HDSR
HARVARD DATA SCIENCE REVIEW

Volume 3 Issue 1
DOI: 10.1162/99608f92.edcee38f
ISSN: 2644-2353

Combining Statistical, Physical, and Historical Evidence to Improve Historical Sea-Surface Temperature Records

Duo Chan[†,*]

[†] Department of Earth and Planetary Sciences, Harvard University

Abstract. Reconstructing past sea-surface temperatures (SSTs) from historical measurements containing more than 100 million ship-based observations taken by over 500,000 ships from more than 150 countries using a variety of methodologies creates a wide range of historical, scientific, and statistical challenges. The reconstruction of historical SSTs for studying climate change is particularly challenging because SST measurements are uncertain and contain systematic biases of order 0.1°C to 1°C—these systematic biases are in the range of the historical global warming signal of approximately 1°C. The biases are complicated and have generally been addressed using simplified corrections. In this review, I introduce a history of SST observations, review a statistical method developed for quantifying SST biases, and illustrate scientific insights obtained from adjusted SSTs. This article also documents the scientific journey of my Ph.D. work. As a result, I report personal stories on both successes, difficulties, and setbacks along the way. The statistical method for correcting SSTs (i.e., a linear-mixed-effect intercomparison framework) depends on identifying systematic offsets between intercomparable groups of SST observations. Combining estimated offsets with physical and historical evidence has allowed for correcting discrepancies associated with SSTs, including the North Atlantic warming twice as fast as the North Pacific in the early 20th century and anomalously warm SSTs during World War II. Corrections also permit better hindcasting of Atlantic hurricanes. I conclude with some discussion on how the SST records might be further improved. Given the importance of SSTs for understanding historical changes in climate, I hope that this review can help others appreciate challenges that are present and spark some interest and ideas for further improvement.

Keywords: climate reconstruction, sea-surface temperature, bias correction, data homogenization, linear-mixed-effect model.

[*] duochan@g.harvard.edu

Media Summary

To better predict what climate change will look like in the future, it is crucial to know how and why climate has changed in the past. One essential component of climate change is the ocean, for which we have more than 200 years of ship-based temperature measurements made at the ocean surface. However, biases in early sea-surface temperatures have limited their usage in climate studies. These biases are similar in magnitude to historical warming, and they vary with measurement methods, instruments, protocols, and even postprocessing and data-keeping practices. The question is, therefore, can we remove the complicated biases and obtain a sea-surface temperature estimate that is accurate enough to study past climate change?

In this article, I review recent progress aimed at correcting sea-surface temperatures for individual nations and data-collecting groups. I introduce a statistical framework that compares nearby measurements and estimates systematic offsets in temperatures among groups. Physical and historical evidence is then combined to understand the origins of significant groupwise differences detected by the statistical method. Correcting data leads to spatially more homogeneous warming in the early 20th century and removes anomalously warm sea temperatures during World War II, which reconciles existing model-data discrepancies and brings observations into consistency with current knowledge of climate forcing and variability. Beyond the ocean itself, adjusted sea-surface temperatures also allow atmospheric models to simulate more realistic historical variations in North Atlantic hurricanes, showing potential for improving predictions of these high-impact events. This review also demonstrates the importance of understanding the social context and history of how data are collected and postprocessed. When data and models disagree, keeping an awareness of potential flaws in the quality of data appears to be a necessary practice.

1. Introduction

Sea-surface temperature (SST), typically defined at ocean depth of 20–30 cm (Kennedy et al., 2019), is a crucial quantity for studying the Earth's climate. Estimates of historical SSTs to an accuracy of 0.05°C at the global scale and 0.1°C at regional scales are required for a wide range of climate applications (Kent & Berry, 2008), which include depicting past climate change (Hartmann et al., 2013), attributing anthropogenic versus internal climate variability (Bindoff et al., 2013), and understanding changes in climate and weather events that have far-reaching societal impacts, such as El Niño (Yeh et al., 2009) and hurricanes (Vecchi et al., 2011). Moreover, SSTs are often used as boundary conditions in numerical models to reproduce or hindcast a variety of meteorological phenomena (e.g., Gates et al., 1999).

Despite their importance for climate sciences, estimates of historical SSTs remain highly uncertain (P. Jones, 2016), with disagreements existing between observational estimates and climate-model simulations. One example of such data-model disagreement would be the recent warming hiatus, which refers to a slowdown in the increase of the observed global-mean surface temperature since the late 1990s (Easterling & Wehner, 2009). The hiatus was one of the most popular climate-related research topics in the first half of the 2010s. At that time, state-of-art climate models that were used in the latest Intergovernmental Panel on Climate Change report (IPCC AR5, Taylor et al., 2012) simulated significantly faster warming ($p < 0.05$) than observational estimates (Fyfe et al., 2013; Medhaug et al., 2017). One possible explanation, as suggested by many studies, involves natural climate fluctuations that can uptake heat from the surface into the deep ocean (e.g., X. Chen & Tung, 2014; Kosaka & Xie, 2013). Another plausible explanation, however, is that

recent trends in observed SSTs were underestimated by 0.064°C/decade from 2000 to 2014 due to biases in ship-based measurements (Karl et al., 2015). Correcting SSTs reduces the so-called recent warming hiatus, making the estimate of warming rates since the early 2000s consistent with the rapid warming since the late 1970s (Hausfather et al., 2017; Karl et al., 2015). In another example, after removing contributions from major physical modes of climate variations, Thompson et al. (2008) detected a sudden drop of about 0.3°C in global-mean SSTs immediately after World War II, which they attributed to insufficient corrections of instrumental SST biases.

When data and models disagree, a common practice is to assume that data reflect reality and to look for new theories to enrich the model and explain the data (e.g., using natural climate fluctuations to explain the recent warming hiatus). However, there is always a second and often overlooked possibility: that data contain undetected problems. As we shall see in detail in later sections, in addition to the recent warming hiatus and the artificial temperature drop at the end of World War II, major data problems also exist in SST estimates in terms of patterns of warming in the early 20th century and temperature evolution during World War II. Observed historical SSTs are particularly likely to contain data problems because of complicated biases associated with using various crude methods to collect early measurements and also because of simplified bias corrections employed when generating SST estimates.

1.1. **A Brief History of Measuring SSTs since the 1800s.** Instrumental SSTs have been measured on ships at the ocean surface for more than 200 years, yielding more than 130 million ship-based measurements since the 1850s (Freeman et al., 2017). Such a history is longer than that of studies on anthropogenic climate change; the first estimate of equilibrium warming once doubling atmospheric CO_2 was made by Svante Arrhenius in 1896 (Lapenis, 1998). The history of SST measurement is also much longer than that of dedicated scientific efforts to systematically monitor ocean temperatures, which began in the late 1970s.

Most of these early SST measurements were made not by dedicated researchers but by voluntary and nonscientist sailors from different countries who sailed, for example, as soldiers, merchants, or fishermen (Kennedy, 2014; Kent et al., 2017). These early SST measurements were made for variety of purposes, including pure scientific interest, facilitating navigation, predicting stormy weather, and mapping a climatological summary of the marine environment (Kennedy, 2014). Although not made for purposes of monitoring climate change, ship-based in-situ observations are the only available source of direct measurements of past states at the ocean surface.

Historical SST values were originally recorded in ship logs and were rescued and digitized by a variety of projects and institutions (e.g., Wilkinson et al., 2011). Digitized data from various sources were later put together to construct the International Comprehensive Ocean-Atmosphere Data Set (ICOADS, the most comprehensive modern compilation of available marine meteorological measurements since the 1700s). The digitization of ship logs and the construction of ICOADS spanned several decades (Freeman et al., 2017; Woodruff et al., 1998; Woodruff et al., 1987; Woodruff et al., 2011; Worley et al., 2005), and this initiative was accompanied by revolutions in computer and data-storage technologies. ICOADS3 is the latest version, and efforts continue to recover lost historical data sets and to include missed metadata during initial digitizations by reprocessing existing data banks (Kent et al., 2017).

In addition to the changing purposes of measurements and record-keeping efforts, instruments and associated systematic SST biases during measurement have also experienced major changes. The first instruments to systematically measure SSTs across large spatial scales were buckets and

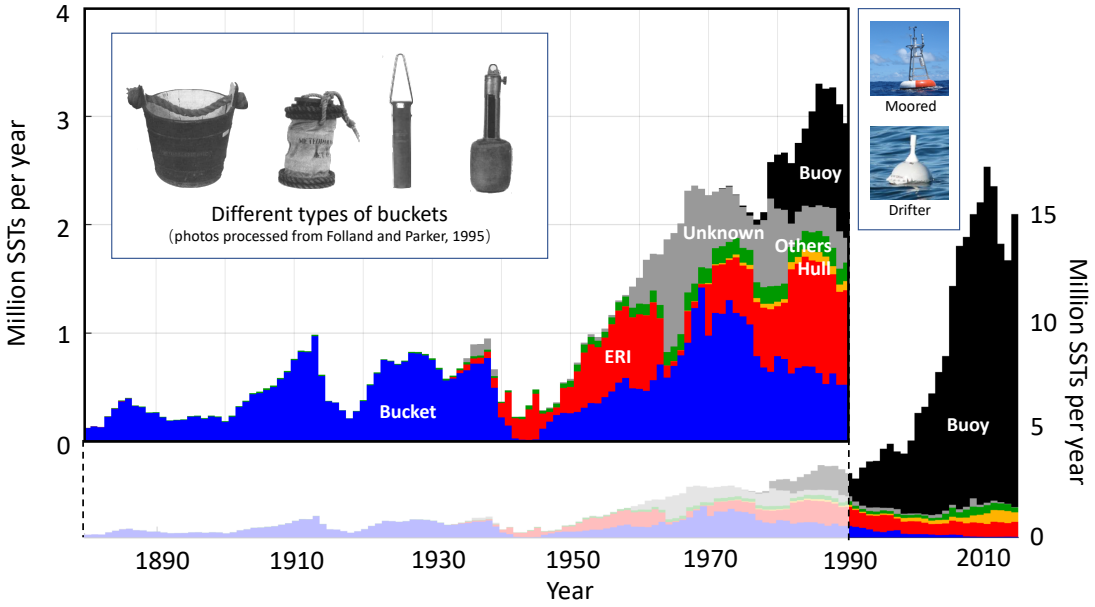


Figure 1. Distinct methods used to take in-situ SSTs compiled under the International Comprehensive Ocean Atmosphere Data Set (ICOADS). The overall number of in-situ SSTs collected by different measurement methods (stacked bars) in individual years from 1880 to 2014. In addition to bucket (blue), engine-room intake (ERI, red), hull sensor (orange), and buoy (black), other methods (green) include radiation thermometer, reversing thermometer, and electronic sensors, which are, however, not thought to be representative of SSTs due to their limited numbers (Kent et al., 2010). Results are based on version 3.0 of ICOADS. Method information is inferred for some unknown measurements, following Kennedy et al. (2011b). For example, SSTs before 1941 come from buckets, if not explicitly indicated otherwise, and U.S. and U.K. Naval SSTs during World War II are assumed to be ERI measurements (Kennedy et al., 2011b). Also shown are photos of some types of buckets used in SST collections, as well as images of moored and drifting buoys that have been widely deployed since the 1980s.

thermometers. The procedure to measure SSTs involved hauling buckets of water from the ocean surface and measuring the temperature of water in buckets on ship decks. SSTs made by this method (hereafter bucket SSTs) are thought to dominate ICOADS before the 1940s, and the number of bucket SSTs gradually decreased after the mid-1970s (Figure 1). During the measurement process, water temperatures in buckets will generally become colder due to wind-induced evaporation, as well as sensible heat loss in the tropics and the subtropics (Folland & Parker, 1995). In midlatitudes, bucket biases are still expected to be cold in winter. During summer, evaporation is suppressed in humid air, and the direction of sensible heat flux can be reversed as air temperatures become warmer than SSTs, leading to less heat loss (Folland & Parker, 1995). Sometimes, bucket water can be heated by the sun, especially on a calm summer afternoon (Kennedy et al., 2019). When averaged annually and over the globe, early bucket SSTs are estimated to be biased cold to an order of 0.4°C (Folland & Parker, 1995; Kennedy et al., 2011b). However, buckets of different materials and designs have been used under different protocols in history, which could lead to distinct biases among groups of bucket SSTs (Folland & Parker, 1995; Kent & Taylor, 2006). For example, a

less-insulated canvas bucket can be colder than a more-insulated wooden bucket by around 0.5°C even measured under the same conditions (Folland & Parker, 1995). The time gap between water retrieval and measurement can also affect bucket bias.

After the emergence of engine ships in the late-19th century, a second method of measuring SSTs was introduced, which is a byproduct when monitoring the temperature of inlet water before entering and cooling ship engines. SSTs made by the engine-room intake (ERI) method first appear in ICOADS in the 1930s (Figure 1) and are mainly from U.S. ships that dominated the Atlantic and the northeast Pacific. Later, the ERI method was adopted by more nations and gradually became the preferred method because of safety concerns associated with hauling buckets on fast engine ships (Kennedy, 2014). ERI SSTs typically come from a depth of 5–15 m where the ocean is less affected by solar heating and should consequently be colder than SSTs defined at 20–30 cm (Carella et al., 2018; Chan & Huybers, 2020b). However, because of the absorption of heat from ship engines, ERI measurements are estimated to have warm biases of 0.1°C to 0.3°C, depending on ship design and cargo (Kennedy et al., 2011b).

In the modern era, a variety of new methods that give more reliable SSTs have been used (Figure 1). Since the 1970s, an increasing number of ships are equipped with specialized digital sensors (Kennedy, 2014). SSTs from hull sensors should be free of engine heating and are therefore expected to be less biased. Scientists have also been deploying drifting and moored buoys since the late 1970s, which has become the dominant data source since the 1990s (Figure 1). Similar to hull sensors, buoys make contact with seawater directly and are expected to give less biased SST measurements, although individual buoys could be problematic due to instrumental drift or biofouling (Kennedy et al., 2012; Kent et al., 2017). Biofouling refers to the accumulation of small ocean organisms on the wet surface of instruments, leading to structural or functional deficiencies. Whereas drifting buoys typically measure at a depth of 20–30 cm, most moored buoys measure at around 1 m deep (Kennedy, 2014). The deployment of drifting buoys substantially increased the spatial coverage of the observing system, especially in the southeastern Pacific and the Southern Ocean, which nonresearch ships rarely traverse. The majority of moored buoys are installed along the coastal United States as marine weather stations and over the tropical Pacific and the Indian Ocean to monitor El Niño evolution (Hervey, 2014). Combining different types of buoys, which sample at different depth, can result in biases due to vertical temperature gradients that often exist near the ocean surface. One cause of these gradients is solar heating in low-wind conditions, which can exceed 3°C in some extreme cases (Kennedy et al., 2007). This depth effect may be damped by ship-induced turbulent mixing for ship-based SSTs and may appear small relative to bucket and ERI biases when averaged over seasons and weather conditions. However, recent ship-based SSTs are reported to be, on average, systematically warmer than collocated buoy SSTs on an order of 0.1°C (e.g., Huang et al., 2017; Karl et al., 2015), which needs to be accounted for when combining SSTs from both sources.

In addition to in-situ observations collected at the ocean surface, satellite and other remote-sensing techniques became available in the 1980s. Remote-sensing techniques further increase the spatial and temporal coverage of SST measurements. Note that satellites observe the skin temperature in the upper several millimeters of the ocean (Kennedy et al., 2007), and this difference in sampling depth, again, needs to be accounted for when homogenizing with in-situ SSTs. Additionally, SSTs retrieved from satellites can be biased due to changes in atmospheric optical depth associated with volcanic and anthropogenic aerosols and, therefore, have to be calibrated and corrected against in-situ measurements (T. M. Smith et al., 2008).

In addition to instruments dedicated to measuring SSTs, near-surface temperatures are also available from ocean profiling instruments. Historical profiles have been made on research vessels for more than a hundred years (Meyssignac et al., 2019). Since the late 1990s, scientists have been deploying Argo floats that profile temperature and conductivity as functions of pressure (Roemmich et al., 1999). Since 2006, Argo floats have been able to provide temperatures within the upper five meters of the ocean with nearly global coverage (Huang et al., 2017). Some of the most recent Argo floats can provide temperatures at 0.1-meter resolution in the upper 200 meters of the ocean. Currently, there are approximately 4,000 Argo floats providing nearly global information on near-surface temperatures at a frequency of once every 10 days.

1.2. **Insufficient Metadata and Simplified Corrections.** The shift from measuring SSTs from buckets to ERI to buoys is, therefore, accompanied by systematic biases varying on the order of 0.5°C. Such a variation in bias has a similar magnitude to the less-than-1°C global warming that is thought to have happened in the 20th century. On account of the irreplaceable nature of these early SST measurements, adjusting biases becomes crucial for quantifying and interpreting historical climate change. Such a problem, however, is difficult because biases associated with SSTs coming from the same method can be distinct due to different instrumental designs (e.g., different bucket materials) and measurement protocols used by various subsets of ships, which will interact with the uneven sampling to create regionally varying biases.

Although biases are complicated, lack of metadata by which to make specific corrections has necessitated simplifying assumptions regarding the spatial and temporal structure of SST biases, which inevitably lead to insufficient corrections and SST estimates having higher uncertainty than land surface temperatures (P. Jones, 2016). SST products from the U.K. Met Office, for example, assumed that a transition from early wooden buckets to less-insulated canvas buckets happened with the percentage of canvas buckets increasing linearly, from 35% in 1880 to 100% in 1920 over the entire ocean (Folland & Parker, 1995; Kennedy et al., 2019; Kennedy et al., 2011b). In other words, all bucket measurements in the same year are assumed to be biased in the same way, as if they were measured by the same person using the same bucket. In other SST estimates, corrections do not distinguish between measurement methods. Rather, biases for SSTs from all methods are represented using a large-scale fixed pattern, with the amplitude of the pattern estimated by comparing SSTs with other temperature estimates, for example, nighttime marine air temperatures (Huang et al., 2015; Huang et al., 2017) or coastal station-based air temperatures (Cowtan et al., 2018).

In addition to biases introduced during measurement, problems may also occur in the record-keeping and data-processing stage, as information is transferred over time and across technologies. One example is inaccurate metadata that leads to ERI SSTs being misclassified as coming from buckets (Carella et al., 2018; Kennedy et al., 2011b). Other postmeasurement problems may also exist but have not yet been quantified systematically.

2. Toward refined corrections for individual nations and groups of data

During my Ph.D. study, I aimed to refine SST corrections by resolving the regional biases that arise from different measurement and postprocessing characteristics due to distinct physical and historical reasons. Because ships from the same nation and data-collecting group would have used similar instruments, followed similar protocols, and experienced similar postprocessing practices, I corrected biases for individual nations and data-collecting groups, assuming data within the same

group have similar bias characteristics. Specifically, nation information is mainly identified using ICOADS country code (Freeman et al., 2017) and metadata from the World Meteorology Organization No. 47 publication (Kent et al., 2007). Collecting groups are assigned using ICOADS deck number, which is the primary field to track ICOADS data collection and postprocessing (Freeman et al., 2017). Because available metadata is insufficient for physically constraining biases in individual groups, I turned the research question into a big-data problem and developed a statistical method to estimate corrections required for individual groups.

Ship-based SSTs contain information of both physical SST variations and biases. Because nearby measurements tend to have similar physical SST variations, examining differences between nearby SSTs allows for focusing on data heterogeneity associated with distinct biases. SST measurements were, therefore, first paired if they came from different groups and were within 300 km and two days of one another. These scales were chosen to keep expected physical variability with biases on the order of tenths of a degree Celsius. The results are not qualitatively sensitive to scales used in pairing SSTs (Chan & Huybers, 2019). To prevent error covariance between pairs, each measurement was used only once, with an algorithm prioritizing measurements closest in space. Specifically, the method rank-orders all potential pairs within a given month according to distance and selects the closest pair. The next closest pair is selected after removing previously selected measurements. The process repeats until all paired measurements are selected (Chan & Huybers, 2019).

Because measurements in a pair are not perfectly collocated in space and time, we first removed expected physical differences arising from displacements in geographical locations, seasonality, and day-night differences (Chan & Huybers, 2019). For example, SSTs closer to the Equator or during daytime are expected to be warmer. Expected differences were estimated from the 1982–2014 climatology of high-resolution satellite-based retrievals (T. M. Smith et al., 2008). Remaining differences in reported SSTs ($\delta\mathbf{T}$) were represented using a linear-mixed-effect (LME) model,

$$\delta\mathbf{T} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}_r\boldsymbol{\beta}_r + \mathbf{Z}_y\boldsymbol{\beta}_y + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\delta\mathbf{T}$ is represented as a fixed-effect term describing offsets between groups ($\boldsymbol{\alpha}$) and random effects describing regional ($\boldsymbol{\beta}_r$) and temporal ($\boldsymbol{\beta}_y$) variations. We constrain $\boldsymbol{\alpha}$ such that the average offset of all compared measurements is zero. $\mathbf{X}$, $\mathbf{Z}_r$, and $\mathbf{Z}_y$ are design matrices that specify, respectively, common pairs of groups, years, and regions. See Figure 2 for an element-wise illustration of the LME model. Such a model is similar to an ANOVA approach. It makes use of random effects to give more conservative estimates. As the number of pairs available for constraining a random effect decreases, the estimate is relaxed toward zero such that estimates of regional and yearly variations in groupwise offsets are robust against noise (Chan & Huybers, 2019). The model is flexible in terms of controlling for specific effects and is easily extendable to account for variations in offsets associated with seasonality and day-night differences.

In practice, to reduce computational cost, SST differences are aggregated according to combinations of pairs of groups, regions, and years before estimating offsets (Chan & Huybers, 2019). Pairs in an aggregate are assigned equal weights when taking the average. Errors of aggregated SST differences ($\boldsymbol{\epsilon}$) are budgeted to account for errors from different sources and heteroscedasticity associated with distinct group size (Chan & Huybers, 2019). See Figure 3a for an example of uneven numbers of pairs between different combinations of groups. The error of each aggregated pair (ϵ_k)

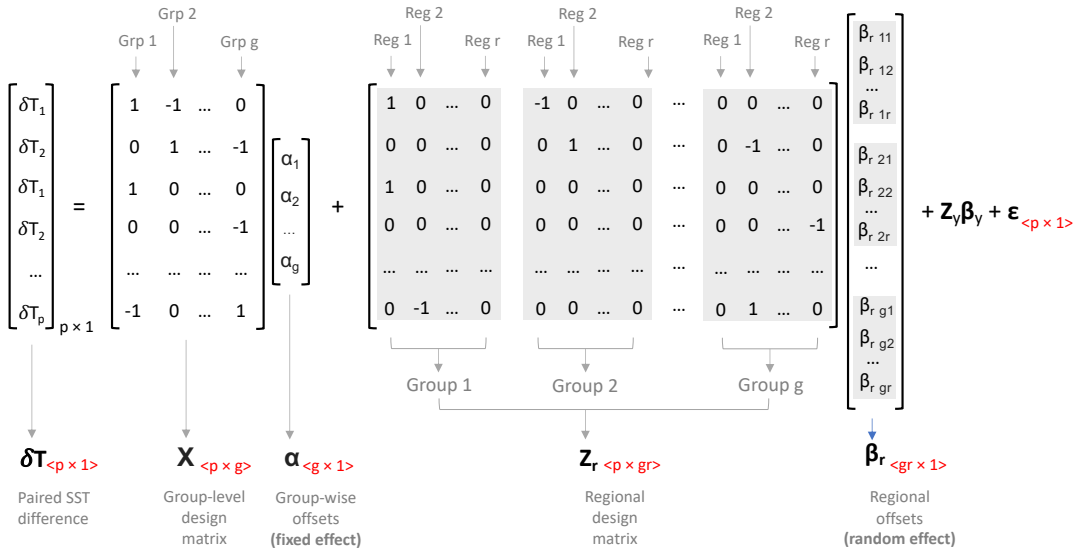


Figure 2. An element-wise illustration of the LME model in Eq. 2.1. Also shown is the dimensionality of matrices and vectors (red), where p, g, and r are, respectively, numbers of pairs, groups, and regions. $\mathbf{X}$, $\mathbf{Z}_r$, and $\mathbf{Z}_y$ are design matrices whose entries are one, zero, or minus one. Regional effects ($\boldsymbol{\beta}_r$) are estimated for individual groups. These regional effects are assigned as random effects and are assumed to follow a Gaussian distribution such that each $\beta_{r_{ij}} \sim N(0, \sigma_r^2)$. Yearly effects, $\mathbf{Z}_y\boldsymbol{\beta}_y$, are also estimated for individual groups and have a similar structure to $\mathbf{Z}_r\boldsymbol{\beta}_r$. Higher-order interactions that involve group, year, and regions are not accounted for in this model to limit the number of free parameters.

is assumed to follow $N(0, \bar{\sigma}_k^2)$, where

$$\bar{\sigma}_k^2 = \frac{2\sigma_o^2}{n_k} + \frac{2\sigma_s^2}{n_k^x} + \frac{\sum \sigma_c^2(l)}{n_k^2}. \tag{2.2}$$

$\frac{2\sigma_o^2}{n_k}$ denotes the contribution of random observational errors, where n_k is the number of pairs in the k^{th} aggregate. Random observational error is denoted by σ_o^2 and is estimated to be 0.86±0.18°C (2 *SD*, Chan et al., 2019) for individual bucket SSTs. Contributions of partially correlated observational errors are denoted by $\frac{2\sigma_s^2}{n_k^x}$, with σ_s^2 estimated to be 0.38±0.14°C (Chan et al., 2019). One possible source of σ_s^2 is systematic errors associated with individual ships. Because ship information is not always available in ICOADS, n_k^x is used to approximate effective numbers of ships within the k^{th} aggregate, with x estimated to be 0.57 (Chan et al., 2019). Finally, $\sigma_c(l)$ denotes uncertainties associated with physical SST variations for the l^{th} out of n_k pairs. The estimation of σ_c accounts for interannual variance and covariance of physical SSTs as a function of location, month, and displacement, with more details documented in section 5.a.1 of Chan and Huybers (2019). The robustness of offset estimates to a variety of model formulations and assumptions was explored in section 5.b of Chan and Huybers (2019).

In the following sections, I will show that the LME method detects significant offsets among groups classified by both nation and deck number. Accounting for these systematic groupwise offsets improves the quality of historical SST estimates at regional and sub-basin scales, resolves

several existing data-model discrepancies, and brings in new opportunities for understanding extreme weather events and climate variations. Beyond the nation-and-deck level, the model in Equation 2.1 can be extended to resolve offsets associated with individual ships for more refined SST corrections. However, metadata of ship information and the algorithm used to fit the LME model need to be improved before estimating ship-level offsets. These steps will be carried out in future works, and associated plans will be discussed in section 7.1.

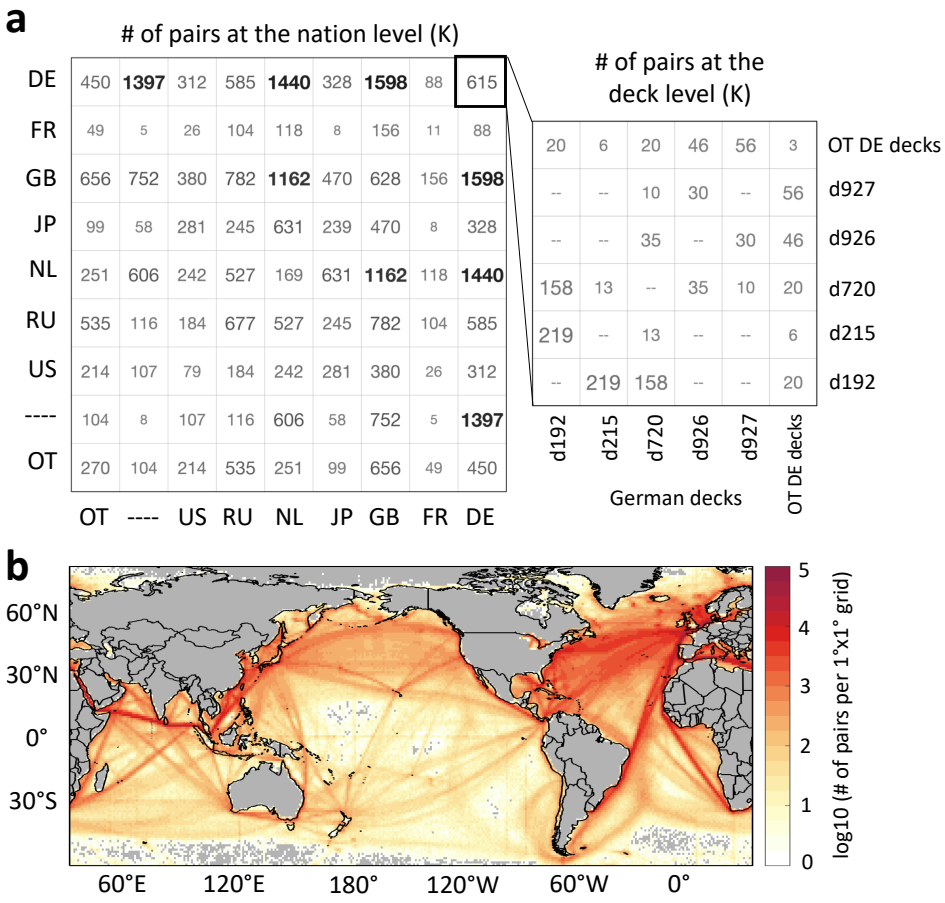


Figure 3. Schematics of an LME intercomparison. (**a**) Numbers of SST pairs between nation-deck groups (in the unit of one thousand pairs) in the analysis that intercompares SSTs thought to come from buckets. The comparison happens not only between different nations (left) but also between distinct decks from the same nation (e.g., right; zooming in the top-right box in the left and showing the comparison between German decks). The number of pairs can be very different across combinations of groups, with '- -' denoting that no pairs are found between corresponding groups. Nation abbreviations are for Germany (DE), France (FR), Great Britain (GB), Japan (JP), the Netherlands (NL), Russia (RU), the United States (US), and unknown (- -). Nations that contribute to fewer than 500,000 are labeled as "other nations" (OT) for this visualization but are distinguished in the LME analysis. Similarly, Germany decks that contribute to fewer than 50,000 pairs are shown as "OT DE decks." (**b**) The spatial distribution of paired measurements follows major ship tracks.

3. Offsets among bucket groups and more uniform early-twentieth-century warming

When applied to measurements thought to come from buckets, the LME analysis intercompares 17.8 million pairs coming from 162 groups from 1850 to 2014 (Figure 3).[1] The LME methodology detects significant ($p < 0.05$) offsets between major collecting nations and ships sailing for different purposes (Chan & Huybers, 2019). Around 15% of groups remain highly significant after controlling for family-wise error rates using a Bonferroni correction (Chan et al., 2019). A file listing individual offsets and associated uncertainty estimates can be found in the supplement to this article. The identification of significant differences among nation-and-deck groups indicates that we can resolve region- and time-varying SST biases arising from groupwise offsets and varying sampling coverage of individual groups. We, therefore, can perform more detailed corrections that have not yet been accounted for in previous studies (Cowtan et al., 2018; Huang et al., 2017; Kennedy et al., 2011b).

Removing these statistically constrained offsets provides refined SST corrections at regional scales. Central estimates of adjusted SSTs show higher interannual correlations (Pearson's r) with nearby air temperatures from coastal land stations. For example, the correlation in the early 20th century increases from 0.67 to 0.85 after groupwise bucket adjustments over coastal East Asia (Chan et al., 2019). Station-based air temperatures are independently measured using more homogeneous instruments and are expected to contain fewer spatially and temporally varying systematic biases (P. Jones, 2016). Moreover, previous corrections may have missed errors associated with groupwise offsets and, therefore, underestimated SST uncertainties at regional scales. Accounting for uncertainties of groupwise offsets increases the standard error of trend estimates to more than three times at basin scales, which, despite being higher, is a more comprehensive description of our current knowledge of uncertainties.

More importantly, accounting for groupwise bucket SST offsets reconciles a long-standing data-model discrepancy regarding the spatial heterogeneity of the early-20th-century warming. Before groupwise bucket adjustments, whereas the North Pacific warmed by around 0.3°C, the North Atlantic warming exceeded 0.8°C (Figure 4; Hegerl et al., 2018). Such a big difference in warming rate, however, cannot be reproduced by any of the IPCC AR5 models given current knowledge of external forcing and internal climate variability.

The magnitude of adjustments from individual groups depends both on the magnitudes of offsets and on the spatial and temporal coverage of distinct groups. In the early 20th century, one group that determines basin-scale SST estimates is a major subset of Japanese measurements compiled under the Kobe collection, which dominated the North Pacific before World War II (Uwai & Komura, 1992). Interestingly, Japanese Kobe SSTs, compared with nearby measurements from other nations, show a drop of around 0.35±0.07°C (2 *SD*) in the 1930s (Chan et al., 2019). Such a drop could lead to a significant underestimation of the early-20th-century warming in the Pacific. It is, therefore, crucial to understand why Japanese Kobe SSTs experienced this drop. Is it because Japan used a new type of bucket in the 1930s or did something change in the postprocessing of Japanese data? To disentangle this mystery, I combined historical approaches and physical methods.

Retrospectively, figuring out the cause is like detective work: it requires effort, persistence, and some good luck. My first hypothesis was that Japanese sailors measured SSTs on larger ships that

[1]Data associated with this analysis are available at https://doi.org/10.7910/DVN/DXJIGA.

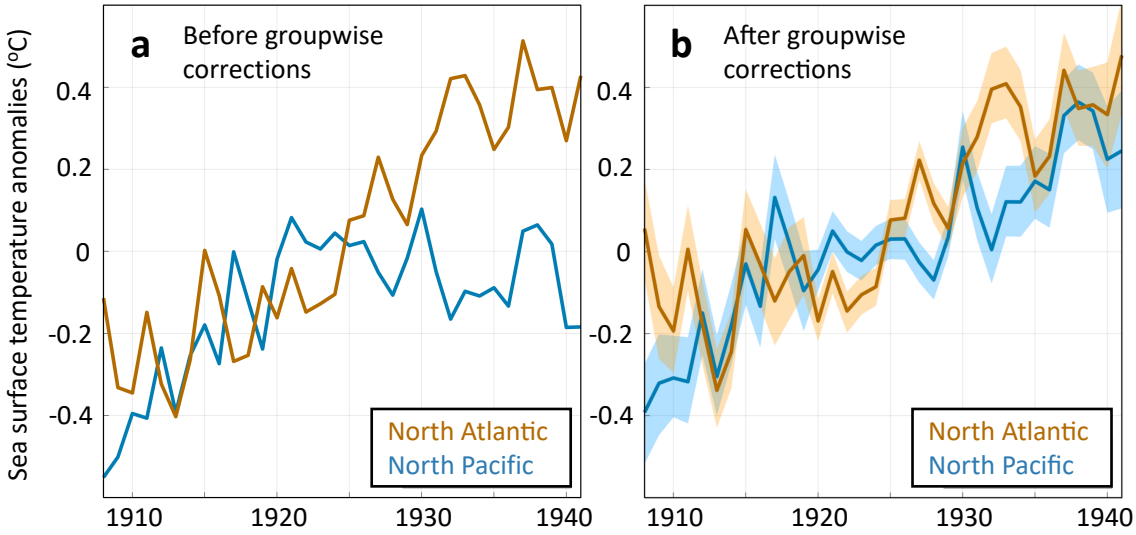
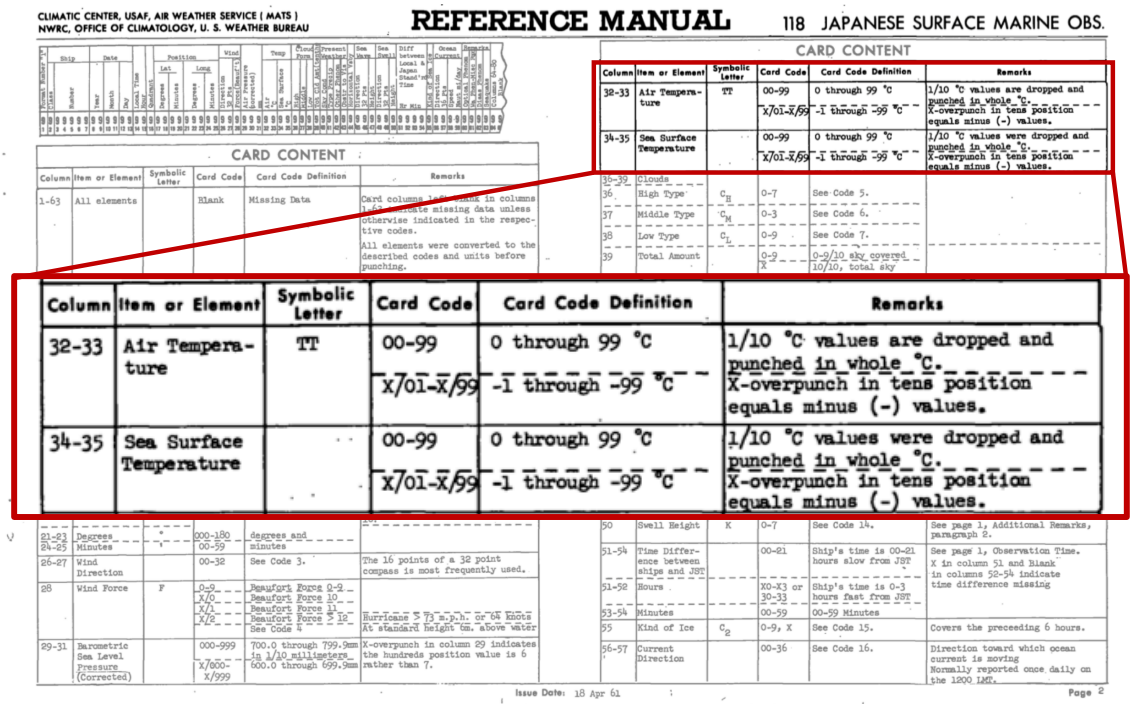


Figure 4. Basin-scale SSTs in the early twentieth century (Chan et al., 2019). (**a**) Without groupwise corrections, the annual-mean SST in the North Atlantic (red, 20°N poleward) warms more than twice as fast as that in the North Pacific (blue, 20°N poleward). Shown SSTs are based on ICOADS3.0 bucket measurements with only bulk bucket corrections that do not distinguish groupwise offsets, following Kennedy et al. (2011b). SSTs are shown as anomalies relative to the 1920–1929 average of each basin. (**b**) As (**a**) but after adjusting for groupwise offsets in bucket SSTs. Uncertainties (2 *SD*, shadings) are for annual average SSTs in each basin and are from a 1000-member ensemble of random adjustments that perturb groupwise offsets using their error estimates from the LME analysis in keeping with covariance and spatial structures.

have higher decks. Japanese ships could have increased size as circumstances deviated from the 1922 Washington Naval Treaty that limited ship displacement, and as World War II approached, the Imperial Japanese Navy required large ships for longer voyages across the Pacific. When taking bucket measurements on higher decks, it generally takes longer to haul buckets, and SSTs tend to be collected in stronger winds, leading to colder biases. To test this hypothesis, I first went through individual ships in the Imperial Japanese Navy and mapped out the evolution of the average displacement of Japanese naval ships since the 1920s. The initial result was promising: Japanese ships increased in displacement at a rate that was approximately 40% faster than the U.S. and U.K. ships in the 1930s. [2]

Despite the confirmation of a rapid increase in the displacement of Japanese naval ships, follow-up analyses served to disprove the initial hypothesis, with two pieces of evidence. First, I used a thermal model of a bucket (Folland & Parker, 1995) to simulate the influence of higher ship decks on bucket bias by increasing the hauling time and the ambient wind. The bucket model indicates that water temperatures will be further biased cold by less than 0.1°C, a magnitude that is insufficient to explain the 0.35±0.07°C drop seen in Japanese SSTs. Second, most of these large Japanese naval

[2]This is an unpublished result from my Ph.D. research, and it does not explain the drop in Japanese SSTs in the 1930s. The displacement data are from the World War II database (P. Chen, 2004).



CLIMATIC CENTER, USAF, AIR WEATHER SERVICE (MATS)
NWRC, OFFICE OF CLIMATOLOGY, U. S. WEATHER BUREAU

REFERENCE MANUAL 118 JAPANESE SURFACE MARINE OBS.

CARD CONTENT

Column	Item or Element	Symbolic Letter	Card Code	Card Code Definition	Remarks
1-63	All elements		Blank	Missing Data	Card columns left blank in columns 1-63 indicate missing data unless otherwise indicated in the respective codes. All elements were converted to the described codes and units before punching.

CARD CONTENT

Column	Item or Element	Symbolic Letter	Card Code	Card Code Definition	Remarks
32-33	Air Temperature	TT	00-99	0 through 99 °C	1/10 °C values are dropped and punched in whole °C.
			X/01-X/99	-1 through -99 °C	X-overpunch in tens position equals minus (-) values.
34-35	Sea Surface Temperature		00-99	0 through 99 °C	1/10 °C values were dropped and punched in whole °C.
			X/01-X/99	-1 through -99 °C	X-overpunch in tens position equals minus (-) values.
35-39	Clouds				
36	High Type	C_H	0-7	See Code 5.	
37	Middle Type	C_M	0-3	See Code 6.	
38	Low Type	C_L	0-9	See Code 7.	
39	Total Amount		0-9	0-9/10 sky covered	
			X	10/10, total sky	

Column	Item or Element	Symbolic Letter	Card Code	Card Code Definition	Remarks
32-33	Air Temperature	TT	00-99	0 through 99 °C	1/10 °C values are dropped and punched in whole °C.
			X/01-X/99	-1 through -99 °C	X-overpunch in tens position equals minus (-) values.
34-35	Sea Surface Temperature		00-99	0 through 99 °C	1/10 °C values were dropped and punched in whole °C.
			X/01-X/99	-1 through -99 °C	X-overpunch in tens position equals minus (-) values.

21-23	Degrees		000-180	degrees and	
24-25	Minutes		00-59	minutes	
26-27	Wind Direction		00-32	See Code 3.	The 16 points of a 32 point compass is most frequently used.
28	Wind Force	F	0-9	Beaufort Force 0-9	
			X/0	Beaufort Force 10	
			X/1	Beaufort Force 11	
			X/2	Beaufort Force 12	Hurricane 12 Bf a.p.h. or 64 Knots
				See Code 4	At standard height 6m. above water
29-31	Barometric Sea Level Pressure (Corrected)		000-999	700.0 through 799.9mm in 1/10 millimeters	X-overpunch in column 29 indicates the hundreds position value is 6 rather than 7.
			X/000-X/999	600.0 through 699.9mm	
50	Swell Height	K	0-7	See Code 14.	See page 1, Additional Remarks, paragraph 2.
51-54	Time Difference between ships and JST		00-21	Ship's time is 00-21 hours slow from JST	See page 1, Observation Time. X in column 51 and Blank in columns 52-54 indicate time difference missing
51-52	Hours		X0-X3 or 30-33	Ship's time is 0-3 hours fast from JST	
53-54	Minutes		00-59	00-59 Minutes	
55	Kind of Ice	C_2	0-9, X	See Code 15.	Covers the preceeding 6 hours.
56-57	Current Direction		00-36	See Code 16.	Direction toward which ocean current is moving Normally reported once daily on the 1200 LMT.

Issue Date: 18 Apr 61 Page 2

Figure 5. Image of a U.S. Air Force Weather Service document detailing how data from Japanese Kobe Collection deck 118 were digitized (Wilkinson et al., 2011). On the right-upper corner, it indicates that when both SSTs and marine air temperatures from this group were digitized, the data was floored to whole degrees Celsius, with all decimals dropped (https://icoads.noaa.gov/reclaim/pdf/dck118.pdf).

ships sank during battles with the U.S. Navy, and the replacement ships were small in size, yet the cold offsets remain in Japanese Kobe SSTs until the 1960s. The failure of my initial hypothesis reflects the difficulty of determining historical fact when faced with many possibilities. Fortunately, I discussed an initial manuscript with Dr. Elizabeth Kent, an expert from the U.K. National Oceanography Center who has been working on ICOADS for more than 30 years. She pointed me to an online library that documents digitization practices of many ICOADS decks. I was not aware of these documents before, and her deep expertise nudged me in the right direction. It turns out that SSTs from the Japanese Kobe collection were digitized during the recovery of logbooks and international marine data (RECLAIM) project (Wilkinson et al., 2011), with the data set divided into three subsets. The U.S. Air Force was in charge of the two parts that span from the 1930s to the early 1960s. During digitization, staff truncated Japanese temperatures and floored values to whole degrees Celsius (Figure 5), leading to the cold offsets that were prevalent in Japanese SSTs since the 1930s. Because SST measurements in ICOADS have a precision of 0.1°C, the expected cold offset due to truncation is -0.45°C when assuming that the 10th of degree digit is uniformly distributed on the values from zero to nine. The detected smaller magnitude of -0.35±0.07°C could reflect the presence of additional offsets and biases between decks.

Accounting for truncation errors in Japanese data, together with removing offsets in other groups, reveals a more homogeneous early-20th-century warming (Figure 4b). The difference in warming rates between the North Atlantic and the North Pacific decreases from 0.54°C to 0.10±0.07°C (2

SD) and becomes consistent with simulations from IPCC models (Chan et al., 2019). With quantification from the statistical method and confirmation from historical documents, we now understand that the long-standing warming discrepancy is not physical but a result of data problems as simple as truncation errors. What has happened is more consistent with physics-based expectations of uniform warming associated with anthropogenic activities (Chan et al., 2019).

4. Tracing the origin of bucket offsets using physical evidence

The reconciliation of discrepancies in the early-20th-century warming demonstrates the power of the LME method. It is also a good example of combining statistical methodologies and historical evidence to make convincing inferences and interpretations. However, despite the cause of the drop in Japanese Kobe SSTs being explained by limited historical metadata, the origin of offsets remains unclear for other groups. Due to limited metadata, the problem is approached by using features of the data. Specifically, we used a physical quantity (i.e., the diurnal cycle of SSTs) to explore the origin of groupwise offsets (Chan & Huybers, 2020b).

The diurnal cycle is variation among individual hours in a day, which can be easily estimated for each group independent of the LME methodology. Diurnal cycles are used because differences in diurnal cycles may reflect differences in measurement characteristics and could affect daily mean SSTs through physical processes. Water in buckets is subject to heat loss from the wind but is heated additionally by the sun during the daytime. As a result, if a bucket stays longer on the ship's deck before temperature is measured, it tends to have a higher day-night SST difference and overall a colder daily mean SST bias. Interestingly, when the amplitude of diurnal cycles and groupwise SST offsets are plotted against one another, the two quantities scale negatively for data in the 1980s and 1990s (Figure 6b; Chan & Huybers, 2020b), which strongly indicates the physicality of groupwise offsets detected by the LME methodology.

However, varying time on the ship's deck is not the only reason negative scalings emerge. To explore other possible origins, I extended the classic thermodynamical bucket model (Folland & Parker, 1995) to resolve bucket biases at individual local hours and simulate diurnal cycles of bucket water temperatures. Model simulations show that a negative scaling between diurnal amplitude and daily mean biases can emerge not only from varying time on deck but also from the type of bucket insulation or misclassification of ERI measurements (Chan & Huybers, 2020b). The latter arises because ERI measurements are biased warm by heat from ship engines and the ERI method samples at a depth of 5–15 m, which is less affected by diurnal variations in radiation from the sun (Carella et al., 2018). Contribution from each of these origins, however, cannot be determined from one single slope because expected slopes associated with individual origins can vary with other unknown factors, including wind and solar exposure.

To further trace the origin, we turned to the historical evolution of observed amplitude–offset relationships, which reveals that negative scalings first emerge in the 1930s (Figure 6c; Chan & Huybers, 2020b). Data before 1930, however, have a smaller range in both amplitudes and offsets and show no significant scalings (Figure 6a). Interestingly, the 1930s is also the advent of ERI measurements in ICOADS. Moreover, groups having the warmest offsets also have amplitudes of diurnal cycles that are smaller than physical SSTs at a depth of 20–30 cm, which is consistent with characteristics of ERI measurements. In other words, the misclassification of ERI measurements, as from buckets, provides the simplest explanation and is also most consistent with the history of changing data characteristics.

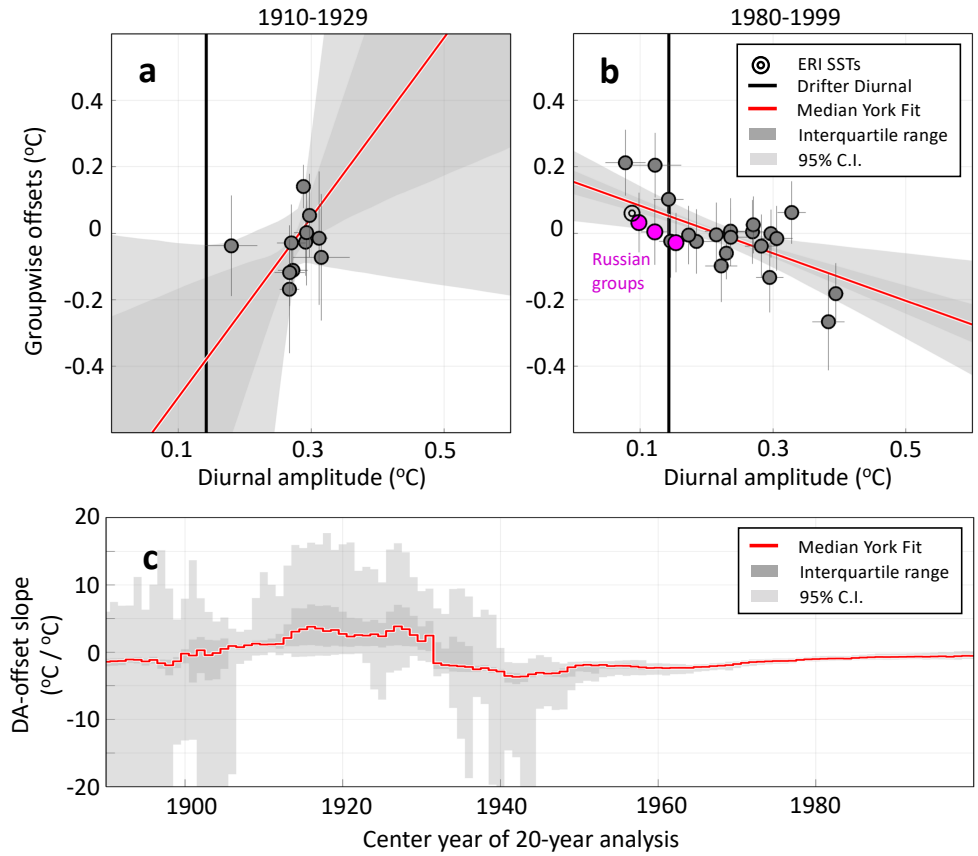


Figure 6. Groupwise SST offsets and diurnal amplitudes for groups thought to contain bucket SSTs (Chan & Huybers, 2020b). Shown results are over the tropics (20°S-20°N) and are for 20-year periods: (**a**) 1910–1929 and (**b**) 1980–1999. Groups (markers) are assigned according to nation (two-letter abbreviations, see the legend of Figure 3) and deck number. Diurnal amplitude is quantified as the amplitude of a once-per-day sinusoid using least-squares fitting. LME analyses shown here include ERI SSTs as a single group such that groupwise bucket offsets are evaluated against ERI measurements (details in Chan & Huybers, 2020b). Slopes between diurnal amplitudes and groupwise offsets (red lines) are based only on bucket groups and are estimated using York regressions (York et al., 2004). In an update to Chan and Huybers (2020b), the uncertainty of regression slopes is estimated using a stratified bootstrapping technique that resamples the entire history of individual groups with replacement (see Appendix for more details). In panel (**b**), note that the regression intersects the offset and diurnal amplitude of ERI measurements (double circles), indicating that bucket groups on the warm end of the slope (such as Russian groups shown in magenta markers) could contain misclassified ERI measurements. Also note that numerous groups show a diurnal amplitude that is similar to or lower than that of drifter SSTs (vertical black lines), which is consistent with the deep sampling depth of the ERI method. (**c**) Evolution of the amplitude–offset relationship, which is based on an analysis that uses a 20-year window and slides annually from 1880–1899 to 1990–2009. Results are shown on the center year of each 20-year analysis. Whereas highly uncertain slopes are found before the 1930s (estimates of the 2.5% quantile can be as negative as -56°C/°C before the 1910s), significant negative slopes are found afterward. Shown are median values (red curve), interquartile CI (dark shading), and 95% CI (light shading).

Analyzing diurnal cycles reveals that the record-keeping problem of incorrect metadata that mixes up bucket and ERI measurements is prevalent in data thought to come from buckets after the 1930s (Chan & Huybers, 2020b). Moreover, examining the decimal distributions of individual groups indicates that in the whole ICOADS deck, truncation is only seen in the Japanese Kobe collection. Remaining offsets, including those before the 1930s, could still arise from differences in physical processes, although, without introducing evidence from further dimensions, the trade-off among different physical factors makes it hard to attribute offsets to individual processes.

5. Beyond bucket-only SSTs — World War II warm anomaly

The LME methodology detects significant groupwise offsets that arose from both physical processes during data collection and record-keeping problems, including truncation and misrecorded metadata. This method can be easily extended beyond bucket SSTs and provides refined internal homogenization for measurements coming from various instruments. When applied to all ship-based SSTs in ICOADS, the LME method resolves another major data-model discrepancy involving excess warming during World War II (Chan & Huybers, 2020a).

Recent SST estimates feature warmer global-mean SSTs during World War II that well exceed climate-model reproductions (Figure 7a, b). Such warm anomalies are at the end of the early-20th-century warming and the beginning of the mid-20th-century hiatus and, therefore, have implications for quantifying decadal climate variations (Hansen et al., 2010; Morice et al., 2012; Vose et al., 2012), constraining uncertain aerosol forcing (Stevens, 2015), and attributing external anthropogenic forcing and internal climate variability in driving past climate change (Bindoff et al., 2013; Hegerl et al., 2018; G. S. Jones et al., 2013; Maher et al., 2014). Moreover, the World War II warm anomaly is the largest remaining data-model discrepancy in the global-mean surface temperature, given current knowledge of forcing and internal variability (Folland et al., 2018).

Such an observed warm anomaly could indicate that current models missed important physical processes. The anomaly could also arise from a 58% drop in the amount of data collected during 1942–1945 (Figure 1). Another plausible explanation involves the warm anomalies reflecting incomplete corrections of SST biases. These biases have been hypothesized to arise from a rapid increase in the number of warm-biased ERI measurements during the war (Thompson et al., 2008). However, tracing the origin of biases to specific sets of SST measurements and estimating the amount of required adjustments has not previously been possible. Existing corrections are limited by not being able to resolve offsets between groups and also due to the fact that more than 80% of wartime measurements have missing method information in raw ICOADS.

The LME method is, however, suitable in this situation of missing metadata because it allows for groupwise quantification of data heterogeneity without the need for specifying method information (although methods can be inferred from offsets and cross-checked using diurnal amplitudes, e.g., as in Figure 1b). In a recent work (Chan & Huybers, 2020a), we extended the estimation of groupwise offsets to all ship-based SST measurements in ICOADS, and the LME model quantifies that SSTs from some U.S. and U.K. naval ships that dominated data collections in World War II are, respectively, around 0.45°C and 0.25°C warmer than other groups before and after the war. These large and fast naval ships were likely to take ERI measurements (Kennedy et al., 2011b). Moreover, when further extended to resolve diurnal differences, the LME model detects an increase in nighttime measurements of around 0.3°C for many wartime non-ERI measurements (Chan & Huybers, 2020a). Such an increase is consistent with warm biases arising from measuring nighttime

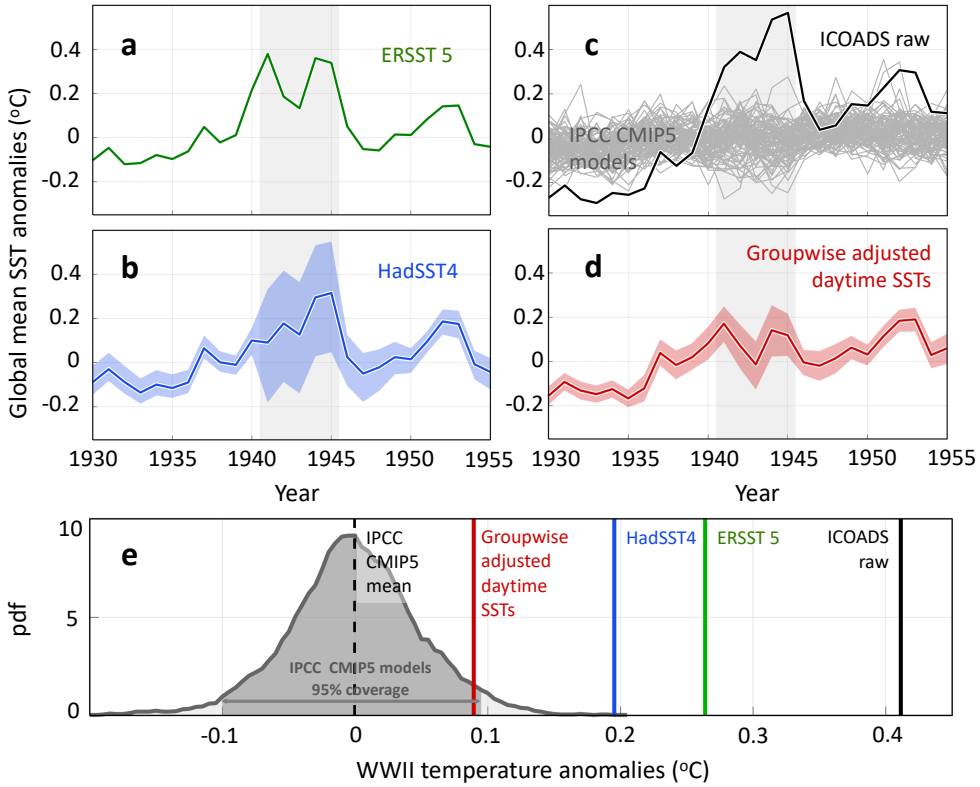


Figure 7. World War II SST anomalies in observational estimates and model simulations (Chan & Huybers, 2020a). Most recent SST estimates from (**a**) the U.S National Oceanic and Atmospheric Administration (ERSST5, green) and (**b**) U.K. Met Office (HadSST4, blue) show warm anomalies in global-mean SSTs that exceed 0.2°C during World War II. The axes for (**c**) and (**d**) are as in (**a**), but (**c**) shows raw ship-based SSTs in ICOADS (black) and (**d**) shows daytime ship-based SSTs after groupwise adjustments (red). Shown time series are global-averaged SST anomalies relative to the 20-year average over 1931–1940 and 1946–1955, where the global average is taken over grid boxes containing major ship tracks in the early and mid-20th century. Uncertainties are 95% CI (blue and red shading) from ensemble corrections of corresponding estimates. Also shown is an ensemble of 94 historical all-forcing simulations from 39 IPCC models (light gray curves in **c**; Taylor et al., 2012). In (**e**), the World War II warm anomaly in groupwise adjusted daytime SSTs (red) becomes consistent with the range of internal variations estimated from IPCC models (gray distribution). The World War II anomaly is quantified as the difference between averages over 1941–1945 and over the surrounding 10 years (1936–1940 and 1946–1950).

bucket SSTs inside ships to avoid detection, a wartime practice documented for the U.K. Navy (Folland et al., 1984).

The effect of groupwise offsets and nighttime bucket biases contributes to, respectively, 0.26°C (95% CI 0.15°C–0.38°C) and 0.05°C (0.02°C–0.08°C) warm anomalies in raw ICOADS (Chan & Huybers, 2020a). Adjustments bring the World War II warm anomaly from 0.41°C in raw ICOADS to 0.09°C (-0.01°C to 0.18°C), which becomes consistent with the ±0.10°C range (95% CI) of internal variability in IPCC models (Figure 7b, c; Chan & Huybers, 2020a). Groupwise adjustments

based on the LME methodology lead to more homogeneous spatial and temporal variations in SSTs and reconcile the largest remaining data-model discrepancy in global-mean surface temperatures.

Fixing problems in the WWII warm anomaly confirms the hypothesis of data biases, as suggested by Thompson et al. (2008). This piece of work also provides us with a lesson that historical data may contain not only physical but also social aspects. When interpreting historical data, especially in the context of comparing with model simulations, it is often implicitly assumed that data reflect the physical, or broadly, the scientific dimension. This assumption could be valid when a small amount of data is calibrated carefully for a specific scientific purpose. But for massive data sets that pool information from heterogeneous sources and take generations to construct, it is crucial to keep an awareness of the social dimension and the people involved in data collection and processing, especially over periods having dramatic social changes.

6. Beyond SSTs — Hindcasting of North Atlantic hurricanes

Adjustments of groupwise offsets have been shown to improve historical SST estimates and reconcile major data-model discrepancies in surface temperature evolution. On account of the broad climatic applications of SSTs, the implication of improved SST corrections, however, is not limited to simple year-to-year variations or linear trends. Beyond surface temperatures, improvements associated with groupwise SST adjustments could also advance other fields in atmospheric and ocean sciences.

One example is the hindcasting of North Atlantic hurricane activities. Decadal variations in the frequency of North Atlantic hurricanes are known to depend on patterns of tropical SSTs (Vecchi, Msadek, et al., 2013; Vecchi et al., 2008). Compared with observational reconstructions of Atlantic hurricane counts, dynamical climate models prescribed with historical SSTs as boundary conditions,[3] however, reproduce too few hurricanes in the late-19th century and too many in the mid-20th century (p <0.05, Figure 8a, Chan et al., 2020). The inability for models to skillfully reproduce a long-term evolution of hurricane counts that are statistically consistent with observational estimates erodes the credibility of future projections based on these models (Vecchi et al., 2019). Possible causes for this low reproducibility include inaccurate hurricane reconstructions (Vecchi & Knutson, 2008) and model deficiency (Zhao et al., 2009). In addition, biases in SSTs may also undermine simulations of hurricane genesis.

The SST value that is thought to affect North Atlantic hurricane frequency is the difference between the subtropical North Atlantic and the whole tropical ocean. This SST difference is also known as 'relative SST' (RSST) and is thought to influence hurricane genesis through affecting convective activities and potential energy (Vecchi, Fueglistaler, et al., 2013; Vecchi et al., 2011). Groupwise bucket SST corrections increase RSST in the late 19th century and decrease RSST in the mid-20th century. In a recent work in which I collaborated with colleagues from Princeton University (Chan et al., 2020), we incorporated groupwise bucket corrections to previous SST estimates and found that simulated hurricane counts also show increases in the late-19th century

[3]Although SSTs are not entirely independent of hurricanes, these simulations should have partially accounted for the effect that hurricanes lower SSTs by using observed monthly SSTs. Moreover, the response of SSTs to hurricanes is not expected to change much with time (Vecchi et al., 2019). Thus, using monthly SSTs as boundary conditions will not alter the active and inactive phases in the simulated decadal variability of hurricane counts.

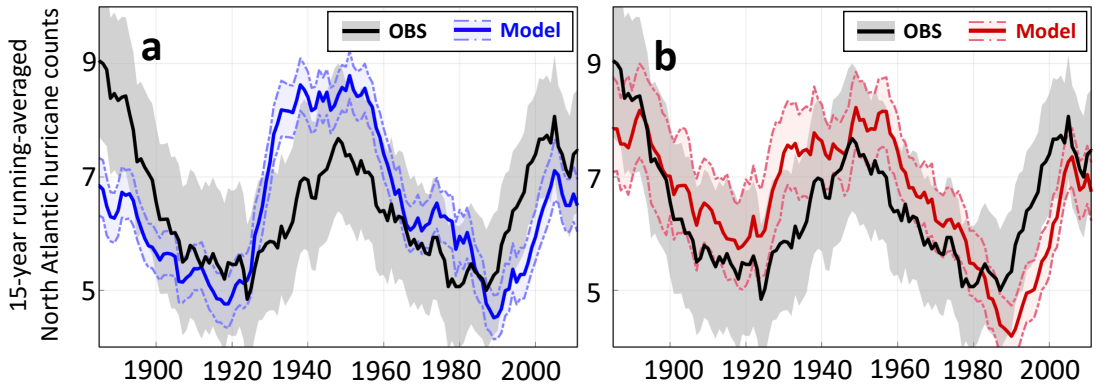


Figure 8. 15-year running-averaged North Atlantic hurricane counts in observational reconstructions and model simulations (Chan et al., 2020). (**a**) Simulations (blue, average of a five-member ensemble) using SST estimates without groupwise bucket corrections give significantly ($p < 0.05$) lower hurricane counts than observational estimates (black) in the late 19th century and higher counts in the mid-20th century. (**b**) Simulated (red, average of a five-member ensemble) and observed (black) hurricane counts become consistent using SSTs that include groupwise bucket SST corrections. Shown curves are 15-year running-averaged rather than raw integer counts because we are interested in the decadal variability of North Atlantic hurricane frequency. Uncertainties (95% CI) account for atmospheric internal variability and errors in hurricane adjustments (gray shading), atmospheric internal variability (blue shading in panel a), and atmospheric internal variability and errors in uncertain groupwise SST corrections (red shading). Distinct types of errors are assumed to be independent of one another. For estimates containing errors from two sources (i.e., black and red lines), shown uncertainties are summations of squared errors from both sources. A detailed description of the error model is in the method section of Chan et al. (2020). Note that atmospheric internal variability arises from perturbations to initial conditions, and that variability in observations is not expected to be reproduced by models because of imperfect initial conditions. It is, therefore, necessary to consider atmospheric internal variability as random error in both observation and simulation.

and decreases in the mid-20th century, consistent with expectations from adjusting RSST. More importantly, simulated hurricane counts become statistically consistent with independently reconstructed observational estimates of hurricane counts after accounting for groupwise SST offsets (Figure 8b; Chan et al., 2020).

Showing that SST biases are the dominant limiting factor for models to recover historical Atlantic hurricane counts is exciting news for both the SST and hurricane communities. The diminishing data-model discrepancy in hurricane variability provides dynamical evidence to buttress the improved quality of SSTs after groupwise corrections. On the other hand, the more stable relationship between observed and simulated hurricane activity increases the credibility of dynamical models in making accurate predictions of future changes in hurricane activities.

7. What is next?

Correcting national and groupwise offsets improves historical SSTs and reconciles a number of data-model mismatches. Despite these significant improvements, estimates of historical SSTs are still far from perfect, and there is much scope for further improvements. In addition to recovering lost data sets and missed metadata (as suggested in, e.g., Kent et al., 2017), opportunities exist to develop new techniques and further analyze existing data sets.

7.1. **Internal Homogeneity to the Level of Individual Ships.**

Further improvements could come from better resolving internal heterogeneity at more refined levels, such as quantifying offsets associated with distinct measurement characteristics of individual ships. Ship-level biases can lead to partially correlated errors across space and time as ships passing through different grid boxes (Kennedy, 2014). Ship-level biases were estimated to have a similar magnitude to random measurement errors by comparing with satellite (Kennedy et al., 2012) or observational-constrained model simulations (Kent & Berry, 2008) using data in recent decades. Ship-level biases, however, have not yet been explicitly quantified for data before the 1970s. In version 3 of the HadSST data set, biases of individual ships were assumed to follow a Gaussian distribution that has a zero mean and a standard deviation of ship-level biases (Kennedy et al., 2011a), where the ship-level standard deviation was estimated by comparing ship-based SSTs with collocated satellite retrievals since the 1990s (Kennedy et al., 2012). Uncertainties associated with ship-level biases were inferred for gridded data sets after estimating the effective number of ships in individual 5° boxes (Kennedy et al., 2011a). Such a treatment better accounted for error covariance but could not remove biases associated with individual ships. In version 4 of HadSST (the latest version), ship-based SSTs were compared against uppermost temperature measurements from ocean profiles for data after World War II, with the assumption that profile measurements are free of bias (Kennedy et al., 2019). Kennedy et al. (2019) assumed that ship-based SSTs and profile temperatures follow a multivariate normal distribution, which allows for estimating biases of gridded ship-based SST fields at the level of individual grid boxes. Such a method, to some extent, has implicitly accounted for ship-level biases.

The LME method appears to be a suitable approach for estimating offsets between individual ships, which further increases internal homogeneity among measurements within ICOADS. Methodologically, quantifying ship-level offsets can be realized by assigning random effects for individual ships. A systematic implementation, however, is currently limited by the quality of ship information. A total of 44% of paired bucket SSTs from 1850 to 1970 do not have ship identifiers in raw ICOADS. Moreover, around 85% of ships having IDs have no more than 25 paired measurements, which is too few for robust offset estimates because random observational errors are estimated to be 0.74°C (Kennedy et al., 2012). To improve ship information, Carella et al. (2017) tried to track measurements with missing ship IDs and combine short tracks into longer ones. The tracking algorithm of Carella et al. (2017), however, is uncertain at ship crossings on which the LME intercomparison entirely relies. A careful examination of the suitability of these tracked ships is, therefore, required before estimating ship-level offsets using the LME method. Other opportunities may come from redigitizing early ship logs or developing more robust tracking algorithms. In addition to improving metadata of ship information, the algorithm of fitting the LME model also needs to be modified to account for hundreds of thousands of additional parameters associated with individual ships.

7.2. **External Consistency with Independent Instrumental Measurements and Paleoproxies.** Potential for further improvements may also come from improving external consistency with independent temperature measurements or proxies. Measurements from marine air temperatures (Huang et al., 2017; T. M. Smith & Reynolds, 2002) and coastal land stations (Cowtan et al., 2018) have been used to quantify SST biases at the global scale. Recent studies also use subsurface temperatures from ocean profiling data to estimate SST biases at both global and regional scales since the 1940s (e.g., Huang et al., 2018; Kennedy et al., 2019). In addition to these external data sources, I would like to call attention to the potential value of ocean temperatures at deeper depth and proxies from coral reefs.

The deep ocean communicates with the surface through convective and diffusive processes (Gebbie & Huybers, 2011). Deep-ocean temperatures largely reflect SST variations at high latitudes where ocean convection is most active (Gebbie & Huybers, 2011). Variability in deep-ocean temperatures, however, needs to be interpreted cautiously on account of possible smoothing associated with eddy diffusion, less constrained variability of ocean circulation, and a time lag between the surface and interior ocean. Alternatively, SSTs may contain information to constrain and reconstruct ocean circulation. Furthermore, similar to the SST problem, profiles that contain deep-ocean temperatures also come from various methods and nations (Meyssignac et al., 2019). Quality controls that involve group- or ship-level examination to profile data using the LME method appear necessary before calibrating SSTs.

There could also be value in paleoclimate proxies, a data source often considered to have higher noise and be less reliable than instrumental measurements. Mechanistically, heavier isotopes tend to enrich in the condensed phase due to kinetic fractionation (Urey, 1947). Heavy oxygen isotopes (e.g., O^{18}) in coral reefs will, therefore, decrease with water temperature, providing long-term and homogeneous approximations of SSTs (e.g., Gagan et al., 2000). Pfeiffer et al. (2017) showed that proxy temperatures from coral reefs in the Indian Ocean do not show abrupt changes during World War II, which is consistent with our groupwise corrections. Coral reefs have the benefit of a broader coverage in tropical oceans, including the eastern Pacific, which is not frequently sampled by ships. Caution is required when interpreting oxygen isotopes. In certain regions that have abundant rainfall, such as the intertropical convergence zone, the concentration of O^{18} in rainwater (thus seawater and coral reefs) decreases strongly with increasing rainfall, which could mask temperature signals (Gagan et al., 2000; Lee & Fung, 2008; Pfeiffer et al., 2017).

7.3. **New Mapping Techniques.** In addition to correcting SST biases, an equally important problem in SST reconstructions involves mapping and infilling grids without observations to have global coverage. Unlike typical kernel functions that have decaying covariance with increasing displacement, kernels preferred in climate sciences should account for covariance associated with large-scale variations in ocean and atmosphere (e.g., El Niño and Southern Oscillation). Most previous SST estimates use principal component analysis (PCA) to learn patterns of covariance from satellite observations since the 1980s and assume stationarity (e.g., Hirahara et al., 2014; Huang et al., 2017; Rayner et al., 2003), even though satellite retrievals show that the details of the SST patterns are distinct across El Niño events in the past 40 years (Timmermann et al., 2018). Variational Bayesian methods have been proposed to learn patterns from ship-based SSTs that have a longer history (Ilin & Kaplan, 2009). Reconstructions from this method, however, contain patterns of ship tracks that we do not expect to exist in physical SSTs (Kennedy et al., 2013). The most recent advance involves using inpainting techniques in artificial intelligence and

learning patterns from climate model simulations (Kadow et al., 2020), but whether such a method gives reliable error estimates remains questionable. As a result, a statistically rigorous mapping technique that accounts for physical climatic covariance assuming potentially nonstationary spatial covariance is necessary for reconstructing past climate variability and budgeting uncertainties.

7.4. **A Unified Statistical Framework.** Quantification and correction of SST biases are often treated as separate steps from mapping and infilling for existing SST estimates (e.g., Hirahara et al., 2014; Huang et al., 2017; Rayner et al., 2003). Moreover, the mapping procedure is often further divided into separately performed substeps according to spatial scales of infilling. Such frameworks, however, make it difficult to budget and synthesize uncertainties in SST estimates arising from distinct analyzing steps.

Developing a holistic statistical framework that unifies distinct steps appears to be a solution. Such a framework should incorporate random error, systematic biases, and physical variations of global SSTs simultaneously with fully resolved covariance. Moreover, on account of potentially large uncertainties in estimates of physical SST covariance and observational errors and biases, a Bayesian method may be a more suitable framework for comprehensive quantification of SST uncertainties. Ideally, this framework should also take in biases and uncertainties we have learned from existing works and other pieces of useful information from independent external measurements. The Bayesian framework developed by Tingley and Huybers (2010) appears to be a valid starting point.

8. Conclusion

Understanding the history of data is crucial, and one needs to be particularly careful when using data outside their historical context. Most historical SSTs were not collected for studying climate change. These measurements contain various biases due to distinct physical and historical reasons during data collection and postprocessing. Although these SSTs have irreplaceable value for understanding past climate variations, they are undercalibrated to have sufficient accuracy for climatic use. Contrary to complicated biases, previous SST corrections that assumed homogeneous bias structures appear oversimplified, which motivated our scrutinizing historical data. Insufficient corrections have led to substantial remaining errors that result in discrepancies between observations and model simulations of the historical period. When data and models disagree, one can almost always adjust the models so that they better reproduce the data, but being aware of the underlying assumption that data reflect reality and being skeptical about data quality appears to be a good practice.

My Ph.D. work is one step forward toward better resolving complicated SST biases and toward a more accurate depiction of the past climate. Our LME method is ignorant of the existence of data-model discrepancies, but accounting for data heterogeneity among nations and groups of collectors reconciles several data-model discrepancies and provides a more comprehensive estimate of SST uncertainties. These improvements will not be achieved and consolidated without combining evidence from statistical, physical, and historical aspects. Even though not assumed or built-in, the updated SST estimates show simpler spatial and temporal variations and are more in line with expected patterns of warming. Bringing observational estimates into accord with our current knowledge of forcing, climate sensitivity, and internal variability leads to greater confidence in future predictions of global warming made by climate models.

Disclosure Statement. D.C. is supported by a grant from the Harvard Global Institute and has no conflicts of interest to disclose.

Acknowledgments

I acknowledge the associate editor, the dataviz editor, and two anonymous reviewers for their constructive comments that greatly improved the quality of this review. I thank my advisor Peter Huybers for advice and helpful discussions throughout my Ph.D. study. I also thank Elizabeth C. Kent, Gabriel A. Vecchi, Wenchang Yang, and David I. Berry for their collaboration and discussions on specific sections of my PhD project. This review was initiated at the 2020 Harvard Horizons program, and I acknowledge Xiaoli Meng, Edward J. Hall, Pamela Pollock, Hardeep Dhillon, the six other fellow scholars, and the staff at the Derek Bok Center for discussions on an initial version of Sections 1–3 of this review.

References

Bindoff, N. L., Stott, P. A., AchutaRao, K. M., Allen, M. R., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., et al. (2013). Detection and attribution of climate change: from global to regional. *Climate Change 2013 – the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.022

Carella, G., Kennedy, J. J., Berry, D., Hirahara, S., Merchant, C. J., Morak-Bozzo, S., & Kent, E. C. (2018). Estimating sea surface temperature measurement methods using characteristic differences in the diurnal cycle. *Geophysical Research Letters*, *45*(1), 363–371. https://doi.org/10.1002/2017GL076475

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

Chan, D., & Huybers, P. (2019). Systematic differences in bucket sea surface temperature measurements among nations identified using a linear-mixed-effect method. *Journal of Climate*, *32*(9), 2569–2589. https://doi.org/10.1175/JCLI-D-18-0562.1

Chan, D., & Huybers, P. (2020a). Identifying and correcting the World War 2 warm anomaly in sea surface temperature measurements. *EarthArXiv preprint*. https://doi.org/10.31223/osf.io/ju26e

Chan, D., & Huybers, P. (2020b). Systematic differences in bucket sea surface temperatures caused by misclassification of engine room intake measurements. *Journal of Climate*, *33*(18), 7735–7753. https://doi.org/10.1175/JCLI-D-19-0972.1

Chan, D., Kent, E. C., Berry, D. I., & Huybers, P. (2019). Correcting datasets leads to more homogeneous early-twentieth-century sea surface warming. *Nature*, *571*(7765), 393. https://doi.org/10.1038/s41586-019-1349-2

Chan, D., Vecchi, G. A., Yang, W., & Huybers, P. (2020). Correcting sea surface temperatures improves simulations of historical hurricane activity. *EarthArXiv preprint*. https://doi.org/10.31223/osf.io/huz73

Chen, P. (2004). World War II Database.

Chen, X., & Tung, K.-K. (2014). Varying planetary heat sink led to global-warming slowdown and acceleration. *Science*, *345*(6199), 897–903. https://doi.org/10.1126/science.1254937

Cowtan, K., Rohde, R., & Hausfather, Z. (2018). Evaluating biases in sea surface temperature records using coastal weather stations. *Quarterly Journal of the Royal Meteorological Society*, *144*(712), 670–681. https://doi.org/10.1002/qj.3235

Easterling, D. R., & Wehner, M. F. (2009). Is the climate warming or cooling? *Geophysical Research Letters*, *36*(8). https://doi.org/10.1029/2009GL037810

Folland, C. K., Boucher, O., Colman, A., & Parker, D. E. (2018). Causes of irregularities in trends of global mean surface temperature since the late 19th century. *Science Advances*, *4*(6), EAAO5297. https://doi.org/10.1126/sciadv.aao5297

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

Folland, C. K., Parker, D., & Kates, F. (1984). Worldwide marine temperature fluctuations 1856–1981. *Nature*, *310*(5979), 670–673. https://doi.org/10.1038/310670a0

Freeman, E., Woodruff, S. D., Worley, S. J., Lubker, S. J., Kent, E. C., Angel, W. E., Berry, D. I., Brohan, P., Eastman, R., Gates, L., et al. (2017). ICOADS Release 3.0: a major update to the historical marine climate record. *International Journal of Climatology*, *37*(5), 2211–2232. https://doi.org/10.1002/joc.4775

Fyfe, J. C., Gillett, N. P., & Zwiers, F. W. (2013). Overestimated global warming over the past 20 years. *Nature Climate Change*, *3*(9), 767–769. https://doi.org/10.1038/nclimate1972

Gagan, M., Ayliffe, L., Beck, J. W., Cole, J., Druffel, E., Dunbar, R. B., & Schrag, D. (2000). New views of tropical paleoclimates from corals. *Quaternary Science Reviews*, *19*(1-5), 45–64. https://doi.org/10.1016/S0277-3791(99)00054-2

Gates, W. L., Boyle, J. S., Covey, C., Dease, C. G., Doutriaux, C. M., Drach, R. S., Fiorino, M., Gleckler, P. J., Hnilo, J. J., Marlais, S. M., et al. (1999). An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). *Bulletin of the American Meteorological Society*, *80*(1), 29–56. https://doi.org/10.1175/1520-0477(1999)080<0029:AOOTRO>2.0.CO;2

Gebbie, G., & Huybers, P. (2011). How is the ocean filled? *Geophysical Research Letters*, *38*(6). https://doi.org/10.1029/2011GL046769

Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, *48*(4). https://doi.org/10.1029/2010RG000345

Hartmann, D. L., Tank, A. M. K., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y. A. R., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., et al. (2013). Observations: atmosphere and surface. *Climate Change 2013 – the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.008

Hausfather, Z., Cowtan, K., Clarke, D. C., Jacobs, P., Richardson, M., & Rohde, R. (2017). Assessing recent warming using instrumentally homogeneous sea surface temperature records. *Science Advances*, *3*(1), e1601207. https://doi.org/10.1126/sciadv.1601207

Hegerl, G. C., Brönnimann, S., Schurer, A., & Cowan, T. (2018). The early 20th century warming: anomalies, causes, and consequences. *Wiley Interdisciplinary Reviews: Climate Change*, *9*(4), e522. https://doi.org/10.1002/wcc.522

Hervey, R. V. (2014). Meteorological and oceanographic data collected from the National Data Buoy Center Coastal-Marine Automated Network (C-MAN) and moored (weather) buoys during

2014-03 (NODC Accession 0117682). Version 1.1. National Oceanographic Data Center, NOAA. Dataset. https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.nodc:0117682

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

Huang, B., Angel, W., Boyer, T., Cheng, L., Chepurin, G., Freeman, E., Liu, C., & Zhang, H.-M. (2018). Evaluating SST analyses with independent ocean profile observations. *Journal of Climate*, *31*(13), 5015–5030. https://doi.org/10.1175/JCLI-D-17-0824.1

Huang, B., Banzon, V. F., Freeman, E., Lawrimore, J., Liu, W., Peterson, T. C., Smith, T. M., Thorne, P. W., Woodruff, S. D., & Zhang, H.-M. (2015). Extended reconstructed sea surface temperature version 4 (ERSST. v4). Part I: Upgrades and intercomparisons. *Journal of Climate*, *28*(3), 911–930. https://doi.org/10.1175/JCLI-D-14-00006.1

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., & Zhang, H.-M. (2017). Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *Journal of Climate*, *30*(20), 8179–8205. https://doi.org/10.1175/JCLI-D-16-0836.1

Ilin, A., & Kaplan, A. (2009). Bayesian PCA for reconstruction of historical sea surface temperatures. *2009 International Joint Conference on Neural Networks*, 1322–1327. https://doi.org/10.1109/IJCNN.2009.5178744

Jones, G. S., Stott, P. A., & Christidis, N. (2013). Attribution of observed historical near–surface temperature variations to anthropogenic and natural causes using CMIP5 simulations. *Journal of Geophysical Research: Atmospheres*, *118*(10), 4001–4024. https://doi.org/10.1002/jgrd.50239

Jones, P. (2016). The reliability of global and hemispheric surface temperature records. *Advances in Atmospheric Sciences*, *33*(3), 269–282. https://doi.org/10.1007/s00376-015-5194-4

Kadow, C., Hall, D. M., & Ulbrich, U. (2020). Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, 1–6. https://doi.org/10.1038/s41561-020-0582-5

Karl, T. R., Arguez, A., Huang, B., Lawrimore, J. H., McMahon, J. R., Menne, M. J., Peterson, T. C., Vose, R. S., & Zhang, H.-M. (2015). Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, *348*(6242), 1469–1472. https://doi.org/10.1126/science.aaa5632

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. https://doi.org/10.1002/2013RG000434

Kennedy, J. J., Brohan, P., & Tett, S. (2007). A global climatology of the diurnal variations in sea-surface temperature and implications for MSU temperature trends. *Geophysical Research Letters*, *34*(5). https://doi.org/10.1029/2006GL028920

Kennedy, J. J., Rayner, N., Atkinson, C., & Killick, R. (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. https://doi.org/10.1029/2018JD029867

Kennedy, J. J., Rayner, N., Smith, R., Parker, D., & Saunby, M. (2011a). Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *Journal of Geophysical Research: Atmospheres*, *116*(D14). https://doi.org/10.1029/2010JD015218

737 Kennedy, J. J., Rayner, N., Smith, R., Parker, D., & Saunby, M. (2011b). Reassessing biases and
738 other uncertainties in sea surface temperature observations measured in situ since 1850:
739 2. Biases and homogenization. *Journal of Geophysical Research: Atmospheres*, *116*(D14).
740 https://doi.org/10.1029/2010JD015220
741 Kennedy, J. J., Rayner, N., Saunby, M., & Millington, S. (2013). Bringing together measurements of
742 sea surface temperature made in situ with retrievals from satellite instruments to create a
743 globally complete analysis for 1850 onwards, HadISST2. *EGU General Assembly Conference*
744 *Abstracts*, *15*.
745 Kennedy, J. J., Smith, R. O., & Rayner, N. A. (2012). Using AATSR data to assess the qual-
746 ity of in situ sea-surface temperature observations for climate studies. *Remote Sensing of*
747 *Environment*, *116*, 79–92. https://doi.org/10.1016/j.rse.2010.11.021
748 Kent, E. C., & Berry, D. I. (2008). Assessment of the marine observing system (ASMOS): final
749 report.
750 Kent, E. C., Kennedy, J. J., Berry, D. I., & Smith, R. O. (2010). Effects of instrumentation changes
751 on sea surface temperature measured in situ. *Wiley Interdisciplinary Reviews: Climate*
752 *Change*, *1*(5), 718–728. https://doi.org/10.1002/wcc.55
753 Kent, E. C., Kennedy, J. J., Smith, T. M., Hirahara, S., Huang, B., Kaplan, A., Parker, D. E.,
754 Atkinson, C. P., Berry, D. I., Carella, G., et al. (2017). A call for new approaches to
755 quantifying biases in observations of sea surface temperature. *Bulletin of the American*
756 *Meteorological Society*, *98*(8), 1601–1616. https://doi.org/10.1175/BAMS-D-15-00251.1
757 Kent, E. C., & Taylor, P. K. (2006). Toward estimating climatic trends in SST. Part I: Methods
758 of measurement. *Journal of Atmospheric and Oceanic Technology*, *23*(3), 464–475. https:
759 //doi.org/10.1175/JTECH1843.1
760 Kent, E. C., Woodruff, S. D., & Berry, D. I. (2007). Metadata from WMO publication no. 47 and an
761 assessment of voluntary observing ship observation heights in ICOADS. *Journal of Atmo-*
762 *spheric and Oceanic Technology*, *24*(2), 214–234. https://doi.org/10.1175/JTECH1949.1
763 Kosaka, Y., & Xie, S.-P. (2013). Recent global-warming hiatus tied to equatorial Pacific surface
764 cooling. *Nature*, *501*(7467), 403–407. https://doi.org/10.1038/nature12534
765 Lapenis, A. G. (1998). Arrhenius and the Intergovernmental Panel on Climate Change. *Eos, Trans-*
766 *actions American Geophysical Union*, *79*(23), 271–271. https://doi.org/10.1029/98EO00206
767 Lee, J.-E., & Fung, I. (2008). ‘‘Amount effect” of water isotopes and quantitative analysis of post-
768 condensation processes. *Hydrological Processes: An International Journal*, *22*(1), 1–8. https:
769 //doi.org/10.1002/hyp.6637
770 Maher, N., Gupta, A. S., & England, M. H. (2014). Drivers of decadal hiatus periods in the 20th
771 and 21st centuries. *Geophysical Research Letters*, *41*(16), 5978–5986. https://doi.org/10.
772 1002/2014GL060527
773 Medhaug, I., Stolpe, M. B., Fischer, E. M., & Knutti, R. (2017). Reconciling controversies about the
774 global warming hiatus. *Nature*, *545*(7652), 41–47. https://doi.org/10.1038/nature22315
775 Meyssignac, B., Boyer, T., Zhao, Z., Hakuba, M. Z., Landerer, F. W., Stammer, D., Köhl, A., Kato,
776 S., L’Ecuyer, T., Ablain, M., et al. (2019). Measuring global ocean heat content to estimate
777 the Earth energy imbalance. *Frontiers in Marine Science*, *6*, 432. https://doi.org/10.3389/
778 fmars.2019.00432
779 Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. (2012). Quantifying uncertainties
780 in global and regional temperature change using an ensemble of observational estimates:

781 The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, *117*(D8). https:
782 //doi.org/10.1029/2011JD017187
783 Pfeiffer, M., Zinke, J., Dullo, W.-C., Garbe-Schönberg, D., Latif, M., & Weber, M. (2017). In-
784 dian Ocean corals reveal crucial role of World War II bias for twentieth century warming
785 estimates. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-14352-6
786 Rayner, N., Parker, D. E., Horton, E., Folland, C. K., Alexander, L., Rowell, D., Kent, E. C.,
787 & Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night ma-
788 rine air temperature since the late nineteenth century. *Journal of Geophysical Research:*
789 *Atmospheres*, *108*(D14). https://doi.org/10.1029/2002JD002670
790 Roemmich, D., Boebel, O., Desaubies, Y., Freeland, H., King, B., LeTraon, P.-Y., Molinari, R.,
791 Owens, B., Riser, S., Send, U., et al. (1999). Argo: The global array of profiling floats.
792 *CLIVAR Exchanges*, *13*(4 (3)), 4–5.
793 Shankar, P. (2020). Tutorial overview of simple, stratified, and parametric bootstrapping. *Engineer-*
794 *ing Reports*, *2*(1), e12096. https://doi.org/https://doi.org/10.1002/eng2.12096
795 Smith, T. M., & Reynolds, R. W. (2002). Bias corrections for historical sea surface temperatures
796 based on marine air temperatures. *Journal of Climate*, *15*(1), 73–87. https://doi.org/10.
797 1175/1520-0442(2002)015<0073:BCFHSS>2.0.CO;2
798 Smith, T. M., Reynolds, R. W., Peterson, T. C., & Lawrimore, J. (2008). Improvements to NOAA's
799 historical merged land-ocean surface temperature analysis (1880-2006). *Journal of Climate*,
800 *21*(10), 2283–2296. https://doi.org/10.1175/2007JCLI2100.1
801 Stevens, B. (2015). Rethinking the lower bound on aerosol radiative forcing. *Journal of Climate*,
802 *28*(12), 4794–4819. https://doi.org/10.1175/JCLI-D-14-00656.1
803 Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment
804 design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/
805 10.1175/BAMS-D-11-00094.1
806 Thompson, D. W., Kennedy, J. J., Wallace, J. M., & Jones, P. D. (2008). A large discontinuity in
807 the mid-twentieth century in observed global-mean surface temperature. *Nature*, *453*(7195),
808 646–649. https://doi.org/10.1038/nature06982
809 Timmermann, A., An, S.-I., Kug, J.-S., Jin, F.-F., Cai, W., Capotondi, A., Cobb, K. M., Lengaigne,
810 M., McPhaden, M. J., Stuecker, M. F., et al. (2018). El Niño–southern oscillation complex-
811 ity. *Nature*, *559*(7715), 535–545. https://doi.org/10.1038/s41586-018-0252-6
812 Tingley, M. P., & Huybers, P. (2010). A Bayesian algorithm for reconstructing climate anomalies
813 in space and time. Part I: Development and applications to paleoclimate reconstruction
814 problems. *Journal of Climate*, *23*(10), 2759–2781. https://doi.org/10.1175/2009JCLI3015.1
815 Urey, H. C. (1947). The thermodynamic properties of isotopic substances. *Journal of the Chemical*
816 *Society (Resumed)*, 562–581. https://doi.org/10.1039/jr9470000562
817 Uwai, T., & Komura, K. (1992). The collection of historical ships' data in Kobe marine observatory.
818 *Proceedings of the International COADS Workshop, Boulder, CO, USA*, 13–15.
819 Vecchi, G. A., Delworth, T. L., Murakami, H., Underwood, S. D., Wittenberg, A. T., Zeng, F.,
820 Zhang, W., Baldwin, J. W., Bhatia, K. T., Cooke, W., et al. (2019). Tropical cyclone sen-
821 sitivities to CO2 doubling: roles of atmospheric resolution, synoptic variability and back-
822 ground climate changes. *Climate Dynamics*, *53*(9-10), 5999–6033. https://doi.org/10.1007/
823 s00382-019-04913-y

824 Vecchi, G. A., Fueglistaler, S., Held, I. M., Knutson, T. R., & Zhao, M. (2013). Impacts of atmo-
825 spheric temperature trends on tropical cyclone activity. *Journal of Climate*, *26*(11), 3877–
826 3891. https://doi.org/10.1175/JCLI-D-12-00503.1
827 Vecchi, G. A., & Knutson, T. R. (2008). On estimates of historical North Atlantic tropical cyclone
828 activity. *Journal of Climate*, *21*(14), 3580–3600. https://doi.org/10.1175/2008JCLI2178.1
829 Vecchi, G. A., Msadek, R., Anderson, W., Chang, Y.-S., Delworth, T., Dixon, K., Gudgel, R.,
830 Rosati, A., Stern, B., Villarini, G., Wittenberg, A., Yang, X., Zeng, F., Zhang, R., &
831 Zhang, S. (2013). Multiyear predictions of North Atlantic hurricane frequency: Promise
832 and limitations. *Journal of Climate*, *26*(15), 5337–5357. https://doi.org/10.1175/JCLI-D-
833 12-00464.1
834 Vecchi, G. A., Swanson, K. L., & Soden, B. J. (2008). Whither hurricane activity? *Science*, 687–689.
835 https://doi.org/10.1126/science.1164396
836 Vecchi, G. A., Zhao, M., Wang, H., Villarini, G., Rosati, A., Kumar, A., Held, I. M., & Gudgel,
837 R. (2011). Statistical–dynamical predictions of seasonal North Atlantic hurricane activity.
838 *Monthly Weather Review*, *139*(4), 1070–1082. https://doi.org/10.1175/2010MWR3499.1
839 Vose, R. S., Arndt, D., Banzon, V. F., Easterling, D. R., Gleason, B., Huang, B., Kearns, E., Law-
840 rimore, J. H., Menne, M. J., Peterson, T. C., et al. (2012). NOAA's merged land–ocean sur-
841 face temperature analysis. *Bulletin of the American Meteorological Society*, *93*(11), 1677–
842 1685. https://doi.org/10.1175/BAMS-D-11-00241.1
843 Wilkinson, C., Woodruff, S. D., Brohan, P., Claesson, S., Freeman, E., Koek, F., Lubker, S. J.,
844 Marzin, C., & Wheeler, D. (2011). Recovery of logbooks and international marine data:
845 the RECLAIM project. *International Journal of Climatology*, *31*(7), 968–979. https://doi.
846 org/10.1002/joc.2102
847 Woodruff, S. D., Diaz, H. F., Elms, J. D., & Worley, S. J. (1998). COADS Release 2 data and meta-
848 data enhancements for improvements of marine surface flux fields. *Physics and Chemistry*
849 *of the Earth*, *23*(5-6), 517–526. https://doi.org/10.1016/S0079-1946(98)00064-0
850 Woodruff, S. D., Slutz, R. J., Jenne, R. L., & Steurer, P. M. (1987). A comprehensive ocean-
851 atmosphere data set. *Bulletin of the American Meteorological Society*, *68*(10), 1239–1250.
852 https://doi.org/10.1175/1520-0477(1987)068<1239:ACOADS>2.0.CO;2
853 Woodruff, S. D., Worley, S. J., Lubker, S. J., Ji, Z., Eric Freeman, J., Berry, D. I., Brohan, P.,
854 Kent, E. C., Reynolds, R. W., Smith, S. R., et al. (2011). ICOADS Release 2.5: extensions
855 and enhancements to the surface marine meteorological archive. *International journal of*
856 *climatology*, *31*(7), 951–967. https://doi.org/10.1002/joc.2103
857 Worley, S. J., Woodruff, S. D., Reynolds, R. W., Lubker, S. J., & Lott, N. (2005). ICOADS re-
858 lease 2.1 data and products. *International Journal of Climatology: A Journal of the Royal*
859 *Meteorological Society*, *25*(7), 823–842. https://doi.org/10.1002/joc.1166
860 Yeh, S.-W., Kug, J.-S., Dewitte, B., Kwon, M.-H., Kirtman, B. P., & Jin, F.-F. (2009). El Niño in
861 a changing climate. *Nature*, *461*(7263), 511. https://doi.org/10.1038/nature08316
862 York, D., Evensen, N. M., Martınez, M. L., & De Basabe Delgado, J. (2004). Unified equations
863 for the slope, intercept, and standard errors of the best straight line. *American Journal of*
864 *Physics*, *72*(3), 367–375. https://doi.org/10.1119/1.1632486
865 Zhao, M., Held, I. M., Lin, S.-J., & Vecchi, G. A. (2009). Simulations of global hurricane climatology,
866 interannual variability, and response to global warming using a 50-km resolution GCM.
867 *Journal of Climate*, *22*(24), 6653–6678. https://doi.org/10.1175/2009JCLI3049.1

Appendix

Stratified Bootstrapping for Estimating Uncertainties in the Evolution of the Amplitude–Offset Relationship. Similar to Figure 6c, Chan and Huybers (2020b) also investigated the evolution of the amplitude–offset relationship using a sliding 20-year window. The bootstrapping in Chan and Huybers (2020b) resamples available groups in each 20-year analysis independently, which gives a reasonable uncertainty estimate within each 20-year analysis but may not be optimal for intercomparing slopes across 20-year analyses. Here, I supplement earlier estimates by resampling groups with their entire history of diurnal amplitudes and groupwise offsets, which also estimates the path-wise uncertainty. Moreover, to account for the reduced number of groups before the 1950s, a stratified resampling scheme (Shankar, 2020) is used to guarantee that the resampled groups better reflect the prevalence of groups throughout the history of marine observation. Specifically, groups are divided into two strata based on whether they were present in 20-year windows before 1940–1959. The resampling is then performed within each stratum with replacement and repeated 10,000 times. On average, updating the bootstrapping technique slightly increases the 95% CI of York fit slopes by 6%, and the interquartile range by 1%. Among the 10,000 time series of bootstrapped York regression slope, 9,306 of them have, on average, positive values over 1910–1929 but negative values afterward, indicating that the relationship between diurnal amplitudes and groupwise offsets changed sign significantly ($p < 0.1$) in the 1930s. A Matlab script to reproduce this updated bootstrapping analysis is in the supplement to this article.