*, **33**, 7735–7753, https://doi.org/10.1175/JCLI-D-19-0972.1.

——, 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.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

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.

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

——, O. Boucher, A. Colman, and D. E. Parker, 2018: Causes of irregularities in trends of global mean surface temperature since the late 19th century. *Sci. Adv.*, **4**, EAAO5297, https://doi.org/10.1126/SCIADV.AAO5297.

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.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, https://doi.org/10.1029/2010RG000345.

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

Hausfather, Z., K. Cowtan, D. C. Clarke, P. Jacobs, M. Richardson, and R. Rohde, 2017: Assessing recent warming using instrumentally homogeneous sea surface temperature records. *Sci. Adv.*, **3**, e1601207, https://doi.org/10.1126/SCIADV.1601207.

Hawkins, E., P. Brohan, G. Compo, K. Wood, and M. Mark, 2020: Old Weather—WW2. Accessed 22 November 2020, https://www.zooniverse.org/projects/krwood/old-weather-ww2/about/research.

Hegerl, G. C., S. Brönnimann, A. Schurer, and T. Cowan, 2018: The early 20th century warming: Anomalies, causes, and consequences. *Wiley Interdiscip. Rev.: Climate Change*, **9**, e522, https://doi.org/10.1002/wcc.522.

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

——, 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, G. S., P. A. Stott, and N. Christidis, 2013: Attribution of observed historical near-surface temperature variations to anthropogenic and natural causes using CMIP5 simulations. *J. Geophys. Res. Atmos.*, **118**, 4001–4024, https://doi.org/10.1002/jgrd.50239.

Kennedy, J., N. Rayner, R. Smith, D. Parker, and M. Saunby, 2011a: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *J. Geophys. Res.*, **116**, D14103, https://doi.org/10.1029/2010JD015218.

——, ——, ——, ——, and ——, 2011b: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *J. Geophys. Res.*, **116**, D14104, https://doi.org/10.1029/2010JD015220.

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

——, N. A. Rayner, D. I. Berry, M. Saunby, B. I. Moat, J. J. Kennedy, and D. E. Parker, 2013: Global analysis of night marine air temperature and its uncertainty since 1880: The HadNMAT2 data set. *J. Geophys. Res. Atmos.*, **118**, 1281–1298, https://doi.org/10.1002/jgrd.50152.

——, 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.

Maher, N., A. Sen Gupta, and M. H. England, 2014: Drivers of decadal hiatus periods in the 20th and 21st centuries. *Geophys. Res. Lett.*, **41**, 5978–5986, https://doi.org/10.1002/2014GL060527.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, https://doi.org/10.1029/2011JD017187.

Pfeiffer, M., J. Zinke, W.-C. Dullo, D. Garbe-Schönberg, M. Latif, and M. Weber, 2017: Indian Ocean corals reveal crucial role of World War II bias for twentieth century warming estimates. *Sci. Rep.*, **7**, 14434, https://doi.org/10.1038/s41598-017-14352-6.

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

Stevens, B., 2015: Rethinking the lower bound on aerosol radiative forcing. *J. Climate*, **28**, 4794–4819, https://doi.org/10.1175/JCLI-D-14-00656.1.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Thompson, D. W., J. J. Kennedy, J. M. Wallace, and P. D. Jones, 2008: A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature*, **453**, 646–649, https://doi.org/10.1038/nature06982.

——, J. M. Wallace, P. D. Jones, and J. J. Kennedy, 2009: Identifying signatures of natural climate variability in time series of global-mean surface temperature: Methodology and insights. *J. Climate*, **22**, 6120–6141, https://doi.org/10.1175/2009JCLI3089.1.

Vose, R. S., and Coauthors, 2012: NOAA's merged land–ocean surface temperature analysis. *Bull. Amer. Meteor. Soc.*, **93**, 1677–1685, https://doi.org/10.1175/BAMS-D-11-00241.1.

Woodruff, S. D., and Coauthors, 2011: ICOADS release 2.5: Extensions and enhancements to the surface marine meteorological archive. *Int. J. Climatol.*, **31**, 951–967, https://doi.org/10.1002/joc.2103.

York, D., N. M. Evensen, M. L. Martinez, 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.